PREDICCIÓN DE TASA DE INTERÉS Y SU RELACIÓN CON INDICADORES ECONÓMICOS MEDIANTE CIENCIA DE DATOS

Sebastián Javier Meza Pastrana
8985595

Proyecto Aplicado para optar al título de

Magister en Ciencia de Datos

Director(a)

David Arango Londoño

FACULTAD DE INGENIERÍA Y CIENCIAS

MAESTRÍA EN CIENCIA DE DATOS

SANTIAGO DE CALI, DICIEMBRE 05 DE 2024

TABLA DE CONTENIDO

INTRODUCCIÓN **1**

1. DEFINICIÓN DEL PROBLEMA **2**

1.1. PLANTEAMIENTO DEL PROBLEMA 2
1.2. FORMULACIÓN DEL PROBLEMA 3

2. OBJETIVOS DEL PROYECTO **4**

2.1. OBJETIVO GENERAL 4
2.2. OBJETIVOS ESPECÍFICOS 4

3. MARCO TEÓRICO Y ANTECEDENTES **5**

3.1. MARCO TEÓRICO 5
3.2. ANTECEDENTES 12

4. IDENTIFICACIÓN Y ADQUISICIÓN DE FUENTES DE DATOS CON INDICADORES ECONÓMICOS **15**

4.1 Introducción 15
4.2 Actividad 1: Revisión de Fuentes de Datos 15
4.3 Actividad 2: Selección de Variables Relevantes 16
4.4 Actividad 3: Adquisición de Datos 20
4.5 Conclusión del Capítulo 21

5. APLICACIÓN DE TÉCNICAS DE PREPROCESAMIENTO DE DATOS **22**

5.1 Introducción 22
5.2 Actividad 1: Limpieza de Datos 22
5.3 Actividad 2: Estandarización de los Datos 22
5.4. Actividad 3: Manejo de Datos Temporales 23

6. EVALUACIÓN Y SELECCIÓN DE MODELOS DE MACHINE LEARNING PARA LA PREDICCIÓN DE TASAS DE INTERÉS **25**

6.1. Introducción 25
6.2. Descripción 26
6.3. Modelo Ridge 28
6.4. Modelo Lasso 32
6.5. Modelo Random Forest 36
6.6. Modelo Gradient Boosting – XGBoost 39
6.7. Modelo Support Vector Regressor (SVR) 42
6.8. Comparación general 45

7. ESTRATEGIAS DE OPTIMIZACIÓN Y SELECCIÓN DE VARIABLES DE MODELOS INICIALES IMPLEMENTADOS **47**

7.1. Introducción 47
7.2. Análisis de correlación 48
7.4. Modelo Ridge RFE 52
7.5. Modelo Lasso RFE 59
7.6. Modelo Random Forest RFE 68
7.7. Modelo Gradient Boosting – XGBoost RFE 77

7.8. MODELO SUPPORT VECTOR REGRESSOR RFE 85
7.9. COMPARACIÓN GENERAL MODELOS 93
8. MODELO ECONOMETRICO Y COMPARACIÓN CON MODELOS DE MACHINE LEARNING 96
9. CONCLUSIONES Y TRABAJOS FUTUROS 102
9.1. CONCLUSIONES 102
9.2. TRABAJOS FUTUROS 103
10. BIBLIOGRAFÍA 104
11. ANEXOS 106

LISTA DE ILUSTRACIONES

Ilustración 1. Resultados multicolinealidad Ridge 29

Ilustración 2. Curva aprendizaje Ridge 30

Ilustración 3. Resultados comparativos Ridge inicial 31

Ilustración 4. Resultados multicolinealidad Lasso 32

Ilustración 5. Resultados optimización grid search Lasso 33

Ilustración 6. Curva aprendizaje Lasso 35

Ilustración 7. Resultados comparativos Lasso inicial 35

Ilustración 8. Resultados multicolinealidad Random Forest 36

Ilustración 9. Curva aprendizaje Random Forest 38

Ilustración 10. Resultados comparativos Random Forest inicial 38

Ilustración 11. Resultados multicolinealidad XGBoost 39

Ilustración 12. Resultados optimización grid search XGBoost 40

Ilustración 13. Resultados validación cruzada XGBoost 41

Ilustración 14. Curva aprendizaje XGBoost 41

Ilustración 15. Resultados comparativos XGBoost inicial 42

Ilustración 16. Resultados multicolinealidad SVR 43

Ilustración 17. Resultados optimización grid search SVR 43

Ilustración 18. Resultados validación cruzada SVR 44

Ilustración 19. Curva aprendizaje SVR 45

Ilustración 20. Resultados comparativos SVR inicial 45

Ilustración 21. Resultado RFE modelo Ridge 52

Ilustración 22. Curva aprendizaje Modelo Ridge 54

Ilustración 23. Análisis sensibilidad modelo Ridge 54

Ilustración 24. Análisis sensibilidad hiperparámetro modelo Ridge 55

Ilustración 25. Análisis residuos modelo Ridge 55

Ilustración 26. Análisis de robustez modelo Ridge..... 57

Ilustración 27. Análisis de interpretabilidad SHAP modelo Ridge 57

Ilustración 28. Selección de características RFE modelo Lasso 60

Ilustración 29. Curva aprendizaje modelo Lasso 62

Ilustración 30. Análisis sensibilidad modelo Lasso..... 63

Ilustración 31. Análisis sensibilidad hiperparámetro modelo Lasso 63

Ilustración 32. Análisis de residuos modelo Lasso 64

Ilustración 33. Análisis robustez modelo Lasso 66

Ilustración 34. Análisis interpretabilidad SHAP modelo Lasso 66

Ilustración 35. Selección características RFE modelo Random Forest..... 69

Ilustración 36. Curva aprendizaje modelo Random Forest..... 71

Ilustración 37. Análisis sensibilidad hiperparámetro modelo Random Forest 73

Ilustración 38. Análisis residuos modelo Random Forest 73

Ilustración 39. Análisis robustez modelo Random Forest 74

Ilustración 40. Análisis interpretabilidad SHAP modelo Random Forest 75

Ilustración 41. Selección características modelo XGBoost 77

Ilustración 42. Curva aprendizaje modelo XGBoost 80

Ilustración 43. Análisis sensibilidad hiperparámetro modelo XGBoost 81

Ilustración 44. Análisis residuos modelo XGBoost 82

Ilustración 45. Análisis robustez modelo XGBoost 82

Ilustración 46. Análisis interpretabilidad SHAP modelo XGBoost 83

Ilustración 47. Selección características modelo SVR 86

Ilustración 48. Curva aprendizaje modelo SVR 87

Ilustración 49. Análisis sensibilidad hiperparámetro modelo SVR 88

Ilustración 50. Análisis residuos modelo SVR 89

Ilustración 51. Análisis robustez modelo SVR 90

Ilustración 52. Análisis interpretabilidad SHAP modelo SVR 91

Ilustración 53. Multicolinealidad modelo econométrico inicial 97

Ilustración 54. Métricas modelo econométrico inicial 97

Ilustración 56. Multicolinealidad modelo econométrico ajustado 98

Ilustración 57. Métricas modelo econométrico ajustado 99

LISTA DE TABLAS

Tabla 1. Indicadores económicos 17

Tabla 2. Tabla comparativa métricas implementación inicial modelos 46

Tabla 3. Comparación general modelos optimizados 93

Tabla 4. Métricas modelos de machine learning y modelo econométrico 100

INTRODUCCIÓN

La predicción de la tasa de interés de política monetaria ha sido uno de los desafíos más complejos en la economía contemporánea, especialmente en economías emergentes como la colombiana. En este contexto, la tasa de interés no solo constituye un instrumento esencial de la política monetaria, sino que también tiene implicaciones profundas en la estabilidad macroeconómica, las decisiones de inversión y el bienestar económico de los agentes. Su correcta estimación puede reducir incertidumbres y mejorar la planificación estratégica en diversos sectores. El Banco de la República, en su rol como entidad rectora de la política monetaria en Colombia, ha utilizado la tasa de interés como un mecanismo para mantener bajo control la inflación y estabilizar las fluctuaciones económicas. Sin embargo, la complejidad inherente a la dinámica económica y la interacción de variables nacionales e internacionales ha hecho que esta tarea sea altamente desafiante. Las aproximaciones tradicionales, como los modelos econométricos, han sido efectivas en ciertos contextos, pero han mostrado limitaciones importantes cuando se enfrentan a relaciones no lineales, alta dimensionalidad de los datos y volatilidad en las series temporales.

Con la evolución de la ciencia de datos, se han desarrollado herramientas avanzadas como el Machine Learning (ML), que permiten procesar grandes volúmenes de información y capturar patrones complejos en los datos. Este enfoque ha representado una oportunidad para superar las limitaciones de las técnicas convencionales, al tiempo que mejora la precisión predictiva y proporciona una comprensión más profunda de las dinámicas económicas subyacentes.
Este trabajo tiene como objetivo evaluar y comparar modelos avanzados de Machine Learning con enfoques econométricos tradicionales para la predicción de la tasa de interés de política monetaria en Colombia. Se implementaron modelos de aprendizaje automático, tales como Random Forest, Gradient Boosting (XGBoost), Ridge, Lasso y Support Vector Regression (SVR), complementados con un modelo econométrico de referencia. La metodología empleada abarcó desde la selección y preprocesamiento de datos hasta la optimización de hiperparámetros y el análisis de interpretabilidad, garantizando así la solidez de los resultados obtenidos.

Este proyecto contribuye a la literatura existente al explorar dos enfoques, el respaldo teórico económico de modelos econométricos con la flexibilidad, precisión e interpretabilidad de los modelos de Machine Learning. Los resultados obtenidos proporcionan una base sólida para el diseño de políticas económicas más informadas y para la toma de decisiones estratégicas en entornos de alta incertidumbre. Este trabajo tiene implicaciones prácticas para formuladores de políticas, instituciones financieras y agentes económicos en general, marcando un camino hacia un entendimiento más profundo y eficiente de los determinantes económicos en Colombia.
.

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

La tasa de interés es un instrumento central en la política monetaria de Colombia y juega un papel crucial en la estabilidad económica del país. Su correcta administración por parte del Banco de la República permite controlar la oferta monetaria y regular el costo del crédito, afectando directamente el comportamiento económico de todos los sectores productivos [1]. Sin embargo, prever la evolución de la tasa de interés sigue siendo un desafío debido a la complejidad de los factores que la influyen, tanto a nivel nacional como internacional. Esta incertidumbre genera problemas tanto para los formuladores de política económica como para los agentes económicos que dependen de estas tasas para tomar decisiones financieras estratégicas [1].

¿Por qué es un problema prever la tasa de interés?

Predecir la tasa de interés es problemático debido a la multiplicidad de factores que la afectan y la naturaleza altamente volátil de algunos de estos elementos. Entre los factores más críticos se encuentran la inflación, el crecimiento económico, la tasa de cambio y el comportamiento de los mercados internacionales [2], [3]. Estos indicadores no siempre siguen patrones lineales o predecibles, lo que aumenta la dificultad de anticipar con precisión los cambios en la tasa de interés [3]. Los modelos econométricos tradicionales, aunque efectivos en ciertos contextos, a menudo fallan al manejar grandes volúmenes de datos o detectar relaciones no lineales entre variables [2].

Esta falta de precisión en la predicción de las tasas de interés tiene consecuencias significativas. Las instituciones financieras, como bancos y entidades de crédito, dependen de estas predicciones para gestionar sus estrategias de liquidez y definir las tasas de los préstamos y depósitos [2], [3]. De igual manera, las empresas necesitan prever los costos de financiamiento para planificar inversiones a largo plazo [2]. Por otro lado, los individuos, especialmente aquellos con préstamos o hipotecas, ven afectadas sus decisiones sobre el endeudamiento y el ahorro [2] [3].

Impacto de un modelo de predicción de tasas de interés avanzado

El desarrollo de un modelo de predicción de tasas de interés basado en técnicas de Machine Learning puede tener un impacto significativo en diversos sectores. Para los formuladores de política económica, como el Banco de la República, este tipo de modelo ofrecería herramientas más precisas para ajustar la política monetaria en función de condiciones económicas en constante cambio, permitiendo intervenciones más efectivas y oportunas [1].

En el sector financiero, las entidades bancarias y las casas de bolsa podrían utilizar estas predicciones para ajustar sus ofertas de productos financieros, como tasas de interés en préstamos y productos de inversión [2]. Las empresas, por su parte, estarían en mejor posición para planificar inversiones y estrategias de financiamiento, minimizando los riesgos asociados a

los costos variables de capital [3].

Para los individuos, un modelo más preciso les permitiría tomar decisiones informadas sobre préstamos, hipotecas y ahorro, al contar con una mejor comprensión de cómo podrían evolucionar las tasas de interés [2], [3]. Esto es particularmente relevante para aquellos con préstamos a tasas variables, que podrían estar expuestos a fluctuaciones significativas en sus costos financieros [2].

El problema de predecir la tasa de interés, por tanto, no solo es un desafío técnico, sino que también tiene profundas implicaciones económicas y sociales. La incertidumbre en esta predicción afecta tanto a grandes instituciones financieras como a pequeños inversionistas y hogares [1]. El enfoque propuesto en este proyecto busca cerrar esta brecha de incertidumbre, utilizando técnicas avanzadas de ciencia de datos para mejorar la precisión de las predicciones y, con ello, ofrecer herramientas más robustas para la toma de decisiones en el ámbito financiero y empresarial [2], [3].

1.2. FORMULACIÓN DEL PROBLEMA

¿Cómo se puede estimar la tasa de interés de política monetaria en la economía colombiana mediante la aplicación de modelos avanzados de Machine Learning, utilizando variables e indicadores económicos tanto nacionales como internacionales?

Preguntas de sistematización

- ¿Cuáles son las fuentes de datos más relevantes y confiables para la recopilación de variables e indicadores económicos, tanto a nivel nacional como internacional, necesarios para la estimación de tasas de interés en la economía colombiana?
- ¿Cuáles son las técnicas más apropiadas de preprocesamiento de datos para la calidad y coherencia de la información utilizada en los modelos?
- ¿Qué modelos de Machine Learning son los más adecuados para la predicción de las tasas de interés?
- ¿Cómo se puede optimizar la selección de variables económicas para maximizar la precisión del modelo de estimación de tasas de interés?
- ¿Cuáles son las métricas de evaluación más apropiadas para medir la precisión y robustez de los modelos de Machine Learning en el contexto de la estimación de las tasas de interés?
- ¿Cuál es la eficacia relativa de los modelos de Machine Learning en comparación con los modelos econométricos tradicionales, para predecir la tasa de interés en la economía colombiana mediante la incorporación de variables e indicadores económicos tanto nacionales como internacionales, con el objetivo de proporcionar estimaciones más precisas y decisiones informadas?

2. OBJETIVOS DEL PROYECTO

2.1. OBJETIVO GENERAL

Desarrollar y evaluar modelos avanzados de Machine Learning con el propósito de estimar la tasa de interés de política monetaria en la economía colombiana utilizando diversas variables e indicadores económicos.

2.2. OBJETIVOS ESPECÍFICOS

- Identificar y adquirir fuentes de datos con indicadores económicos para la estimación de las tasas de interés.
- Aplicar técnicas de preprocesamiento de datos con el objetivo de garantizar la calidad, consistencia y coherencia de la información utilizada en los modelos de Machine Learning.
- Evaluar y seleccionar modelos específicos de Machine Learning que sean apropiados para la predicción de tasas de interés.
- Implementar estrategias de optimización de la selección de variables económicas con el fin de maximizar la precisión y relevancia del modelo de estimación de tasas de interés.
- Realizar una comparación entre modelos de Machine Learning seleccionados y los modelos econométricos tradicionales, incorporando variables e indicadores económicos elegidos. Evaluar los resultados obtenidos a través de métricas específicas para cada uno de estos enfoques, proporcionando así una evaluación detallada del rendimiento de cada modelo.

3. MARCO TEÓRICO Y ANTECEDENTES

3.1. MARCO TEÓRICO

El presente marco teórico proporciona una visión integral para comprender la complejidad de la estimación de las tasas de interés abordando tanto los fundamentos de la política monetaria en Colombia como el uso de modelos de ciencia de datos para abordar los desafíos particulares de la economía, proporcionando una base sólida para comprender las dinámicas económicas y explorar posibles soluciones.

3.1.1. Economía

3.1.1.1. Política Monetaria y Tasas de Interés

En el marco de la economía colombiana, la tasa de interés se configura como una herramienta esencial en la política monetaria, siendo administrada mediante operaciones ejecutadas por el Banco de la República. Este banco sigue un proceso estructurado para tomar decisiones sobre la acción de Política Monetaria, en el cual participan la Junta Directiva del Banco de la República (JDBR), el Gerente Técnico, el Subgerente de Política Monetaria e Información Económica, así como los departamentos asociados a esta Subgerencia. A lo largo de este procedimiento, se busca abordar cuestionamientos fundamentales sobre la situación económica actual, la dirección que está tomando la economía y, por último, determinar la senda de la tasa de interés de referencia que conduce hacia la meta de inflación establecida y estabiliza el producto en sus niveles de largo plazo [4]. La meta primordial de la política monetaria es ejercer influencia sobre las condiciones económicas y financieras del país. Este control se logra mediante la regulación de la oferta monetaria y las tasas de interés, con el propósito central de gestionar y mantener bajo control la tasa de inflación.

Las tasas de interés, como instrumento de la política monetaria, simbolizan la tasa mínima que el Banco de la República (BanRep) establece para las entidades financieras [5]. Esta tasa, también conocida como tasa de intervención o tasa de referencia, desempeña un papel central en la regulación económica al influir en la disponibilidad y el costo del dinero en la economía colombiana [2]. Al establecer esta tasa, BanRep busca no solo controlar la liquidez del sistema financiero sino también incidir en el comportamiento económico y financiero a nivel más amplio [6]. Este mecanismo de política monetaria impacta directamente en la dinámica económica y social del país. Una tasa de interés baja tiende a estimular el endeudamiento y la inversión, fomentando así el crecimiento económico y la creación de empleo [7]. Por otro lado, una tasa alta puede desincentivar el endeudamiento, pero también puede afectar el consumo y la inversión. En este sentido, las tasas de interés no solo son un indicador económico, sino una herramienta estratégica que moldea las decisiones financieras de individuos y empresas, con repercusiones significativas en la actividad económica y el bienestar social. La relación entre las tasas de interés y la sociedad se manifiesta en diversos aspectos, desde el acceso al crédito por parte de los ciudadanos hasta la rentabilidad de las inversiones empresariales [8]. Un entorno de tasas de interés estables y predecibles puede generar un ambiente propicio para la planificación financiera y el desarrollo económico sostenible [8]. Por otro lado, variaciones bruscas o extremas en las

tasas pueden generar incertidumbre y afectar la toma de decisiones, influyendo en la capacidad de las personas para acceder a préstamos, invertir o planificar sus finanzas a largo plazo [8].

3.1.1.2. Factores determinantes en las decisiones de tasas de interés

Según lo indicado por el Banco de la República, las tasas de interés se utilizan como instrumento para gestionar la inflación, estableciendo así una conexión intrínseca entre ambas. En el contexto colombiano, la medición de la inflación se lleva a cabo a través del índice de precios al consumo (IPC). Según Mankiw, la inflación se describe como "incremento general en el nivel de precios" y en términos técnicos como el "cambio porcentual en el índice de precios al consumidor (IPC), el deflactor del PIB o algún otro índice del nivel general de precios" [9]. La inflación conlleva la pérdida del poder adquisitivo de la moneda, generando la necesidad de una mayor cantidad de dinero con el tiempo para adquirir la misma cantidad de bienes y servicios. Según Mishkin [2], en el largo plazo, la teoría del efecto Fisher sostiene que las tasas de interés nominales son equivalentes a la tasa de interés real más la inflación esperada. Así, si las expectativas de inflación son al alza, las tasas de interés nominales deben aumentar para compensar la depreciación del valor del dinero. En el corto plazo, la inflación puede influir en las tasas de interés a través de la pérdida de poder adquisitivo del dinero y el incremento de la demanda de dinero. Este aumento en la demanda significa que individuos y empresas prefieren mantener más dinero en efectivo como medida de protección contra la pérdida de valor de la moneda, lo cual reduce la oferta de préstamos y ejerce presión al alza sobre las tasas de interés. Cuando la inflación es elevada, el Banco de la República suele elevar las tasas de interés con el objetivo de desincentivar el consumo y la inversión, contribuyendo así a controlar la inflación.

Por otro lado, el crecimiento económico de un país se refiere al incremento en la producción de bienes y servicios en una economía. El indicador principal utilizado para medir este crecimiento es el Producto Interno Bruto (PIB). Según Mauricio Cárdenas "mide el valor de los bienes y servicios finales que se producen dentro de los límites geográficos de un territorio en un período determinado de tiempo" [10]. La cuantificación del PIB involucra la multiplicación de los precios de cada producto final por su precio de mercado, y la suma total de esta producción se determina como el PIB nominal. El PIB nominal se ve afectado por los cambios en los precios, lo que significa que un aumento en los precios de los bienes finales también resulta en un crecimiento del PIB en términos de precios corrientes. Sin embargo, para evaluar de manera precisa el verdadero crecimiento de una economía en relación con la expansión de las cantidades producidas de esos bienes finales, se recurre al PIB real (ajustado por precios constantes). En cuanto a la relación entre tasas de interés y crecimiento económico, Alan [11] sostiene que, en líneas generales, la evidencia empírica sugiere que tasas de interés más bajas están asociadas a un mayor crecimiento económico. Esta asociación se explica por la capacidad de las tasas de interés reducidas para estimular la inversión y el consumo, factores que, a su vez, podrían desencadenar un aumento en la producción y el empleo. Sin embargo, es importante destacar que la conexión entre tasas de interés y crecimiento económico es compleja y depende de diversos elementos, como el nivel de desarrollo económico del país, la estructura de su economía y las políticas económicas implementadas por el gobierno.

En cuanto a las tasas de cambio, Banrep las caracteriza como el valor de una moneda en relación con otra, expresado como la cantidad de unidades de la moneda local necesarias para adquirir una unidad de la moneda extranjera [12]. Según Mankiw [7], las tasas de interés y el tipo de cambio suelen tener una relación inversa. Esto significa que, cuando las tasas de interés aumentan, el tipo de cambio tiende a depreciarse, y viceversa. Esta relación se explica por el hecho de que las tasas de interés son un factor que determina la rentabilidad de los activos financieros. Cuando las tasas de interés aumentan, los inversores tienen más incentivos para invertir en activos de renta fija, como bonos y letras del tesoro. Esto aumenta la demanda de moneda extranjera, lo que hace que el tipo de cambio se aprecie. Por el contrario, cuando las tasas de interés disminuyen, los inversores tienen menos incentivos para invertir en activos de renta fija. Esto reduce la demanda de moneda extranjera, lo que hace que el tipo de cambio se deprecie. Sin embargo, Mankiw también señala que esta relación no siempre se cumple. En algunos casos, las tasas de interés y el tipo de cambio pueden moverse en la misma dirección. Esto puede ocurrir, por ejemplo, si los inversores esperan que las tasas de interés aumenten en el futuro. En este caso, los inversores pueden comprar moneda extranjera con la expectativa de que su valor aumente cuando las tasas de interés realmente aumenten.

En lo que respecta al empleo, el empleo se define como el número de personas que desempeñan un trabajo remunerado, y su medición abarca diversas formas. Una de estas formas es la tasa de desempleo, la cual representa el porcentaje de la población en edad de trabajar que busca empleo, pero no lo encuentra. Asimismo, otra métrica relevante es la tasa de participación en la fuerza de trabajo, que indica el porcentaje de personas en edad de trabajar que están empleadas o activamente buscan oportunidades laborales. Estas medidas ofrecen perspectivas clave para evaluar el estado y la dinámica del mercado laboral. El empleo es un factor importante para la economía porque contribuye al crecimiento económico. Mankiw [13] sostiene que las tasas de interés y el empleo suelen tener una relación negativa. Esto se debe a que las tasas de interés más altas hacen que sea más caro para las empresas pedir prestado dinero, lo que reduce su inversión y contratación. Por el contrario, las tasas de interés más bajas hacen que sea más barato para las empresas pedir prestado dinero, lo que aumenta su inversión y contratación. Esta relación es importante para los bancos centrales, que pueden utilizarla como herramienta para influir en el empleo. Al aumentar las tasas de interés, los bancos centrales pueden tratar de reducir la inflación, pero a costa de reducir el empleo. Por el contrario, al reducir las tasas de interés, los bancos centrales pueden tratar de aumentar el empleo, pero a costa de aumentar la inflación.

Finalmente, las tasas de interés de un país extranjero ejercen influencia sobre las tasas de interés del país local. En el libro International Economics, Paul Krugman [8] explica que las tasas de interés de un país pueden tener un impacto significativo en las tasas de interés de otros países. Esto se debe a que las tasas de interés son un factor que determina el costo del crédito. Cuando las tasas de interés de un país son más altas que las tasas de interés de otros países, el costo del crédito en ese país es también más alto. Esto hace que sea más caro para las empresas y los consumidores pedir prestado dinero, lo que puede reducir la inversión y el consumo. Krugman identifica tres

mecanismos clave que explican la relación entre las tasas de interés de un país y las de otros. Primero, el mecanismo de los flujos de capital señala que tasas más altas atraen inversores internacionales, elevando la demanda de la moneda local y afectando la inflación y el déficit comercial. En segundo lugar, el mecanismo de los tipos de cambio esperados indica que la anticipación de aumentos futuros en las tasas de interés motiva la compra de la moneda del país en cuestión, generando resultados similares al mecanismo de flujos de capital. Finalmente, el mecanismo de integración financiera destaca la influencia de los mercados financieros internacionales, cuya creciente integración facilita la transferencia de capital entre países, resultando en cambios rápidos y significativos en las tasas de interés globales.

3.1.2. Ciencia de Datos

3.1.2.1. Aproximación económica de la Ciencia de Datos

En la aplicación de modelos de ciencia de datos a la teoría económica, se enfrentan desafíos particulares debido a la complejidad inherente de la economía como Ciencia. Los modelos de aprendizaje profundo se han convertido en herramientas muy eficaces en el campo de la macroeconomía, ofreciendo ventajas significativas en diversos aspectos. Dado que muchos fenómenos económicos presentan patrones complejos y no lineales, esta capacidad es esencial para una comprensión más precisa de la dinámica económica. En la tesis doctoral de Silva Guimaraes [14], se implementaron modelos de aprendizaje profundo para hacer frente a desafíos específicos en el campo macroeconómico. Se destacó especialmente el empleo de algoritmos de Deep Learning, conocidos por su habilidad para discernir patrones complejos a partir de datos, particularmente en el contexto de series temporales económicas. Estos modelos se centraron en la predicción de series temporales, con un enfoque específico en anticipar ciclos económicos y estimar brechas económicas. La elección de modelos de aprendizaje profundo se fundamenta en su capacidad para gestionar relaciones no lineales complejas presentes en los datos macroeconómicos. La tesis aborda la noción de transferencia de aprendizaje, una estrategia que consiste en aprovechar conocimientos adquiridos en un dominio para mejorar el rendimiento del modelo en otro dominio relacionado. Estos modelos tienen como objetivo principal clasificar de manera precisa las fases de los ciclos económicos y estimar rápidamente las brechas económicas. Estos objetivos revisten importancia crucial para comprender la salud general de una economía, identificar tendencias y formular políticas efectivas.

3.1.2.2. Machine Learning

Es un campo de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos que permiten a las computadoras aprender patrones y realizar tareas específicas sin ser programadas explícitamente. En lugar de seguir instrucciones específicas, los sistemas de machine learning utilizan datos para mejorar su rendimiento en una tarea particular a lo largo del tiempo. Este enfoque permite a las máquinas adaptarse y mejorar su desempeño sin intervención humana constante. Para modelos con enfoque de predecir la tasa de interés tenemos:

- **Ridge Regression:** aborda la problemática asociada a la alta variabilidad y la presencia de variables correlacionadas en modelos de regresión lineal. Según Hastie, Tibshirani y Friedman [15], el enfoque de regresión Ridge surge como una estrategia efectiva para mitigar la

indeterminación de coeficientes y la elevada varianza resultante de variables predictoras altamente correlacionadas. Al imponer una penalización sobre el tamaño de los coeficientes, controlada por el parámetro de complejidad λ, la regresión Ridge promueve la contracción de los coeficientes hacia cero, atenuando así la variabilidad y mejorando la estabilidad del modelo. Este método, descrito en detalle en el capítulo 3 del libro "The Elements of Statistical Learning," se presenta como una herramienta valiosa para prevenir la inestabilidad de los modelos de regresión lineal en el contexto de la predicción de tasas de interés. Asimismo, los autores subrayan la importancia de la estandarización de las variables predictoras antes de aplicar la regresión Ridge, dado que las soluciones no son invariantes ante cambios en la escala de las entradas. Esta conceptualización, respaldada por la formulación matricial del problema en la siguiente ecuación:

$$RSS(\lambda) = (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$$

Donde,

γ es el vector de respuestas o valores observados.

X Matriz de diseño que contiene variables predictoras. Cada fila es una observación y cada columna una variable predictora.

β Vector de coeficientes de regresión a estimar. Estos coeficientes multiplican las variables predictoras en la matriz X para predecir Y.

$RSS(\gamma)$ la suma de cuadrados de los residuos penalizada. Es decir, la diferencia entre valores observados ((γ) y predichos $(X\beta)$.

$(y - X\beta)^T(y - X\beta)$ término error cuadrático. Suma de los cuadrados de las diferencias entre valores observados y predichos.

$\lambda\beta^T\beta$ término de penalización Ridge. λ es el parámetro de complejidad que controla la penalización.

Ofrece una base teórica sólida para la incorporación de la regresión Ridge en el contexto de la predicción de tasas de interés y la gestión de la multicolinealidad en modelos econométricos.

- **Lasso Regression:** La esencia fundamental del modelo radica en la introducción de una penalización en la función objetivo del problema de regresión. Al hacerlo, Lasso no solo busca ajustar el modelo para lograr una buena predicción de los datos observados, sino que también incorpora una restricción que limita la magnitud absoluta de los coeficientes de regresión. En otras palabras, mientras ajusta el modelo, el Lasso simultáneamente realiza una selección automática de variables, siendo capaz de descartar algunas de ellas al asignar coeficientes exactamente a cero. Este enfoque único es particularmente valioso en situaciones donde se manejan conjuntos de datos con múltiples variables predictoras. Al penalizar la magnitud absoluta de los coeficientes, Lasso tiende a "apagar" algunas variables menos relevantes,

simplificando así el modelo final. La capacidad de realizar selección de variables de manera automática contribuye a la interpretabilidad del modelo y puede ser esencial en escenarios donde se busca identificar las características más significativas para la predicción. Según la perspectiva de Hastie, Tibshirani y Friedman [16], se define como la solución al siguiente problema de optimización:

$$\beta^{lasso} = argmin_{\beta} \sum_{i=1}^{N} (\gamma_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

Sujeto a la restricción,

$$\sum_{j=1}^{p} |\beta_1| \leq t$$

Esta formulación destaca la introducción de la penalización L1, representada por la restricción en la magnitud absoluta de los coeficientes (β_j).

Donde,

N es el número de observaciones.

ρ es el número de características.

γ_1 es la respuesta para la observación i.

x_{ij}es el valor de la característica j para la observación i.

β_0 es el intercepto.

β_j son los coeficientes de regresión estimados

- **Random Forest:** el concepto de bosques aleatorios, según Leo Breiman en su artículo "Random Forests" [17], se basa en la combinación de predictores de árboles de decisión. En este enfoque, cada árbol en el bosque depende de los valores de un vector aleatorio, muestreado de manera independiente y con la misma distribución para todos los árboles en el bosque. Breiman destaca que la selección aleatoria de características para dividir cada nodo contribuye a tasas de error comparativamente favorables en comparación con métodos como Adaboost, al tiempo que demuestra ser más robusta frente al ruido. El autor señala que el error de generalización de los bosques aleatorios tiende a converger a un límite a medida que se incrementa el número de árboles en el bosque. Este error, según Breiman, está influenciado por la fuerza de los árboles individuales y la correlación entre ellos. Además, el artículo aborda la utilidad de las estimaciones internas para supervisar el error, la fuerza y la correlación, destacando su papel en la evaluación y mejora del modelo. La aplicación de estas ideas a problemas de regresión también es explorada por Breiman, quien destaca la capacidad de los bosques aleatorios para proporcionar estimaciones internas valiosas.

- **Gradient Boost:** el modelo conceptualizado por Jerome H. Friedman en su conferencia de Reitz en 1999 [18], revoluciona la estimación de funciones al adoptar una perspectiva no paramétrica y centrarse en la optimización numérica en el espacio de funciones. En lugar de explorar el espacio de parámetros, el método se basa en expansiones aditivas por etapas y descenso de gradiente "boosting", ofreciendo robustez e interpretabilidad. Es especialmente eficaz para conjuntos de datos menos limpios, abordando la optimización en función del ajuste criterios variados como mínimos cuadrados y pérdida Huber-M. El enfoque no paramétrico se adapta a conjuntos de datos finitos, abordando la suavidad de la solución cuando la distribución conjunta de las variables es estimada.

- **Support Vector Machine:** son un tipo de algoritmo de aprendizaje automático diseñado para la clasificación y regresión de datos. La idea central de las SVM es encontrar un hiperplano en un espacio de características de alta dimensión que mejor separe las diferentes clases de datos, maximizando el margen entre ellas. Este margen es definido por los "vectores de soporte", que son los puntos más cercanos al hiperplano y son fundamentales para la construcción del clasificador. Dado un conjunto de datos de entrenamiento $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$, donde x_i es el vector de características y $y_1$es la etiqueta númerica correspondiente, la formulación del modelo SVM de regresión busca encontrar la función de regresión $f(x)$ que minimiza el error de regresión y, al mismo tiempo, mantiene el margen [19]. La fórmula de la función de regresión se expresa como

  $$f(x) = w * x + b$$

  donde,

  x es el vector de características de entrada.

  w es el vector pesos.

  b es el término de sesgo.

  La formulación del problema de optimización para SVM de regresión implica minimizar una función de costo que penaliza las desviaciones de las predicciones con respecto a las etiquetas reales y al mismo tiempo busca maximizar el margen. A menudo, se utilizan funciones de pérdida específicas para regresión, como la pérdida cuadrática. El problema de optimización se expresa como

  $$min_{w,b,\delta,\delta^*} \frac{1}{2}||w||^2 + C\sum_{i=1}^{N}(\delta_i + \delta_i^*)$$

  sujeto a,

  $$y_i - (w * x_i + b) \leq \varepsilon + \delta_i$$
  $$(w * x_i + b) - y_i \leq \varepsilon + \delta_i^*$$

$$C_i, C_i^* \geq 0$$

Donde,

C es el parámetro de regularización

ε es el margen de tolerancia para errores

$\delta_i \; y \; \delta_i^*$ son variables de holgura que permiten que algunos puntos estén dentro del margen o tengan errores

3.2. ANTECEDENTES

3.2.1. Efectos de las tasas de interés sobre el precio de las acciones en el mercado colombiano [23]

Este trabajo investiga la influencia de las tasas de interés del Banco de la República de Colombia en el mercado de renta variable durante un periodo de 5 años. La metodología emplea modelos de regresión lineal utilizando tanto Excel como Python, con técnicas de optimización como la validación cruzada. Se extiende el análisis a productos de renta fija y variable, evaluando modelos mediante back-testing para seleccionar el más efectivo. El enfoque aborda un vacío en investigaciones al considerar eventos como la pandemia del COVID-19. Se eligen dos activos específicos, las acciones de ISA y un bono del estado colombiano a 2 años, para determinar si la información del bono permite prever los precios de la acción. A lo largo de los 5 años, se observan 22 modificaciones en las tasas de intervención (tasa de interés) del Banco de la República, influenciadas por la recesión económica de 2020. A pesar de aplicar un modelo de regresión lineal simple, los resultados indican una débil incidencia de las tasas en los precios de acciones. Al ampliar el análisis al índice COLEQTY y optimizar parámetros con Python Cross Validation, la correlación entre las tasas de interés y el índice MSCI COLCAP no resulta significativa. El modelo muestra una capacidad de predicción del 79.2% en back-testing de dos años, pero se señalan limitaciones, como la restricción temporal y la consideración de una sola variable, sugiriendo futuras investigaciones con más variables para evaluar su impacto en el mercado de renta variable colombiano. Presenta un notable aporte para el proyecto aplicado al proporcionar un caso práctico asociado al uso de las tasas de interés como mecanismo de intervención del Banco de la República. Resalta por su competencia en el manejo de modelos de ciencia de datos y la realización de una evaluación de los modelos y variables escogidas. Su perspectiva integral se evidencia en la consideración de eventos inesperados, como la pandemia del COVID-19, agregando un valor significativo al trabajo.

3.2.2. Redes neuronales regularizadas para la predicción de la inflación en Colombia [24]

El proyecto se centra en desarrollar un modelo de machine learning para predecir la inflación colombiana, utilizando técnicas de regularización y siguiendo la metodología CRISP-DM. Comienza con la caracterización de diversas fuentes de datos relacionadas con la inflación y otras

variables económicas. Luego, implementa y valida un modelo de red neuronal regularizada, evaluándolo frente a un conjunto de prueba y comparándolo con un modelo de referencia ARIMA. Dado que la inflación y las tasas de interés se encuentran relacionadas, los resultados y enfoques metodológicos desarrollados en este proyecto podrían proporcionar insights valiosos y ser aplicables en la estimación de tasas de interés. La capacidad del modelo propuesto para anticipar cambios en la inflación podría tener implicaciones directas en la toma de decisiones financieras, incluida la gestión de tasas de interés. Al comprender y modelar de manera más precisa la dinámica de la inflación, el modelo podría contribuir a mejorar la precisión de las proyecciones relacionadas con las tasas de interés, facilitando así la toma de decisiones informadas. Adicionalmente, contribuye al proyecto aplicado en la aplicación de técnicas avanzadas de machine learning, específicamente redes neuronales regularizadas, para la predicción de la inflación y suministra un gran aporte al comparar modelos de ciencia de datos con modelos econométricos tradicionales.

3.2.3. Nowcasting de la tasa de desempleo de Colombia por medio de Machine Learning y pronóstico de la inflación y representación por medio de la curva de Phillips proyectada [25]
La investigación se enfoca en afrontar la demora en la publicación de cifras económicas, específicamente la tasa de desempleo en Colombia, utilizando un enfoque de NowCast (predicción inmediata) mediante modelos de Machine Learning. La metodología que propone implica la extracción en tiempo real de datos de Google Trends relacionados con la búsqueda de empleo en Colombia. Se implementan distintos modelos de Machine Learning, siendo XGBOOST el seleccionado por su mejor ajuste. Además, se proyecta la tasa de inflación y se explora la relación entre la tasa de desempleo y la inflación, representándola a través de una curva de Phillips clásica. Contribuye de forma significativa al proyecto aplicado dado que, la inflación y las tasas de interés están interrelacionadas, el modelo desarrollado en este proyecto para predecir la tasa de inflación puede proporcionar información valiosa para la estimación de tasas de interés.

3.2.4. Deep Learning para la predicción de tasas de interés en mercados emergentes [23]
Este trabajo implementa modelos de Deep Learning para predecir la tasa de interés en mercados emergentes, incluyendo Colombia. Utilizando una base de datos económica que abarca variables macroeconómicas como inflación, desempleo y tipo de cambio, el estudio evalúa la capacidad predictiva de redes neuronales profundas en comparación con modelos econométricos tradicionales como VAR y ARIMA. Los resultados muestran que los modelos de Deep Learning ofrecen una mayor precisión en la predicción de tasas de interés a largo plazo en comparación con los enfoques tradicionales [26]. Este estudio destaca la importancia de incorporar técnicas avanzadas de inteligencia artificial en mercados emergentes, donde la volatilidad y la interrelación de las variables económicas requieren modelos más complejos y robustos. La investigación respalda el enfoque del proyecto al mostrar cómo los modelos de Machine Learning y Deep Learning pueden mejorar significativamente la precisión en la predicción de tasas de interés en economías similares a la colombiana.

3.2.5. Predicción de tasas de interés mediante modelos híbridos de econometría y Machine Learning en el mercado europeo [24]

Este artículo se enfoca en la combinación de técnicas econométricas tradicionales con enfoques de Machine Learning para mejorar la predicción de tasas de interés en el mercado europeo. Los autores integran modelos de regresión lineal múltiple con técnicas de Machine Learning, como el Random Forest y el Gradient Boosting, para mejorar la precisión predictiva en comparación con los métodos puramente econométricos. Los resultados muestran que el enfoque híbrido logra una mayor precisión en escenarios donde las relaciones entre variables económicas no son lineales, como en momentos de alta volatilidad económica [27]. Este estudio ofrece un marco metodológico relevante para el proyecto, ya que resalta cómo la combinación de enfoques tradicionales y avanzados puede mejorar las predicciones de tasas de interés, lo que es aplicable en el contexto colombiano.

3.2.6. Impacto de las tasas de interés en la economía global y su predicción mediante redes neuronales artificiales [25]

Este estudio examina el impacto de las tasas de interés en diversas economías a nivel mundial y explora la predicción de las mismas utilizando redes neuronales artificiales. El enfoque se centra en la predicción de las tasas de interés a corto y mediano plazo utilizando datos de mercados globales, como el mercado de bonos de Estados Unidos, Europa y Asia. El estudio demuestra que las redes neuronales ofrecen ventajas en términos de precisión y capacidad de adaptación a cambios abruptos en los mercados financieros [28]. El impacto de las tasas de interés en el crecimiento económico y las decisiones de inversión es particularmente destacado, haciendo de este trabajo una contribución valiosa para investigaciones futuras en el ámbito de la política monetaria. Este estudio respalda el uso de modelos avanzados de Machine Learning en la predicción de tasas de interés, similar a lo que se busca en el presente proyecto.

4. IDENTIFICACIÓN Y ADQUISICIÓN DE FUENTES DE DATOS CON INDICADORES ECONÓMICOS

4.1 Introducción

La identificación y adquisición de fuentes de datos confiables es un paso fundamental para la construcción de modelos predictivos que permitan la estimación de la tasa de interés. En este capítulo, se detalla el proceso de recopilación y tratamiento de los datos necesarios para desarrollar los modelos propuestos en este proyecto. El objetivo específico que guía este capítulo es identificar y adquirir los indicadores económicos clave que influyen en la tasa de interés, asegurando la calidad, integridad y coherencia de los datos a lo largo del tiempo.

Para cumplir este objetivo, se llevaron a cabo tres actividades principales: la revisión de fuentes de datos relevantes, la selección de variables económicas de impacto en la tasa de interés y la adquisición y tratamiento de los datos de las fuentes identificadas. En las siguientes secciones, se detallan cada una de estas actividades y sus resultados.

4.2 Actividad 1: Revisión de Fuentes de Datos

4.2.1 Fuentes de Datos Identificadas

Para asegurar la precisión en la predicción de la tasa de interés, se identificaron y revisaron las siguientes fuentes de datos, consideradas como las más relevantes y confiables en el contexto económico colombiano:

- Banco de la República (BanRep)
- Departamento Administrativo Nacional de Estadística (DANE)
- Ministerio de Hacienda y Crédito Público
- Reserva Federal de Estados Unidos (FED)
- Investing

Estas fuentes ofrecen acceso a información macroeconómica y financiera de alta calidad, con cobertura temporal suficiente para realizar un análisis robusto. En particular, el Banco de la República (BanRep), el Departamento Administrativo Nacional de Estadística (DANE) y el Ministerio de Hacienda y Crédito Público, como entidades públicas del gobierno colombiano, son altamente confiables debido a su mandato institucional de generar y publicar datos oficiales de relevancia nacional. Estas instituciones están encargadas de monitorear y regular diversos aspectos económicos del país, garantizando la transparencia y exactitud de los datos que producen.

Además, la Reserva Federal de los Estados Unidos (FED), como el banco central de EE.UU., es una entidad pública internacionalmente reconocida por la calidad de sus datos y su rol en la política monetaria global, lo que refuerza la confiabilidad de los datos obtenidos de esta fuente.

Finalmente, Investing, aunque no es una entidad pública, es una plataforma reconocida a nivel mundial que proporciona información financiera actualizada, consolidada y precisa sobre una amplia gama de mercados e indicadores económicos. Su fiabilidad radica en el uso de datos provenientes de fuentes oficiales y su actualización constante, lo que la convierte en una fuente complementaria importante para el análisis.

4.2.2 Resultados de la Revisión

Como resultado de esta revisión, se logró compilar una lista de fuentes de datos relevantes y confiables para la obtención de los indicadores económicos necesarios para el análisis de la tasa de interés. Estas fuentes proporcionan información con diferentes frecuencias temporales (diaria, mensual, trimestral, y anual), lo que requiere un tratamiento específico para cada variable, según se detalla en las secciones siguientes.

4.3 Actividad 2: Selección de Variables Relevantes

4.3.1 Identificación de Variables

Tras la revisión de las fuentes de datos, se identificaron 35 indicadores económicos relevantes, los cuales se descomponen en 54 variables. Estas variables fueron seleccionadas con base en su respaldo teórico dentro de la literatura económica y su impacto en la tasa de interés, la variable objetivo del estudio. Las variables seleccionadas se agruparon en las siguientes áreas clave:

- **Política monetaria:** Incluye variables directamente relacionadas con la intervención del Banco de la República y otros bancos centrales, como la tasa de interés de política monetaria, las tasas OMA de expansión y contracción, y otras tasas de referencia como la tasa de interés interbancaria y la tasa de captación DTF.
- **Mercado de valores:** Comprende variables relacionadas con el comportamiento del mercado financiero, como el Índice COLCAP (bolsa de valores colombiana), los índices bursátiles internacionales (S&P 500, Dow Jones, Nasdaq, FTSE Colombia), y los precios de futuros de bienes como el oro, el petróleo Brent y WTI, y el café.
- **Inflación y precios:** Incluye variables como la inflación total, la meta de inflación, y el índice de precios al productor (IPP), tanto de la producción nacional como de la oferta interna.
- **Indicadores macroeconómicos generales:** Aquí se agrupan variables como el Producto Interno Bruto (PIB) real y nominal, la balanza comercial, el saldo de deuda interna y externa, y el balance fiscal.
- **Mercado laboral y salarios:** Incluye variables como la tasa de desempleo, la tasa de ocupación, y el salario mínimo mensual.
- **Sector financiero:** Comprende variables como el crédito doméstico neto (tanto del sector público como privado), el crédito bruto del sector financiero, y las reservas internacionales netas.

- **Tasas de cambio:** Se incluyen variables como la tasa representativa del mercado (TRM) y el índice de tasa de cambio real (IPP e IPC), que reflejan la relación del peso colombiano con otras monedas.

4.3.2 Tratamiento de los Datos

Cada variable fue tratada de manera específica según su frecuencia y disponibilidad. A continuación, se detalla el proceso de tratamiento aplicado a cada una de las variables seleccionadas:

Tabla 1. *Indicadores económicos*

#	Variable	Descripción	Frecuencia Original	Tratamiento Realizado
1	Tasa de intervención política monetaria	Variable objetivo, cambia 2 o 3 veces al año	Irregular (2-3 veces/año)	Rellenado mensual con el último dato disponible
2	Tasa OMA ventanilla de expansión (%)	Varios datos mensuales	Variable	Promedio mensual y rellenado con el último dato visto en meses sin publicación
3	Tasa OMA ventanilla de contracción (%)	Varios datos mensuales	Variable	Promedio mensual y rellenado con el último dato visto en meses sin publicación
4	Inflación total	Inflación al consumidor	Mensual	Sin modificación
5	Meta de inflación	Meta establecida por el Banco de la República	Mensual	Sin modificación
6	Índice de precios del productor (producción nacional)	Indicador de precios de la producción nacional	Mensual	Sin modificación
7	Índice de precios del productor (oferta interna)	Indicador de precios de la oferta interna	Mensual	Sin modificación
8	Precio de compra de oro	Precio diario de compra de oro	Diario	Promedio mensual

9	Precio de venta de oro	Precio diario de venta de oro	Diario	Promedio mensual
10	Oferta monetaria agregada (M1, M2, M3)	Suma de M1, M2 y M3	Mensual	Suma de las tres ofertas monetarias para obtener el indicador total
11	Índice COLCAP	Índice bursátil colombiano, con varios datos mensuales	Variable (20-21 veces/mes)	Promedio mensual
12	Reservas internacionales netas	Reservas en el Banco de la República	Mensual	Sin modificación
13	Tasa representativa del mercado (TRM)	Valor del dólar frente al peso colombiano	Diario	Promedio mensual
14	Tasa de desempleo	Tasa de desempleo en Colombia	Mensual	Sin modificación
15	Tasa de ocupación	Porcentaje de la población empleada	Mensual	Sin modificación
16	Salario mínimo mensual	Salario mínimo legal en Colombia	Anual	Rellenado mensual con el mismo valor para todo el año
17	Producto interno bruto real (PIB real)	Valor ajustado por inflación del PIB	Trimestral	Mensualización dividiendo cada trimestre entre 3
18	Producto interno bruto nominal (PIB nominal)	Valor no ajustado por inflación del PIB	Trimestral	Mensualización dividiendo cada trimestre entre 3
19	Balanza comercial	Diferencia entre exportaciones e importaciones	Mensual	Sin modificación
20	Balance fiscal	Balance entre ingresos y gastos del gobierno	Mensual	Sin modificación
21	Saldo deuda interna	Deuda interna del gobierno	Trimestral	Mensualización dividiendo cada trimestre entre 3

22	Saldo deuda externa	Deuda externa del gobierno	Trimestral	Mensualización dividiendo cada trimestre entre 3
23	Ingresos Presupuesto General de la Nación	Ingresos totales del presupuesto público	Anual	Mensualización dividiendo el dato anual entre 12
24	Gasto social	Gasto del gobierno en programas sociales	Anual	Mensualización dividiendo el dato anual entre 12
25	Tasa de interés Prime Rate	Tasa de interés en Estados Unidos	Mensual	Sin modificación
26	Tasa de interés interbancaria	Tasa de préstamos entre bancos	Diario	Promedio mensual
27	Tasa de captación depósitos a término fijo (DTF)	Tasa de interés para depósitos a término fijo	Mensual	Sin modificación
28	Tasa de interés de colocación	Tasa de interés que se cobra por los préstamos	Mensual	Sin modificación
29	Reserva monetaria	Cantidad de dinero en reservas	Mensual	Sin modificación
30	Crédito doméstico neto (sector público)	Cantidad total de crédito otorgado al sector público	Mensual	Sin modificación
31	Crédito doméstico neto (sector privado)	Cantidad total de crédito otorgado al sector privado	Mensual	Sin modificación
32	Índice de tasa de cambio real (IPP e IPC, nacional e internacional)	Indicadores que reflejan el valor real del peso frente a otras monedas	Mensual	Sin modificación
33	Activos netos exteriores	Cantidad de activos externos netos en poder del país	Mensual	Sin modificación
34	Futuros de petróleo Brent	Precios de futuros del petróleo Brent	Mensual	Se tomó el valor de cierre

35	Futuros de oro	Precios de futuros del oro	Mensual	Se tomó el valor de cierre
36	Futuros de petróleo WTI	Precios de futuros del petróleo WTI	Mensual	Se tomó el valor de cierre
37	Futuros del café C	Precios de futuros del café C	Mensual	Se tomó el valor de cierre
38	Índice S&P 500	Índice bursátil de Estados Unidos	Mensual	Se tomó el valor de cierre
39	Índice Dow Jones	Índice bursátil Dow Jones	Mensual	Se tomó el valor de cierre
40	Índice Nasdaq	Índice bursátil Nasdaq	Mensual	Se tomó el valor de cierre
41	Índice FTSE Colombia	Índice bursátil FTSE Colombia	Mensual	Se tomó el valor de cierre
42	Tasa de interés de la Reserva Federal (FED)	Tasa de interés de la Reserva Federal (EE.UU.)	Variable	Rellenado mensual con el último dato visto

4.3.3 Resultados de la Selección de Variables

Como resultado de este proceso, se seleccionaron las variables más relevantes basadas en su respaldo teórico económico y su impacto en las decisiones de la tasa de interés, la variable objetivo. El conjunto de variables seleccionadas constituye la base para la creación de los modelos predictivos que se desarrollarán en los capítulos siguientes.

4.4 Actividad 3: Adquisición de Datos

4.4.1 Descarga y Procesamiento de Datos

Una vez seleccionadas las fuentes de datos y las variables relevantes, se procedió a la adquisición de los datos. Para ello, se descargaron los datos de las plataformas identificadas, asegurando que la temporalidad fuera coherente (mensual en su mayoría) y que la calidad de la información fuera adecuada para el análisis. Se aplicaron técnicas de limpieza y tratamiento para asegurar que no hubiera datos faltantes o inconsistencias que pudieran afectar los resultados.

4.4.2 Resultados de la Adquisición de Datos

Se completó la adquisición de un conjunto de datos robusto, alineado a nivel mensual. Este conjunto de datos será utilizado en los próximos capítulos para el desarrollo de los modelos predictivos que permitirán estimar las tasas de interés. Se definió el período de análisis comprendido entre enero de 2008 y diciembre de 2023, seleccionándolo estratégicamente para

garantizar la disponibilidad de datos completos y consistentes para todas las variables consideradas desde su fecha de inicio.

4.5 Conclusión del Capítulo

En este capítulo se desarrolló el proceso de identificación y adquisición de fuentes de datos e indicadores económicos clave para la estimación de las tasas de interés. La revisión de fuentes confiables, la selección de variables y el adecuado tratamiento y adquisición de los datos aseguran la solidez de la base de datos que alimentará los modelos predictivos en los siguientes capítulos. Este proceso es fundamental para garantizar la calidad y precisión de las predicciones futuras, proporcionando así una plataforma sólida para el análisis posterior.

5. APLICACIÓN DE TÉCNICAS DE PREPROCESAMIENTO DE DATOS

5.1 Introducción

El preprocesamiento de datos es un paso crucial en la construcción de modelos de Machine Learning, ya que garantiza la calidad, consistencia y coherencia de la información utilizada. Sin un preprocesamiento adecuado, los modelos pueden generar resultados inexactos o poco confiables debido a la presencia de datos faltantes, valores atípicos o incoherencias en las variables. El objetivo de este capítulo es aplicar técnicas de preprocesamiento para asegurar que los datos utilizados en los modelos predictivos sean de alta calidad y estén alineados temporalmente.

El proceso de preprocesamiento incluyó tres actividades principales: limpieza de datos, estandarización de las variables, y ajuste temporal de las series de datos. A continuación, se describen las actividades realizadas y sus resultados.

5.2 Actividad 1: Limpieza de Datos

5.2.1 Identificación y Manejo de Valores Atípicos

Durante el análisis exploratorio de los datos, se aplicó el modelo de Isolation Forest para detectar y eliminar valores atípicos. Este modelo no supervisado identificó observaciones anómalas clasificándolas como normales (1) o atípicas (-1). Posteriormente, se creó una serie booleana para filtrar estas observaciones y eliminarlas del conjunto de datos. Sin embargo, se decidió no excluir estos valores, ya que representan comportamientos normales en los datos económicos. La eliminación de estos valores podría haber eliminado información valiosa de las fluctuaciones económicas.

5.2.2 Resultados de la Limpieza de Datos

A través del proceso de detección de valores atípicos utilizando Isolation Forest, se identificaron varias observaciones que se desviaban significativamente del comportamiento esperado. Sin embargo, se decidió no eliminarlas, ya que su impacto en el rendimiento predictivo de los modelos fue considerado importante. Esta decisión permitió mantener la integridad del conjunto de datos y preservar la representación completa de los patrones económicos. Como resultado, el conjunto de datos quedó preparado de forma para su uso en los análisis y modelos posteriores.

5.3 Actividad 2: Estandarización de los Datos

5.3.1 Técnicas Aplicadas

La estandarización es una técnica fundamental en el preprocesamiento de datos para modelos que son sensibles a la escala de las características. Sin embargo, no se aplicó de manera uniforme

a todos los modelos, ya que algunos no dependen de la magnitud de las variables. A continuación, se describe su aplicación específica:

- **Estandarización (Standard Scaling):** Se aplicó a los modelos Ridge, Lasso y Support Vector Machine Regressor (SVR) debido a su sensibilidad a la escala de las características. Esta técnica ajusta las variables para que tengan una media de 0 y una desviación estándar de 1. Es crucial para garantizar la estabilidad del modelo y la correcta interpretación de los coeficientes.

  Justificación: La estandarización mejora la estabilidad numérica de los modelos lineales y basados en márgenes, ayudando a que los algoritmos de optimización converjan de manera más eficiente. Además, asegura que las regularizaciones L1 y L2 penalicen todas las características de forma uniforme, lo que previene el sobreajuste y optimiza el rendimiento.

- **Random Forest y Gradient Boosting (XGBoost):** No se aplicó estandarización a los datos en el caso de Random Forest y XGBoost, ya que estos algoritmos no son sensibles a la escala de las características. Su funcionamiento se basa en comparaciones relativas en los puntos de corte al particionar el espacio de características, lo que elimina la necesidad de transformar los datos.

  Justificación: Ambos modelos manejan eficientemente datos con diferentes escalas, ya que los árboles de decisión dentro de estos algoritmos no dependen de la magnitud de las variables. Este enfoque los hace ideales para conjuntos de datos heterogéneos sin necesidad de preprocesamiento adicional.

5.3.2 Resultados de la Estandarización

La estandarización fue aplicada a los modelos Ridge, Lasso y Support Vector Machine Regressor (SVR), lo que permitió estabilizar los coeficientes y mejorar el rendimiento del modelo. Por otro lado, el modelo Random Forest y XGBoost no requirió ninguna transformación, manteniendo la estructura original de los datos sin afectar su rendimiento. Este enfoque diferencial permitió que cada modelo se ajustara de manera óptima a las características del conjunto de datos.

5.4. Actividad 3: Manejo de Datos Temporales

5.4.1 Creación de Rezagos sin Estandarización

En esta etapa, se aplicó un manejo de datos temporales mediante la creación de rezagos de las variables predictoras, lo que permite capturar la dependencia temporal en los modelos. Este paso se realizó sin estandarizar los datos, utilizando las siguientes estrategias:

- **Creación de Rezagos:** Se generaron tres rezagos para cada una de las variables predictoras, excluyendo la fecha y la variable objetivo ($Tasa_intervencion_politica_monetaria$). Los

rezagos permiten que el modelo incorpore información histórica de las variables como insumo para las predicciones.

- **Eliminación de Valores Faltantes:** La creación de rezagos genera valores faltantes en las primeras filas del conjunto de datos. Para garantizar la integridad del conjunto, se eliminaron las filas con valores NaN.

5.4.2. Creación de Rezagos con Estandarización

Posteriormente, se repitió el manejo de datos temporales sobre el conjunto de datos previamente estandarizado, utilizando los mismos pasos:

- **Estandarización Previa:** Antes de la creación de rezagos, los datos predictivos fueron estandarizados para garantizar que todas las variables estuvieran en una escala comparable, lo que beneficia a los modelos que son sensibles a la magnitud de las variables.
- **Creación de Rezagos:** Similar a la sección anterior, se generaron tres rezagos para las variables predictoras. Este proceso proporciona información temporal adicional al modelo.
- **Eliminación de Valores Faltantes:** Se eliminaron las filas con valores NaN generados por los rezagos para asegurar la coherencia del conjunto de datos final.

5.4.2 Resultados del Manejo de Datos Temporales

Gracias a la implementación de rezagos, se obtuvieron dos conjuntos de datos con diferentes enfoques: uno sin estandarizar y otro estandarizado. Ambos están completamente alineados temporalmente y listos para ser utilizados en los modelos predictivos.

5.5 Conclusión del Capítulo

Este capítulo detalló el proceso de preprocesamiento de datos, que incluyó la limpieza de los datos y estandarización de las variables, así como el ajuste temporal de las series de datos. Si bien se identificaron outliers, no fueron eliminados, ya que no afectaban significativamente el rendimiento de los modelos. Como resultado, se obtuvo un conjunto de datos robusto y alineado temporalmente mediante rezagos (data estandarizada y sin estandarizar).

6. EVALUACIÓN Y SELECCIÓN DE MODELOS DE MACHINE LEARNING PARA LA PREDICCIÓN DE TASAS DE INTERÉS

6.1. Introducción

La predicción de las tasas de interés constituye un desafío analítico significativo en el ámbito económico debido a la complejidad de las relaciones macroeconómicas. Este capítulo abordó el Objetivo Específico 3, enfocado en evaluar y seleccionar modelos de Machine Learning apropiados para la predicción de tasas de interés de política monetaria. Se sigue una metodología estructurada, que abarca desde la implementación preliminar de modelos hasta su comparación y ajuste final, asegurando la solidez de los resultados obtenidos.

El desarrollo de este capítulo se organiza en tres actividades principales:

- **Implementación inicial y ajuste preliminar de cinco modelos de Machine Learning:** Ridge Regression, Lasso Regression, Random Forest, Gradient Boosting con XGBoost y Support Vector Regression (SVR). Estos modelos fueron seleccionados debido a sus capacidades específicas para abordar problemas inherentes a la predicción económica, como la multicolinealidad, la selección automática de variables relevantes y la modelización de relaciones no lineales y lineales. En esta etapa inicial, los modelos Ridge, Lasso y SVR utilizarán datos previamente estandarizados, mientras que Random Forest y XGBoost operarán sobre los datos en su formato original.
- **Evaluación comparativa de los modelos:** Cada modelo se evaluó utilizando un conjunto de métricas que incluyen el coeficiente de determinación (R^2), el error cuadrático medio (MSE), el error absoluto medio (MAE) y el error porcentual absoluto medio (MAPE). Estas métricas ofrecen una visión integral del desempeño predictivo de los modelos, abarcando tanto la precisión en las predicciones como el manejo de errores relativos. La validación cruzada se implementará mediante Grid Search, lo que permitirá ajustar de manera sistemática los hiperparámetros de cada modelo para optimizar su desempeño.
- **Entrenamiento final de los modelos seleccionados:** Los modelos se entrenaron de manera iterativa utilizando técnicas avanzadas como la validación cruzada con Grid Search. Este proceso permitirá refinar los parámetros de los modelos seleccionados y asegurar su capacidad de generalización en diferentes conjuntos de datos, reduciendo el riesgo de sobreajuste.

El enfoque metodológico adoptado para este capítulo buscó establecer una línea base sólida, evitando inicialmente el uso de técnicas de selección de características o análisis de correlación, que serán abordadas en capítulos posteriores. Este enfoque asegura que los resultados obtenidos en esta etapa sean representativos del desempeño bruto de los modelos, permitiendo un análisis más claro del impacto de las mejoras progresivas.

La selección de los modelos está respaldada por sus fundamentos teóricos y capacidades prácticas. El modelo Ridge Regression, al emplear regularización L_2, es particularmente útil para manejar datos con multicolinealidad, mientras que la Lasso Regression, mediante regularización L_1, facilita la selección automática de características relevantes al reducir a cero los coeficientes de las variables menos significativas. Los modelos basados en árboles, como Random Forest y XGBoost, destacan por su capacidad para capturar relaciones no lineales complejas y manejar grandes volúmenes de datos con alta dimensionalidad. Por último, Support Vector Regression (SVR) ofrece un enfoque robusto para la predicción de relaciones no lineales mediante el uso de funciones kernel, siendo especialmente útil en escenarios con datos no lineales y complejos.

El uso de Grid Search en la validación cruzada reforzó la rigurosidad de esta evaluación, asegurando que los modelos implementados operen bajo configuraciones óptimas de hiperparámetros. Este método sistemático no solo mejora la capacidad predictiva de los modelos, sino que también garantiza la comparabilidad justa entre ellos. En definitiva, este capítulo se alinea directamente con el Objetivo Específico 3 de la investigación, proporcionando una evaluación inicial y una selección fundamentada de los modelos de Machine Learning más prometedores para la predicción de tasas de interés. Las actividades descritas no solo establecen un marco sólido para los análisis avanzados en capítulos posteriores, sino que también aseguran una metodología coherente y transparente en el desarrollo de la investigación.

6.2. Descripción

Preparación de los datos

La preparación de datos se realizó de manera uniforme para garantizar la consistencia en el análisis de todos los modelos:

- Definición de variables: X se definió como el conjunto de características independientes, mientras que Y corresponde a la variable objetivo, la Tasa de Intervención Política Monetaria.
- División de datos: Los datos se dividieron en un conjunto de entrenamiento (80%) y un conjunto de prueba (20%) utilizando la función $train_test_split$. El conjunto de entrenamiento fue utilizado para ajustar los modelos, mientras que el conjunto de prueba se empleó para evaluar su desempeño en datos no vistos.

Revisión de multicolinealidad

Se evaluó la presencia de multicolinealidad mediante el cálculo del $Variance\ Inflation\ Factor$ (VIF) para todas las características independientes en X. Este análisis permitió identificar redundancias entre las variables:

- Modelos sensibles a multicolinealidad: En Ridge y Lasso, este análisis fue crucial, ya que una alta correlación entre variables justifica la aplicación de técnicas de regularización. Mientras

que Ridge mitiga estos efectos mediante penalización L2, Lasso no solo los reduce, sino que también realiza selección automática de características, asignando coeficientes nulos a las variables menos relevantes.

- Modelos robustos a multicolinealidad: Aunque Random Forest, XGBoost y SVR no son directamente afectados por la multicolinealidad debido a su naturaleza no lineal o basada en márgenes, este análisis proporcionó información útil para futuras interpretaciones y posibles reducciones de dimensionalidad.

Entrenamiento inicial

Cada modelo se entrenó inicialmente con hiperparámetros predeterminados para establecer una línea base de desempeño:

- Ridge y Lasso: Ambos utilizaron un valor inicial del parámetro de regularización $\lambda = 1.0$ (representado como *alpha*).
- Random Forest: Se emplearon valores predeterminados como $n_estimators = 100$ (número de árboles) y profundidad ilimitada.
- XGBoost: Hiperparámetros iniciales incluyeron $learning_rate = 0.1$ (tasa de aprendizaje) y $max\ _depth\ = 6$ (profundidad máxima de los árboles).
- SVR: Se configuró con $C = 1.0$ (penalización), $\varepsilon = 0.1$ (tolerancia) y kernel radial (rbf).

Se evaluaron las siguientes métricas para cada modelo en el conjunto de prueba:

- Errores: MSE (Mean Squared Error), MAE (Mean Absolute Error) y MAPE (Mean Absolute Percentage Error).
- Capacidad explicativa: R^2.

Optimización hiperparámetros

Para mejorar el desempeño de los modelos, se implementó una búsqueda sistemática de los mejores valores de hiperparámetros utilizando *Grid Search* con validación cruzada (5 *folds*):

- Ridge y Lasso: Se optimizó el parámetro λ (*alpha*) en un rango de valores $[0.01, 0.1, 1, 10, 100]$.
- Random Forest: Se ajustaron $n_estimators$ (número de árboles), $\max\ _depth$ (profundidad máxima de los árboles) y $\min\ _samples_split$ (número mínimo de muestras para dividir un nodo).
- XGBoost: Se exploraron combinaciones de *learning_rate* (control de la contribución de cada árbol), $n_estimators$ y $\max\ _depth$.
- SVR: Se optimizaron C (parámetro de penalización), ε (tolerancia de error) y el tipo de kernel (lineal o radial).

La combinación óptima de hiperparámetros fue seleccionada en función de la maximización del R^2, asegurando un equilibrio entre el ajuste del modelo y su capacidad de generalización.

Validación cruzada con K-Fold

- Se utiliza K-Fold Cross Validation ($k = 5$) para evaluar la generalización del modelo ajustado con el mejor λ.
- Las métricas evalaudas incluyen MSE, MAE y R^2, reportadas para cada fold y promediadas para obtener la medida global.

Curva de aprendizaje

- Se genera una curva de aprendizaje que muestra cómo varía el desempeño del modelo (en términos de R^2) con diferentes tamaños del conjunto de entrenamiento.
- Se analizan Score de entrenamiento (evalúa el ajuste del modelo a los datos de entrenamiento) y Score de validación (indica la capacidad del modelo para generalizar a datos no vistos)

Este análisis permite determinar si el modelo está sufriendo de sesgo (bajo desempeño tanto en entrenamiento como en validación) o varianza (alto desempeño en entrenamiento, pero bajo en validación).

Resultados y análisis

Los resultados de cada modelo se agruparon en tres etapas clave:

- Evaluación inicial: Desempeño con hiperparámetros predeterminados.
- Optimización: Desempeño tras ajustar los hiperparámetros mediante Grid Search.
- Validación cruzada: Promedios de métricas calculadas en los 5 folds.

Se elaboró una tabla comparativa que incluye las métricas obtenidas (MSE, MAE, MAPE, R^2) en cada etapa. Este enfoque permitió analizar de forma integral cómo la optimización y validación mejoraron el desempeño y la capacidad de generalización de cada modelo.

6.3. Modelo Ridge

Revisión de multicolinealidad

El análisis de multicolinealidad mediante el $Factor\ de\ Inflación\ de\ la\ Varianza$ (VIF) reveló valores extremadamente altos (infinito) para todas las variables en el conjunto de datos. Este resultado fue indicativo de una alta correlación entre las características, lo cual generó problemas de multicolinealidad. La presencia de multicolinealidad podía llevar a que los coeficientes de la regresión lineal ordinaria sean inestables y no interpretables. En este contexto, el uso de Ridge fue teóricamente apropiado, ya que su término de regularización L^2 ayudó a mitigar los efectos

negativos de la multicolinealidad al penalizar las magnitudes de los coeficientes de las características.

```
=== Revisión de Multicolinealidad (VIF) ===
                        feature  VIF
0      ventanilla_expansion_(%)  inf
1    ventanilla_contraccion_(%)  inf
2               inflacion_total  inf
3                meta_inflacion  inf
4        IPP_produccion_nacional  inf
..                           ...  ...
435      s&p_500_lag1_lag2_lag3  inf
436    dow_jones_lag1_lag2_lag3  inf
437       nasdaq_lag1_lag2_lag3  inf
438     ftse_col_lag1_lag2_lag3  inf
439     EFFR_FED_lag1_lag2_lag3  inf
```

***Ilustración 1**. Resultados multicolinealidad Ridge*

Evaluación inicial del modelo

La primera implementación de Ridge, con un valor predeterminado del parámetro $\alpha = 1.0$, mostró un excelente desempeño predictivo en términos de las métricas evaluadas:

- MSE: 0.0386
- MAE: 0.1326
- MAPE: 3.11%
- $R^2\ Score$: 0.9943

El R^2obtenido (0.9943) indica que el modelo es capaz de explicar el 99.43% de la varianza de la variable objetivo en el conjunto de prueba, lo que evidencia un ajuste casi perfecto. Adicionalmente, el bajo MSE y MAE reflejan que las predicciones del modelo están muy cerca de los valores reales. Esto sugiere que el modelo incluso sin una optimización avanzada es altamente efectivo. Sin embargo, no es confiable dado la dimensionalidad producto de los rezagos y el resultado VIF que indica problemas de multicolinealidad.

Optimización con Grid Search

Se implementó una búsqueda de hiperparámetros mediante Grid Search para optmizar el valor del parámetro α, utilizando un rango de valores [0.01, 0.1, 1, 10, 100] y una validación cruzada con 5 particiones. El mejor valor encontrado fue $\alpha = 1$, confirmando que el modelo inicial ya estaba configurado de manera óptima.

Los resultados después de la optimización confirmaron que no hubo mejoras adicionales en las métricas:

- MSE: 0.0386

- MAE: 0.1326
- MAPE: 3.11%
- R^2Score: 0.9943

El modelo es robusto con respecto al parámetro α, dado que valores distintos no alteraron el desempeño predictivo.

Validación cruzada (K-Fold)

Se realizó una validación cruzada con K-Fold ($k = 5$) utilizando el mejor modelo encontrado ($\alpha = 1$). Los resultados promedio para cada métrica son los siguientes:

- El $R^2Score\ promedio$: 0.9953 (varianza mínima entre los folds, lo que indica un modelo estable).
- MSE promedio: 0.0340
- MAE promedio: 0.1350

El desempeño consistente a lo largo de los folds indica que el modelo tiene una excelente capacidad de generalización, lo que refuerza su aplicabilidad en escenarios no vistos.

Curva de aprendizaje

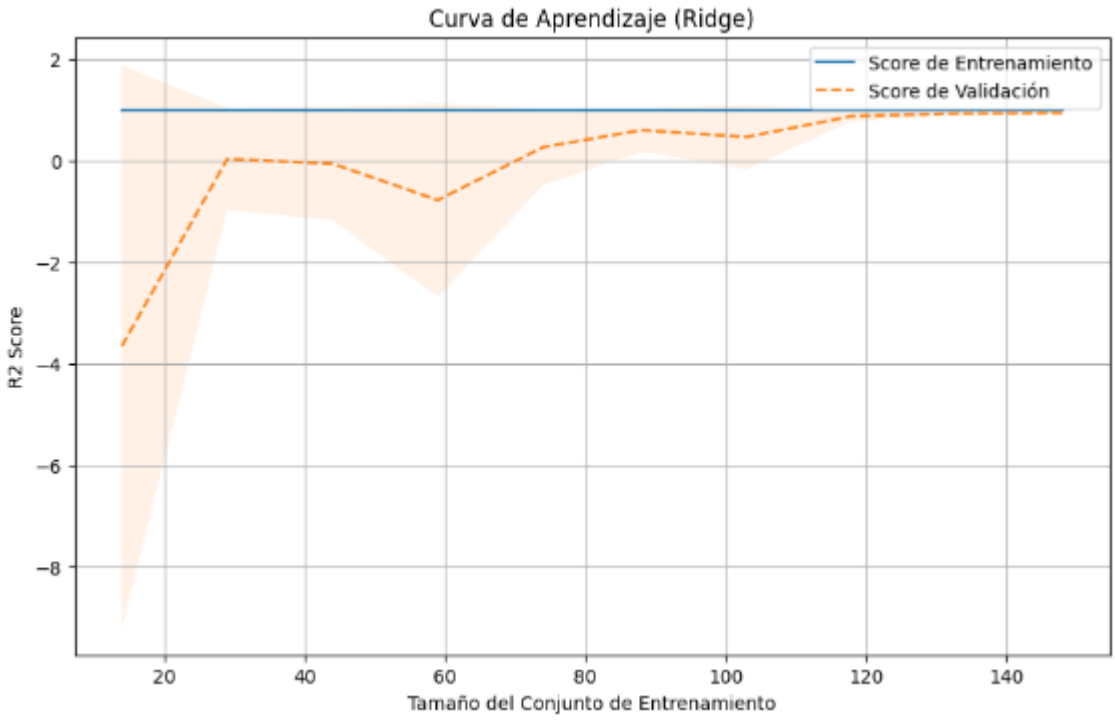

Ilustración 2. *Curva aprendizaje Ridge*

La curva de aprendizaje muestra la relación entre el tamaño del conjunto de entrenamiento y el desempeño del modelo en términos de R^2:

- El score de entrenamiento permanece constante y cercano a 1, lo que indica que el modelo está ajustando perfectamente los datos de entrenamiento.
- El score de validación comienza más bajo, pero se aproxima rápidamente al score de entrenamiento a medida que el tamaño del conjunto de entrenamiento aumenta, alcanzando una estabilidad. Esto sugiere que el modelo no está sufriendo de sobreajuste ni de subajuste.
- El comportamiento observado en la curva confirma que el modelo Ridge es capaz de beneficiarse de un mayor volumen de datos de entrenamiento, pero incluso con un tamaño moderado de muestra, ofrece un desempeño altamente competitivo.

Análisis comparativo

Se presenta un gráfico que resume el desempeño del modelo en las tres etapas clave:

```
=== Tabla de Resultados Comparativos ===
                              Etapa       MSE       MAE      MAPE  R2 Score
0                 Evaluación Inicial  0.038591  0.132614  0.031053   0.99425
1              Después de Grid Search  0.038591  0.132614  0.031053   0.99425
2  Evaluación Final (Cross-Validation)  0.038591  0.132614  0.031053   0.99425
```

Ilustración 3. *Resultados comparativos Ridge inicial*

6.3.1. Conclusiones

El análisis de las métricas y del comportamiento del modelo Ridge indica que, aunque el modelo parece no estar sobreajustado en términos de sus métricas (R^2, MSE, MAE y MAPE) y su comportamiento estable en la validación cruzada, existen riesgos asociados al alto número de variables (442) debido a los rezagos en relación al tamaño de la muestra (186 filas) que deben ser considerados:

Evaluación de sobreajuste basado en métricas

- Resultados del modelo Ridge: El R^2 extremadamente alto (>0.99) y el bajo MSE y MAE son indicadores de un desempeño muy preciso, tanto en los datos de entrenamiento como en los de validación. Esto podría interpretarse como una señal de sobreajuste, especialmente considerando que aún no se ha aplicado ningún análisis de correlación ni selección de características.
- Validación cruzada con K-Fold: La estabilidad de las métricas (R^2promedio de 0.9953 y varianza mínima entre los folds) sugiere que el modelo generaliza bien entre particiones. Sin embargo, la validación cruzada no garantiza completamente que el modelo esté libre de sobreajuste, especialmente en conjuntos de datos con características redundantes o correlacionadas.

Impacto del alto VIF en los datos

El análisis del $Factor\ de\ Inflación\ de\ la\ Varianza$ (VIF) mostró valores infinitos para todas las variables, lo que confirma un alta multicolinealidad en el conjunto de datos. Aunque Ridge mitiga parcialmente los efectos negativos de la multicolinealidad mediante la penalización L_2, esto no elimina por completo las redundancias entre las variables. En este escenario:

- Las 442 variables incluyen redundancias significativas, lo que aumenta la complejidad del modelo sin aportar información adicional útil.
- El alto número de variables en relación al tamaño de la muestra (186 filas) significa que el modelo está ajustándose a muchos parámetros con relativamente pocos datos, lo que puede llevar a un ajuste superficial de las relaciones en los datos sin garantizar su robustez.

Implicaciones de no aplicar técnicas de selección de características

La ausencia de análisis de correlación o técnicas de selección de características en esta etapa inicial implica que:

- El modelo está procesando todas las características, incluidas aquellas que probablemente no tienen relevancia significativa para predecir la variable objetivo.
- Características irrelevantes o redundantes pueden estar influyendo marginalmente en las predicciones, creando la ilusión de un desempeño perfecto (alto R^2).

El riesgo no es tanto el sobreajuste, sino el hecho de que el modelo puede estar utilizando patrones espurios derivados de características redundantes o no informativas.

6.4. Modelo Lasso

Revisión de multicolinealidad

```
=== Revisión de Multicolinealidad (VIF) ===
                        feature  VIF
0      ventanilla_expansion_(%)  inf
1    ventanilla_contraccion_(%)  inf
2                inflacion_total  inf
3                 meta_inflacion  inf
4       IPP_produccion_nacional  inf
..                          ...  ...
435       s&p_500_lag1_lag2_lag3  inf
436     dow_jones_lag1_lag2_lag3  inf
437        nasdaq_lag1_lag2_lag3  inf
438      ftse_col_lag1_lag2_lag3  inf
439      EFFR_FED_lag1_lag2_lag3  inf
```

Ilustración 4. *Resultados multicolinealidad Lasso*

El análisis de $Variance\ Inflation\ Factor$ (VIF) reveló valores extremadamente altos (infinito) para todas las características del conjunto de datos. Esto indica una presencia significativa de multicolinealidad, donde las variables independientes están altamente correlacionadas entre sí. Este escenario puede dificultar la estimación precisa de los coeficientes en métodos de regresión ordinaria. En este contexto, el modelo Lasso es particularmente apropiado, ya que no solo mitiga los efectos de la multicolinealidad mediante regularización L_1, sino que también realiza una selección automática de características, asignando un coeficiente igual a cero a aquellas variables menos relevantes.

Resultados iniciales del modelo Lasso

En la etapa inicial, se utilizó un valor predeterminado de $\lambda = 1.0$ (parámetro alpha), y el modelo fue evaluado mediante las siguientes métricas:

- MSE: 0.8226
- MAE: 0.7077
- MAPE: 18.38%
- R^2: 0.8774

Los resultados iniciales muestran un ajuste deficiente, con un R^2 relativamente bajo (87.74%) y valores elevados de MSE y MAE. Además, el mensaje de advertencia de convergencia sugiere que el modelo no pudo ajustarse adecuadamente con el número predeterminado de iteraciones, probablemente debido a la escala de las características y la alta dimensionalidad del conjunto de datos.

Optimización con Grid Search

Se implementó una búsqueda de hiperparámetros mediante Grid Search para identificar el mejor valor del parámetro λ (alpha) en un rango especifico [0.01, 0.1, 1, 10, 100]. El mejor valor encontrado fue $\lambda = 0.01$, lo que indica que un menor nivel de regularización mejora el ajuste del modelo.

```
=== Resultados Después de Grid Search (Ridge) ===
Mejor Alpha: 1
Mean Squared Error (MSE): 0.03859075741031517
Mean Absolute Error (MAE): 0.13261394586953895
Mean Absolute Percentage Error (MAPE): 0.031053488889220226
R2 Score: 0.9942503411889901
```

***Ilustración 5**. Resultados optimización grid search Lasso*

Tras optimizar el modelo, se obtuvo un desempeño significativamente mejorado:

- MSE: 0.0275
- MAE: 0.0991

- MAPE: 1.94%
- R^2: 0.9959

El R^2 aumentó de 87.74% a 99.59%, reflejando una mejora sustancial en la capacidad explicativa del modelo. Las métricas de error (MSE, MAE, MAPE) también disminuyeron considerablemente, indicando que el modelo optimizado predice con mayor precisión la variable objetivo.

Validación cruzada (K-Fold)

La validación cruzada con K-Fold ($k = 5$) permitió evaluar la capacidad de generalización del modelo optimizado. Los resultados promedio fueron:

- R^2 promedio: 0.9973, con una varianza mínima entre los folds, lo que indica estabilidad en el desempeño del modelo.
- MSE promedio: 0.0207, evidenciando un error cuadrático bajo en las predicciones.
- MAE promedio: 0.0930, lo que refuerza la precisión del modelo para predecir valores cercanos a los observados.

La validación cruzada confirmó que el modelo es robusto y generaliza bien a diferentes particiones del conjunto de datos.

Curva de aprendizaje

La curva de aprendizaje del modelo Lasso muestra la relación entre el tamaño del conjunto de entrenamiento y el desempeño del modelo (R^2):

- El score de entrenamiento se estabiliza rápidamente cerca de $R^2 = 1.0$, lo que indica que el modelo ajusta perfectamente los datos de entrenamiento incluso con conjuntos pequeños.
- El score de validación mejora sustancialmente con un aumento en el tamaño del conjunto de entrenamiento, alcanzando un desempeño consistente con el de entrenamiento. Esto sugiere que el modelo no está sufriendo ni de sesgo (bajo desempeño en entrenamiento) ni de varianza (discrepancia significativa entre entrenamiento y validación).

El comportamiento observado confirma que el modelo Lasso es eficiente en el manejo de alta dimensionalidad y mejora su desempeño con más datos.

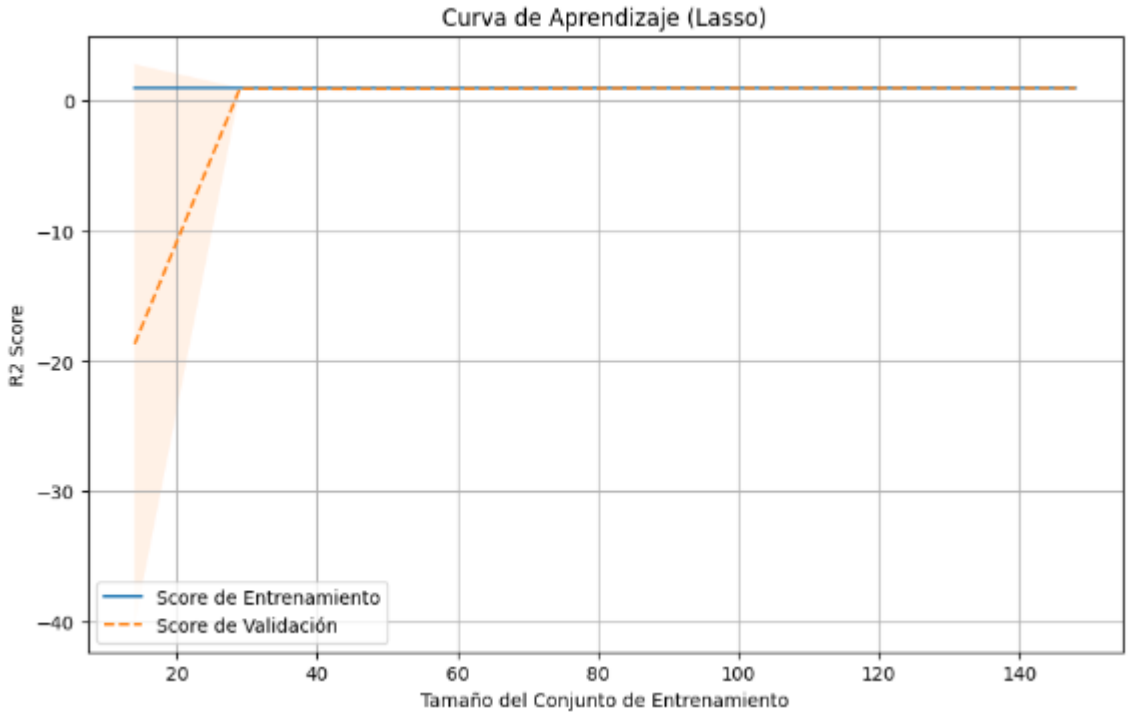

Ilustración 6. *Curva aprendizaje Lasso*

Análisis comparativo

Se presenta un gráfico que resume el desempeño del modelo en las tres etapas clave:

```
=== Tabla de Resultados Comparativos ===
                                Etapa       MSE       MAE      MAPE  R2 Score
0                  Evaluación Inicial  0.822580  0.707724  0.183773  0.877443
1               Después de Grid Search  0.027475  0.099117  0.019379  0.995906
2  Evaluación Final (Cross-Validation)  0.027475  0.099117  0.019379  0.995906
```

Ilustración 7*. Resultados comparativos Lasso inicial*

Conclusiones

- Mejora sustancial tras optimización: El ajuste del parámetro λ a través de Grid Search mejoró significativamente el desempeño del modelo Lasso, logrando un R^2 cercano al ideal (99.59%) y reduciendo drásticamente los errores de predicción.
- Manejo eficiente de multicolinealidad: Lasso no solo mitigó los efectos negativos de la multicolinealidad, sino que también seleccionó automáticamente las características más relevantes al asignar coeficientes cero a aquellas no informativas.
- Capacidad de generalización: La validación cruzada demostró que el modelo optimizado generaliza bien a diferentes particiones de los datos.

- Eficiencia en alta dimensionalidad: A pesar de la elevada relación entre características y observaciones, el modelo logró identificar patrones significativos y robustos en los datos.

6.5. Modelo Random Forest

Revisión de multicolinealidad

```
=== Revisión de Multicolinealidad (VIF) ===
                        feature  VIF
0      ventanilla_expansion_(%)  inf
1    ventanilla_contraccion_(%)  inf
2                inflacion_total  inf
3                 meta_inflacion  inf
4       IPP_produccion_nacional  inf
..                          ...  ...
435      s&p_500_lag1_lag2_lag3  inf
436    dow_jones_lag1_lag2_lag3  inf
437       nasdaq_lag1_lag2_lag3  inf
438     ftse_col_lag1_lag2_lag3  inf
439     EFFR_FED_lag1_lag2_lag3  inf
```

Ilustración 8. *Resultados multicolinealidad Random Forest*

El análisis de $Variance\ Inflation\ Factor$ (VIF) mostró valores extremadamente altos (infinito) para todas las características, indicando un alta multicolinealidad en el conjunto de datos. Aunque la multicolinealidad suele ser problemática para los modelos lineales debido a su impacto en la estabilidad de los coeficientes, el modelo Random Forest no se ve afectado directamente por esta condición. Esto se debe a que los árboles de decisión subyacentes seleccionan características de manera aleatoria y no dependen de combinaciones lineales de las variables. En este sentido, Random Forest es una herramienta robusta frente a la multicolinealidad.

Resultados iniciales del modelo Random Forest

El modelo Random Forest fue entrenado con los hiperparámetros predeterminados y evaluado en el conjunto de prueba. Las métricas obtenidas fueron:

- MSE: 0.02797
- MAE: 0.1006
- MAPE: 2.48%
- R^2: 0.9958

El R^2 obtenido (99.58%) indica que el modelo explica el 99.58% de la varianza de la variable objetivo, lo que refleja un excelente desempeño inicial. Las métricas de error (MSE, MAE, MAPE) son bajas, indicando que el modelo predice con alta precisión los valores de la tasa de intervención política monetaria.

Optimización con Grid Search

La búsqueda identificó la mejor configuración de parámetros como:

- $n_estimators$ (número de árboles): 200
- $\max\ _depth$ (profundidad máxima de los árboles): 10
- $\min\ _samples_split$ (número mínimo de muestras requeridas para dividir un nodo): 5

El modelo optimizado fue evaluado nuevamente en el conjunto de prueba, obteniendo las siguientes métricas:

- MSE: 0.03190
- MAE: 0.1026
- MAPE: 2.40%
- R^2: 0.9952

Aunque el desempeño general sigue siendo excelente, se observa un ligero incremento en el error (MSE y MAE) y una pequeña disminución en el R^2. Esto puede deberse a que la optimización introdujo restricciones adicionales en los árboles (como la menor profundidad), reduciendo ligeramente la capacidad del modelo para ajustarse perfectamente a los datos.

Validación cruzada con K-Fold

Para evaluar la capacidad de generalización del modelo optimizado, se realizó una validación cruzada con K-Fold ($k = 5$). Los resultados promedio de las métricas fueron:

- R^2: 0.9945
- MSE promedio: 0.0433
- MAE promedio: 0.1337

El R^2 promedio refleja que el modelo mantiene un excelente desempeño al generalizar en diferentes particiones del conjunto de datos. Sin embargo, la mayor variabilidad en las métricas, como el MAE promedio (0.1337) en comparación con el conjunto de prueba (0.1026), indica que el modelo podría ser sensible a ciertas divisiones de los datos. Esto es común en problemas con alta dimensionalidad y un número limitado de observaciones (199 filas).

Curva de aprendizaje

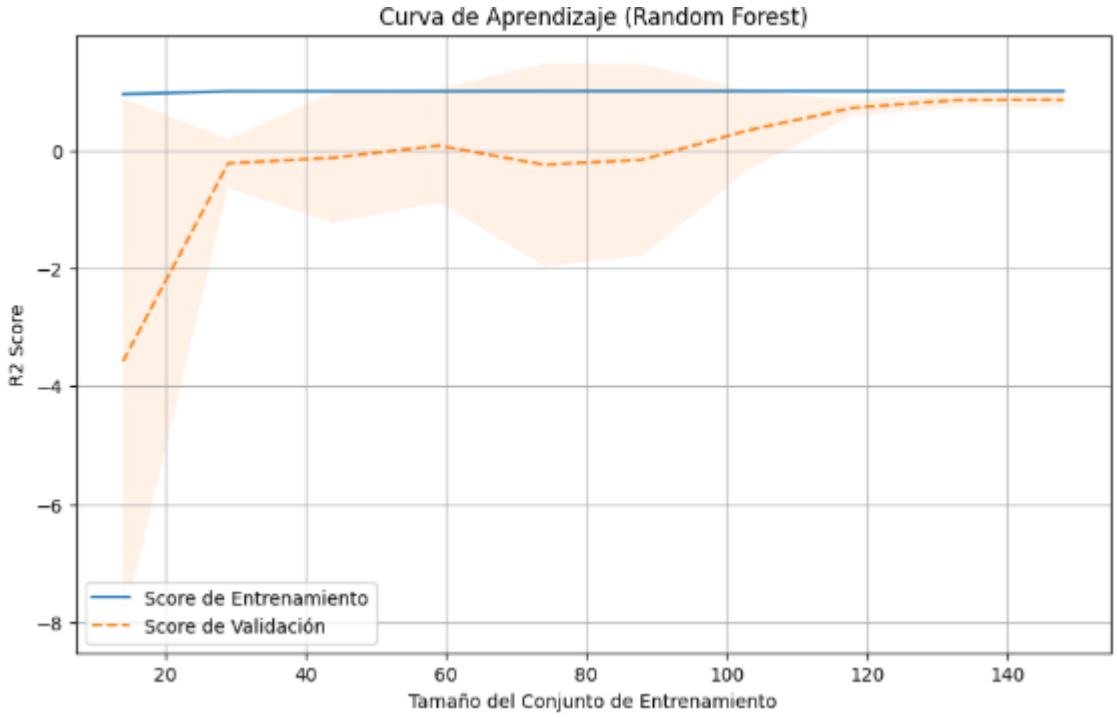

***Ilustración 9**. Curva aprendizaje Random Forest*

La curva de aprendizaje muestra cómo varía el desempeño del modelo (R^2) con diferentes tamaños del conjunto de entrenamiento:

- El score de entrenamiento permanece constante y cercano a 1.0, indicando que el modelo ajusta perfectamente los datos de entrenamiento.
- El score de validación mejora significativamente con el aumento del tamaño del conjunto de entrenamiento, estabilizándose cerca del score de entrenamiento. Esto evidencia que el modelo no sufre de sobreajuste ni subajuste, pero requiere una cantidad adecuada de datos para alcanzar su mejor desempeño.

Análisis comparativo

Se presenta un resumen de las métricas del modelo en las tres etapas:

```
=== Tabla de Resultados Comparativos ===
                                Etapa       MSE       MAE      MAPE  R2 Score
0                 Evaluación Inicial  0.027965  0.100606  0.024771  0.995833
1              Después de Grid Search  0.031904  0.102596  0.023990  0.995247
2  Evaluación Final (Cross-Validation)  0.031904  0.102596  0.023990  0.995247
```

***Ilustración 10.** Resultados comparativos Random Forest inicial*

Conclusiones

1. Capacidad predictiva sobresaliente:
    - El modelo mostró un buen desempeño en todas las etapas, explicando más del 99% de la varianza de la variable objetivo (R^2 >0.994).
    - Los errores (MSE, MAE, MAPE) son bajos, lo que indica predicciones precisas.
2. Generalización robusta: la validación cruzada confirma que el modelo generaliza bien en diferentes particiones del conjunto de datos, aunque con una ligera variabilidad en las métricas.
3. Impacto de la optimización: La optimización con Grid Search ajustó los hiperparámetros para balancear la capacidad predictiva y la simplicidad del modelo. Sin embargo, esto introdujo una pequeña reducción en el R^2y un aumento en los errores en comparación con el modelo inicial.
4. Ventajas frente a la multicolinealidad: A pesar del alta multicolinealidad (valores de VIF infinitos), Random Forest no se vio afectado, gracias a su método de selección aleatoria de características.
5. Eficiencia en alta dimensionalidad: el modelo manejó con bien el conjunto de datos con 440 características y 199 observaciones, proporcionando un modelo robusto y escalable.

6.6. Modelo Gradient Boosting – XGBoost

Revisión de multicolinealidad

```
=== Revisión de Multicolinealidad (VIF) ===
                        feature  VIF
0      ventanilla_expansion_(%)  inf
1    ventanilla_contraccion_(%)  inf
2                inflacion_total  inf
3                 meta_inflacion  inf
4       IPP_produccion_nacional  inf
..                          ...  ...
435      s&p_500_lag1_lag2_lag3  inf
436    dow_jones_lag1_lag2_lag3  inf
437       nasdaq_lag1_lag2_lag3  inf
438     ftse_col_lag1_lag2_lag3  inf
439     EFFR_FED_lag1_lag2_lag3  inf
```

***Ilustración 11**. Resultados multicolinealidad XGBoost*

El análisis de $Variance\ Inflation\ Factor$ (VIF) mostró valores infinitos para todas las características, indicando un alta multicolinealidad en el conjunto de datos. Sin embargo, el modelo XGBoost no depende de relaciones lineales entre variables, ya que utiliza árboles de decisión como base. Este enfoque lo hace robusto frente a la redundancia en las características, seleccionando de manera jerárquica las variables más relevantes en cada iteración.

Resultados iniciales del modelo

El modelo inicial de XGBoost fue entrenado con los hiperparámetros predeterminados, obteniéndose las siguientes métricas en el conjunto de prueba:

- MSE: 0.03266
- MAE: 0.10088
- MAPE: 2.07%
- R^2: 0.99513

El modelo explicó el 99.51% de la varianza de la variable objetivo (R^2), lo que indica un excelente ajuste inicial. Las métricas de error (MSE, MAE, MAPE) son bajas, reflejando una alta precisión en las predicciones.

Optimización con Grid Search

Los mejores hiperparámetros encontrados fueron:

- $learning_rate$: 0.1
- $n_estimators$: 100
- $\max\ _depth$: 3

```
=== Resultados Después de Grid Search (XGBoost) ===
Mejores Parámetros: {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100}
Mean Squared Error (MSE): 0.02783755988794955
Mean Absolute Error (MAE): 0.09801124696618235
Mean Absolute Percentage Error (MAPE): 0.01866890547328081
R2 Score: 0.9958524661803092
```

Ilustración 12. *Resultados optimización grid search XGBoost*

El modelo optimizado fue evaluado nuevamente en el conjunto de prueba, obteniéndose las siguientes métricas:

- MSE: 0.02784
- MAE: 0.09801
- MAPE: 1.87%
- R^2: 0.99585

La optimización mejoró todas las métricas respecto al modelo inicial, destacando una reducción en el error absoluto medio (MAE) y el error porcentual (MAPE). El R^2 aumentó a 99.59%, lo que refleja un ajuste aún más preciso.

Validación cruzada con K-Fold

Se realizó una validación cruzada con K-Fold ($k = 5$) para evaluar la capacidad de generalización del modelo optimizado. Los resultados promedio fueron:

```
=== Resultados de K-Fold (XGBoost) ===
R2 por cada fold: [0.99585247 0.99565577 0.9955958  0.98777139 0.99557209]
R2 promedio: 0.9941
MSE por cada fold: [0.02783756 0.03853125 0.02251029 0.08833798 0.05250844]
MSE promedio: 0.0459
MAE por cada fold: [0.09801125 0.11561525 0.09557503 0.1317072  0.13976048]
MAE promedio: 0.1161
```

Ilustración 13. *Resultados validación cruzada XGBoost*

- R^2: 0.9941
- MSE promedio: 0.0459
- MAE promedio: 0.1161

Aunque el desempeño del modelo se mantuvo alto, la validación cruzada mostró un aumento leve en las métricas de error (MSE y MAE promedio) en comparación con el conjunto de prueba. Esto sugiere una ligera sensibilidad del modelo a ciertas particiones del conjunto de datos, algo común en escenarios con alta dimensionalidad y un número limitado de observaciones.

Curva de aprendizaje

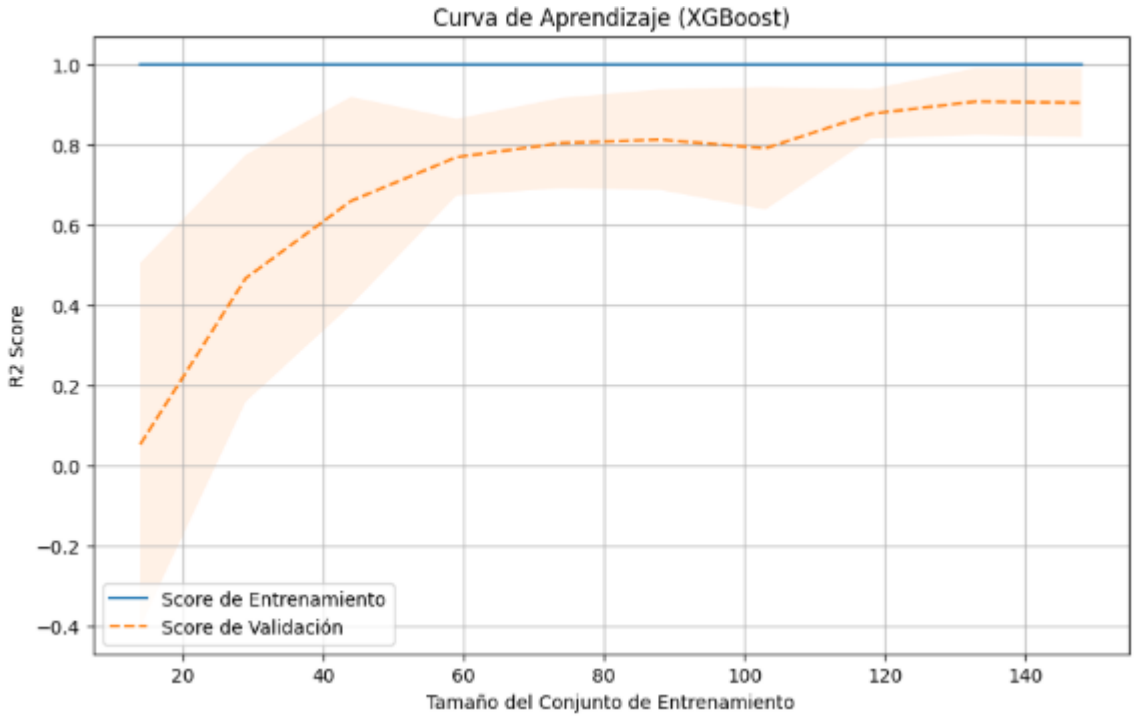

Ilustración 14. *Curva aprendizaje XGBoost*

La curva de aprendizaje muestra cómo varía el desempeño del modelo (R^2) en función del tamaño del conjunto de entrenamiento:

- El score de entrenamiento permanece cercano a 1.0, lo que indica que el modelo ajusta perfectamente los datos de entrenamiento.
- El score de validación mejora significativamente a medida que aumenta el tamaño del conjunto de entrenamiento, estabilizándose cerca del score de entrenamiento para tamaños mayores. Esto sugiere que el modelo no sufre de sobreajuste ni subajuste, aunque requiere suficientes datos para alcanzar su mejor desempeño.

Comparación resultados

A continuación, se resumen las métricas del modelo en las tres etapas:

```
=== Tabla de Resultados Comparativos ===
                              Etapa       MSE       MAE      MAPE  R2 Score
0                  Evaluación Inicial  0.032663  0.100881  0.020688  0.995134
1              Después de Grid Search  0.027838  0.098011  0.018669  0.995852
2  Evaluación Final (Cross-Validation)  0.027838  0.098011  0.018669  0.995852
```

***Ilustración 15**. Resultados comparativos XGBoost inicial*

Conclusiones

- Capacidad predictiva sobresaliente: El modelo XGBoost optimizado mostró un desempeño altamente preciso, explicando más del 99% de la varianza (R^2) en todas las etapas y las métricas de error son bajas, indicando predicciones consistentes y cercanas a los valores observados.
- Impacto de la optimización: La optimización con Grid Search mejoró las métricas de error (MSE, MAE, MAPE) y el ajuste general (R^2) y el modelo optimizado es más eficiente, con un menor número de árboles y de una profundidad limitada, lo que reduce el riesgo de sobreajuste.
- Generalización robusta: La validación cruzada confirma que el modelo generaliza bien a diferentes particiones del conjunto de datos. Aunque las métricas de error promedio aumentaron ligeramente, el desempeño sigue siendo consistente.
- Ventajas frente a alta dimensionalidad: XGBoost manejó bien las 440 características del conjunto de datos, aprovechando su capacidad para manejar relaciones no lineales y multicolinealidad.

6.7. Modelo Support Vector Regressor (SVR)

Revisión de multicolinealidad

El análisis del $Variance\ Inflation\ Factor$ (VIF) reveló valores infinitos para las características, indicando un alta multicolinealidad en el conjunto de datos. Sin embargo, el modelo SVR no se ve directamente afectado por la multicolinealidad, ya que no depende de relaciones lineales entre las características. En cambio, utiliza funciones kernel para mapear los datos a espacios de mayor

dimensión, lo que le permite capturar relaciones complejas sin verse limitado por la redundancia en las variables.

```
=== Revisión de Multicolinealidad (VIF) ===
                          feature  VIF
0      ventanilla_expansion_(%)  inf
1    ventanilla_contraccion_(%)  inf
2               inflacion_total  inf
3                meta_inflacion  inf
4       IPP_produccion_nacional  inf
..                          ...  ...
435      s&p_500_lag1_lag2_lag3  inf
436    dow_jones_lag1_lag2_lag3  inf
437       nasdaq_lag1_lag2_lag3  inf
438     ftse_col_lag1_lag2_lag3  inf
439     EFFR_FED_lag1_lag2_lag3  inf
```

Ilustración 16. *Resultados multicolinealidad SVR*

Resultados iniciales del modelo

El modelo inicial de SVR fue entrenado con los hiperparámetros predeterminados **(**$C = 1.0, \varepsilon = 0.1, kernel\ rbf$), obteniendo las siguientes métricas en el conjunto de prueba:

- MSE: 0.08948
- MAE: 0.24086
- MAPE: 5.79%
- R^2: 0.98667

El modelo explicó el 98.67% de la varianza (R^2), lo que indica un buen desempeño inicial, aunque las métricas de error (MSE, MAE, MAPE) son más altas en comparación con otros modelos estudiados. Este resultado sugiere que una optimización de hiperparámetros puede mejorar significativamente el desempeño.

Optimización con Grid Search

Los mejores hiperparámetros seleccionados fueron:

- $C = 0.1$
- $\varepsilon = 0.1$
- Kernel: linear

```
=== Resultados Después de Grid Search (SVM) ===
Mejores Parámetros: {'C': 0.1, 'epsilon': 0.1, 'kernel': 'linear'}
Mean Squared Error (MSE): 0.03717302681385872
Mean Absolute Error (MAE): 0.14156877065591333
Mean Absolute Percentage Error (MAPE): 0.03572057537993032
R2 Score: 0.9944615696738027
```

Ilustración 17. *Resultados optimización grid search SVR*

El modelo optimizado fue evaluado en el conjunto de prueba, obteniendo:

- MSE: 0.03717
- MAE: 0.14157
- MAPE: 3.57%
- R^2: 0.99446

Tras la optimización, el R^2 mejoró a 99.45%, indicando que el modelo explica un mayor porcentaje de la varianza. Además, las métricas de error (MSE, MAE y MAPE) se redujeron significativamente, destacando una mejora notable en la precisión del modelo.

Validación cruzada con K-Fold

Para evaluar la capacidad de generalización del modelo, se realizó una validación cruzada con K-Fold ($k = 5$), obteniendo:

- R^2promedio: 0.9954
- MSE promedio: 0.0334
- MAE promedio: 0.1374

```
=== Resultados de K-Fold (SVM) ===
R2 por cada fold: [0.99446586 0.99615759 0.99351346 0.9958573  0.99722967]
R2 promedio: 0.9954
MSE por cada fold: [0.03714425 0.03408037 0.03315331 0.0299264  0.03285209]
MSE promedio: 0.0334
MAE por cada fold: [0.14151573 0.14206928 0.14288437 0.12501844 0.1353735 ]
MAE promedio: 0.1374
```

Ilustración 18. *Resultados validación cruzada SVR*

La validación cruzada confirma la robustez del modelo optimizado, con un desempeño consistente en las diferentes particiones del conjunto de datos. Las métricas de error promedio son ligeramente más bajas en comparación con los resultados del conjunto de prueba, lo que evidencia que el modelo generaliza bien.

Curva de aprendizaje

La curva de aprendizaje muestra cómo el desempeño (R^2) varía en función del tamaño del conjunto de entrenamiento:

- El score de entrenamiento permanece estable y cercano al máximo ($R^2 \approx 1$), lo que indica que el modelo ajusta bien los datos de entrenamiento.
- El score de validación mejora progresivamente con el aumento del tamaño del conjunto de entrenamiento, estabilizándose cerca del score de entrenamiento. Esto sugiere que el modelo no sufre de sobreajuste ni subajuste y puede beneficiarse de conjuntos de datos más grandes.

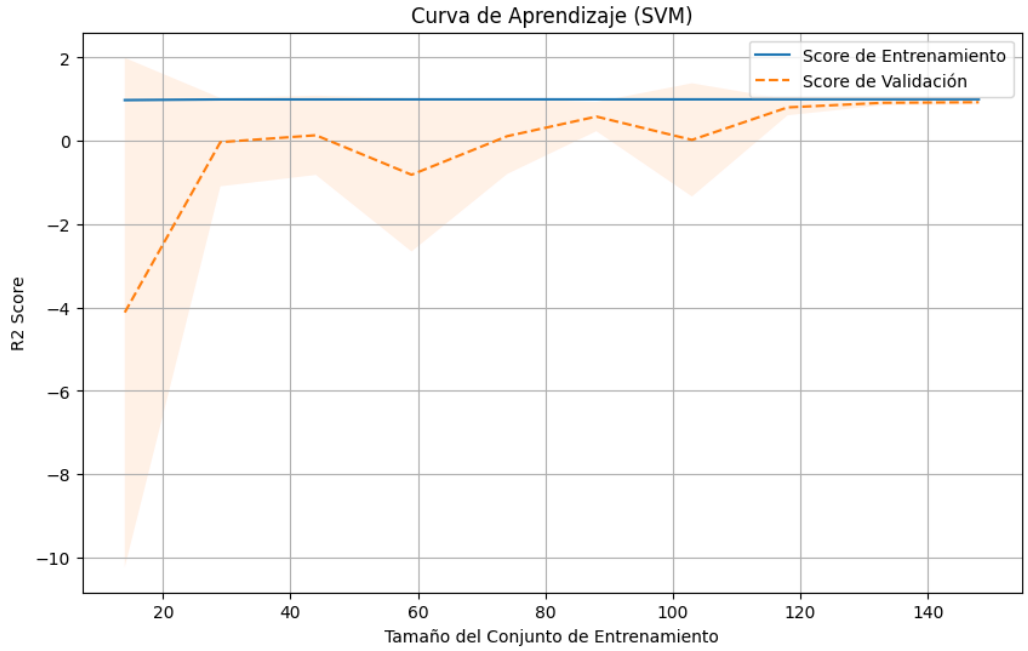

Ilustración 19. *Curva aprendizaje SVR*

Comparación resultados

A continuación, se presentan las métricas del modelo en las tres etapas:

```
=== Tabla de Resultados Comparativos ===
                              Etapa       MSE       MAE      MAPE  R2 Score
0                  Evaluación Inicial  0.089477  0.240862  0.057929  0.986669
1              Después de Grid Search  0.037173  0.141569  0.035721  0.994462
2  Evaluación Final (Cross-Validation)  0.037173  0.141569  0.035721  0.994462
```

Ilustración 20. *Resultados comparativos SVR inicial*

Conclusiones

- Impacto de la optimización: La optimización con Grid Search mejoró notablemente el desempeño del modelo. El R^2aumentó de 98.67% a 99.54%, y las métricas de error (MSE, MAE, MAPE) se redujeron significativamente, lo que indica un modelo más preciso.
- Capacidad de generalización: Los resultados de la validación cruzada confirman que el modelo es robusto y generaliza bien a diferentes particiones del conjunto de datos.
- Eficiencia del kernel lineal: La selección del kernel lineal como el óptimo indica que las relaciones en el conjunto de datos pueden ser aproximadamente lineales, reduciendo la necesidad de transformaciones más complejas como rbf.

6.8. Comparación general

Realizando la comparación del modelo inicial entre Ridge, Lasso, Random Forest, XGBoost y SVR mediante sus métricas principales, tenemos lo siguiente:

Modelo	R^2	MSE	MAE	MAPE
Ridge	0.99425	0.03859	0.13261	3.10%
Lasso	0.87744	0.82258	0.70772	18.38%
Random Forest	0.99583	0.02796	0.10061	2.48%
XGBoost	0.99513	0.03266	0.10088	2.07%
SVR	0.98667	0.08948	0.24086	5.79%

Tabla 2. *Tabla comparativa métricas implementación inicial modelos*

- R^2(Varianza explicada): El mejor modelo inicial es Random Forest (0.99583), seguido de XGBoost (0.99513) y Ridge (0.99425). Por lo que, Random Forest explica la mayor proporción de la varianza de la variable objetivo.
- MSE: El menor error cuadrático medio lo tiene Random Forest (0.02796), lo que lo posiciona como el modelo con predicciones más precisas inicialmente, seguido de XGBoost (0.03266) y Ridge (0.03859).
- MAE: Random Forest (0.10061) y XGBoost (0.10088) tienen errores absolutos medios casi idénticos y significativamente más bajos que los demás modelos, lo que sugiere una alta precisión en las predicciones.
- $MAPE$: El modelo con menor error porcentual relativo es XGBoost (2.07%), seguido de Random Forest (2.48%) y Ridge (3.10%).
- Curva de aprendizaje: Random Forest y XGBoost tienen las curvas más estables y robustas, con buena capacidad de generalización y ajuste a los datos de entrenamiento. Ridge se mantiene competitivo, pero ligeramente inferior, reflejando un desempeño menos eficiente en la generalización comparado con los modelos basados en árboles. SVR y Lasso muestran limitaciones iniciales significativas, con curvas de aprendizaje que indican mayor sensibilidad al tamaño de los datos y la necesidad de optimizaciones adicionales para alcanzar un rendimiento comparable.

Dado que el proyecto no solo busca precisión predictiva, sino también la identificación de las variables más influyentes en la variable objetivo, la alta dimensionalidad del conjunto de datos es un factor crítico a considerar. En el siguiente capítulo, se aplicaron análisis de correlación y técnicas de selección de características para reducir la redundancia entre variables y destacar aquellas que tienen mayor impacto en la predicción de la tasa de intervención política monetaria. Esto permitió no solo mejorar el desempeño de los modelos, sino también interpretar los resultados de manera más clara y alineada con el objetivo final del estudio.

7. ESTRATEGIAS DE OPTIMIZACIÓN Y SELECCIÓN DE VARIABLES DE MODELOS INICIALES IMPLEMENTADOS

7.1. Introducción

Este capítulo se alinea con el Objetivo Específico 4 de la investigación, enfocado en identificar y analizar las variables más relevantes para la predicción de la tasa de intervención política monetaria, utilizando un enfoque sistemático de análisis de correlación, selección de características y evaluación de modelos propuestos en el capítulo 6.

El desarrollo de este capítulo sigue una estructura metodológica clara, organizada en cinco actividades principales:

Análisis de correlación y selección preliminar de variables: Se realizó un análisis de correlación sobre los datos con tres rezagos, identificando las variables con mayor relevancia y eliminando aquellas altamente correlacionadas para reducir la redundancia en los datos. Para los modelos Ridge, Lasso y SVR, este análisis se realizó con datos estandarizados, mientras que Random Forest y XGBoost operarán sobre datos sin estandarizar.

Aplicación de técnicas avanzadas de selección de características: Una vez identificadas las variables relevantes mediante el análisis de correlación, se aplicó el método de Recursive Feature Elimination (RFE) como técnica principal de selección de características para todos los modelos. Este enfoque ajustó iterativamente cada modelo, eliminando las variables menos relevantes en función de su impacto predictivo, hasta obtener un conjunto óptimo de características.

Entrenamiento inicial de modelos: Utilizando las características seleccionadas, se entrenó cada modelo y se evaluó su desempeño inicial mediante métricas como Mean Squared Error (MSE), el Mean Absolute Error (MAE), el Mean Absolute Percentage Error (MAPE) y el coeficiente de determinación (R^2). Este análisis permitió evaluar cómo la selección de características impactó el desempeño de los modelos en comparación con los resultados obtenidos en el capítulo anterior.

Optimización de hiperparámetros y validación cruzada: Para garantizar que los modelos alcancen su máximo desempeño, se optimizó mediante Grid Search para ajustar sus hiperparámetros de manera sistemática. Posteriormente, se implementó validación cruzada con K-Fold, reportando nuevamente las métricas clave, lo que aseguró la robustez y la capacidad de generalización de los modelos.

Análisis avanzado de resultados: Finalmente, se complementó la evaluación con análisis avanzados, incluyendo:

- Curvas de aprendizaje, para estudiar cómo el tamaño del conjunto de entrenamiento impacta el desempeño de los modelos.

- Análisis de sensibilidad, que permitió identificar la variabilidad en las predicciones frente a cambios en las variables seleccionadas.
- Análisis de residuos, para verificar la calidad de las predicciones y detectar posibles patrones no capturados por los modelos.
- Análisis de robustez, evaluando el comportamiento de los modelos frente a diferentes configuraciones y subconjuntos de datos.
- Interpretabilidad mediante SHAP, proporcionando una visión clara del impacto de cada variable en las predicciones y asegurando que los resultados sean comprensibles y útiles para la toma de decisiones.

El enfoque adoptado en este capítulo responde a los desafíos de manejar un conjunto de datos de alta dimensionalidad y garantizar tanto la precisión predictiva como la interpretabilidad. Mientras que en el capítulo anterior se evaluó el desempeño bruto de los modelos, aquí se incorpora un análisis detallado de las características más influyentes, alineándose directamente con el objetivo de identificar las variables clave en la predicción de la tasa de intervención política monetaria.

La metodología descrita no solo mejoró el desempeño de los modelos, sino que también establece un marco sólido para interpretar los resultados y generar conocimiento útil sobre las relaciones entre las variables macroeconómicas y la variable objetivo. Este capítulo, por lo tanto, refuerza los aportes del proyecto al proporcionar una visión integral del impacto de las características en la predicción de tasas de interés.

7.2. Análisis de correlación

7.2.1. Descripción

Preparación de los datos

Se inicia definiendo las variables predictoras (X) y la variable objetivo (y). Las columnas no necesarias para el análisis (la columna fecha) son eliminadas, asegurando que solo se incluyan variables relevantes. Este paso prepara el conjunto de datos para la ejecución del análisis de correlación.

Cálculo de la matriz de correlación

Se calcula la matriz de correlación absoluta ($corr_matrix$) para determinar la relación lineal entre cada par de variables predictoras. Este cálculo permite identificar pares de variables cuya correlación supera un umbral predefinido (0.8 en este caso), lo que indica que estas variables están altamente correlacionadas y podrían aportar información redundante al modelo.

Los pares de variables altamente correlacionadas se identifican utilizando la matriz triangular superior ($np.triu$) de la matriz de correlación, lo que reduce el procesamiento redundante al evitar la consideración de correlaciones duplicadas o de las variables consigo mismas.

Agrupación por nombre base de variables

Las variables que representan el mismo concepto en distintos rezagos (por ejemplo, PIB_real, PIB_real_lag1, PIB_real_lag2) son agrupadas según su nombre base. Este agrupamiento se logra mediante una función que elimina los sufijos de rezagos de los nombres de las variables, dejando únicamente el concepto base. Este paso garantiza que variables que representan el mismo fenómeno sean tratadas conjuntamente durante la selección, evitando que múltiples versiones redundantes sean incluidas.

Selección de variables representativas

De cada grupo de variables correlacionadas, se selecciona la variable con el rezago más reciente. Esto se basa en el supuesto de que los datos más recientes contienen información más relevante para explicar la variación de la variable objetivo (y). La selección se realiza identificando la variable con el mayor número de sufijos (lag), por ejemplo, seleccionando $PIB_real_lag1_lag2_lag3$ en lugar de PIB_real o PIB_real_lag1.

Filtrado de variables redundantes

El conjunto de datos se divide en variables seleccionadas que son las variables representativas de cada grupo correlacionado que se conservan para análisis posteriores y las variables eliminadas que son redundantes y eliminadas para evitar la duplicación de información.

7.2.2. Resultados

El análisis de correlación y la selección de rezagos más recientes resultaron en un conjunto reducido de 55 variables seleccionadas a partir de las 440 variables iniciales. Esto implica que se eliminaron 385 variables redundantes.

7.3. Descripción proceso modelos

Preparación de los datos

En todos los modelos, el proceso de preparación de datos siguió un enfoque estandarizado para garantizar consistencia en las etapas posteriores:

- Definición de variables: X se definió como el conjunto de variables predictoras, excluyendo la columna de fecha y la variable objetivo. Y corresponde a la Tasa de Intervención Política Monetaria.

- División de datos: Los datos fueron divididos en un conjunto de entrenamiento (70%) y un conjunto de prueba (30%) utilizando la función $train_test_split$. Para asegurar reproducibilidad, se utilizó una semilla de aleatoriedad ($random_state = 42$).

Selección de características

Para reducir la dimensionalidad y seleccionar las variables más relevantes, se empleó la técnica de Recursive Feature Elimination (RFE) en todos los modelos. El procedimiento consistió en:

- *Ridge* y *Lasso*: Regularización para penalizar características menos relevantes.
- *Random Forest* y *XGBoost*: Importancia basada en pesos de los árboles.
- *SVR*: Importancia relativa basada en soporte de vectores.

Número de características seleccionadas:

- *Ridge* y *Lasso*: 15 y 7 características, respectivamente.
- *Random Forest, XGBoost* y *SVR*: 7 características cada uno.

El modelo eliminó características menos importantes en cada iteración, conservando las variables con mayor capacidad predictiva.

Entrenamiento inicial

Con las características seleccionadas, se ajustaron los modelos iniciales utilizando los datos de entrenamiento y se evaluó su desempeño tanto en el conjunto de entrenamiento como en el de prueba. Las métricas empleadas para este análisis incluyeron los errores, representados por el error cuadrático medio (MSE), el error absoluto medio (MAE) y el porcentaje de error absoluto medio (MAPE), así como la capacidad explicativa medida a través del coeficiente de determinación (R^2).

Optimización de hiperparámetros

Para mejorar el rendimiento de los modelos, se utilizó Grid Search con validación cruzada (5 o 10 particiones, según el modelo) para encontrar la mejor combinación de hiperparámetros clave:

- Ridge y Lasso: Regularización (α), tolerancia (tol) y número de iteraciones máximas (max_iter).
- Random Forest: Número de árboles ($n_estimators$), profundidad máxima (max_depth), mínimo número de muestras para dividir nodos ($min_samples_split$) y para formar hojas ($min_samples_leaf$).
- XGBoost: Tasa de aprendizaje ($learning_rate$), número de árboles ($n_estimators$), y profundidad máxima (max_depth).
- SVR: Penalización (C), margen de tolerancia (ϵ) y tipo de kernel ($linear, rbf$).

Los modelos optimizados se ajustaron nuevamente y se evaluaron en el conjunto de prueba, obteniendo métricas actualizadas que reflejaron la mejora en su desempeño.

Validación cruzada

Para evaluar la capacidad de generalización de los modelos y garantizar su robustez en datos no vistos, se utilizó un esquema de validación cruzada K-Fold con un número de particiones adaptado a las necesidades de cada modelo:

- Ridge y Lasso: Se aplicó K-Fold ($k = 5$) para calcular métricas promedio como R^2, MSE, MAE y MAPE. Este enfoque ayudó a identificar el valor óptimo del hiperparámetro α en Grid Search, equilibrando sesgo y varianza.
- Random Forest: Validación cruzada con K-Fold ($k = 10$) permitió evaluar diferentes combinaciones de hiperparámetros ($n_stimators$, max_depth, $min_samples_split$), seleccionando el mejor modelo con base en R^2 promedio.
- XGBoost: Se utilizó K-Fold ($k = 10$) integrado en Grid Search para optimizar $learning_rate$, $n_estimators$ y max_depth, asegurando un desempeño consistente y reduciendo el riesgo de sobreajuste.
- SVR: Con K-Fold ($k = 10$) se exploraron valores de C, ε y diferentes kernels (lineal y radial), seleccionando la combinación que maximizó R^2 y minimizó errores promedio.

Curvas de aprendizaje

La curva de aprendizaje se utiliza para analizar cómo varía el desempeño del modelo, medido a través del coeficiente de determinación (R^2), en función del tamaño del conjunto de entrenamiento, permitiendo identificar posibles problemas de sesgo o varianza. En este análisis, el score de entrenamiento refleja cómo se ajusta el modelo a los datos de entrenamiento, mientras que el score de validación evalúa su capacidad de generalización. El gráfico resultante proporciona información clave para comprender la capacidad del modelo de escalar eficientemente con un mayor volumen de datos.

Análisis de sensibilidad

- Ridge y Lasso: Efecto de α en el balance entre sobreajuste y generalización.
- Random Forest: Impacto de $n_estimators$ en el ajuste y precisión.
- XGBoost: Influencia de $learning_rate$ en la convergencia y predicción.
- SVR: Efecto de C en la penalización de errores y capacidad predictiva.

Para cada variación, se calcularon MSE, MAE, MAPE y R^2, organizando los resultados en los gráficos que facilitaron la interpretación.

Análisis de residuos

El análisis de residuos fue fundamental para evaluar las suposiciones del modelo y su desempeño:

- Gráfico Q-Q: Comparación de los residuos con una distribución normal.

Análisis de robustez

Se realizaron perturbaciones en las características seleccionadas, incrementando y disminuyendo su valor en un 1%, para analizar la sensibilidad de las predicciones del modelo:

- Impacto positivo/negativo: Cambios en las predicciones al modificar el valor de una característica.
- Gráficos de barras: Ilustración del efecto de las perturbaciones para identificar variables más influyentes.

Análisis interpretabilidad SHAP

- Gráfico de barras: Muestra el impacto promedio de cada característica.
- Gráfico de puntos: Distribución del impacto de cada característica a nivel individual.
- Gráfico Waterfall: Explicación detallada de una predicción específica, desglosando cómo cada característica influyó en el resultado.

7.4. Modelo Ridge RFE

Selección de características

La técnica RFE permitió identificar 15 características relevantes de entre un conjunto más amplio, seleccionadas por su importancia en el ajuste del modelo Ridge. Estas características abarcan las siguientes variables económicas:

```
Características seleccionadas por RFE: Index(['indice_precios_importaciones_lag1_lag2_lag3',
       'inflacion_total_lag1_lag2_lag3', 'valor_colcap_lag1_lag2_lag3',
       'IPP_produccion_nacional_lag1_lag2_lag3', 'ITCR_IPP_T_lag1_lag2_lag3',
       'oferta_monetaria_agregada_lag1_lag2_lag3', 'EFFR_FED_lag1_lag2_lag3',
       'futuros_petroleo_wti_lag1_lag2_lag3',
       'credito_domestico_neto_sector_privado_banrep_lag1_lag2_lag3',
       'salario_minimo_mensual_lag1_lag2_lag3', 'gasto_social_lag1_lag2_lag3',
       'TRM_lag1_lag2_lag3', 'dow_jones_lag1_lag2_lag3',
       'ITCR_IPC_NT_lag1_lag2_lag3',
       'credito_domestico_neto_sector_publico_secfinan_lag1_lag2_lag3'],
      dtype='object')
```

Ilustración 21. *Resultado RFE modelo Ridge*

Evaluación inicial

El modelo Ridge entrenado inicialmente con las características seleccionadas mostró un desempeño sólido:

- En el conjunto de entrenamiento, se obtuvo un R^2 de 0.9722, con un MSE de 0.2480, un MAE de 0.3831, y un MAPE de 8.41%.
- En el conjunto de prueba, el R^2 fue de 0.9399, con un MSE de 0.3531, un MAE de 0.4694, y un MAPE de 11.26%. Estos resultados reflejan una adecuada capacidad predictiva, aunque evidencian ligeros signos de sobreajuste debido a la mayor precisión en el conjunto de entrenamiento.

Optimización hiperparámetros Grid Search

- Tras la optimización de los hiperparámetros α (regularización), tolerancia y número máximo de iteraciones, el modelo Ridge ajustado mostró mejoras significativas en las métricas:
- En el conjunto de entrenamiento, se alcanzó un R^2 de 0.9777, con una reducción del MSE a 0.1990, un MAE de 0.3528, y un MAPE de 7.90%.
- En el conjunto de prueba, el R^2 aumentó a 0.9449, con un MSE de 0.3239, un MAE de 0.4467, y un MAPE de 11.15%. Esto demuestra que la regularización optimizada ($\alpha = 0.1$) logró reducir la complejidad del modelo, mejorando tanto la precisión como la generalización.

Validación cruzada con K-Fold

Se implementó validación cruzada con 5 particiones para evaluar la estabilidad y robustez del modelo:

- El R^2 promedio fue de 0.9663, lo que confirma la capacidad del modelo para generalizar en diferentes subconjuntos.
- El MSE promedio fue de 0.2646, y el MAE promedio alcanzó 0.4003. La estabilidad de estas métricas en múltiples pliegues valida la consistencia del modelo Ridge optimizado.

Curva de aprendizaje

El análisis de la curva de aprendizaje evidenció:

- Un rendimiento estable y elevado ($R^2 > 0.94$) incluso con tamaños reducidos del conjunto de entrenamiento.
- Una convergencia entre los scores de entrenamiento y validación, lo que indica un balance adecuado entre sesgo y varianza.

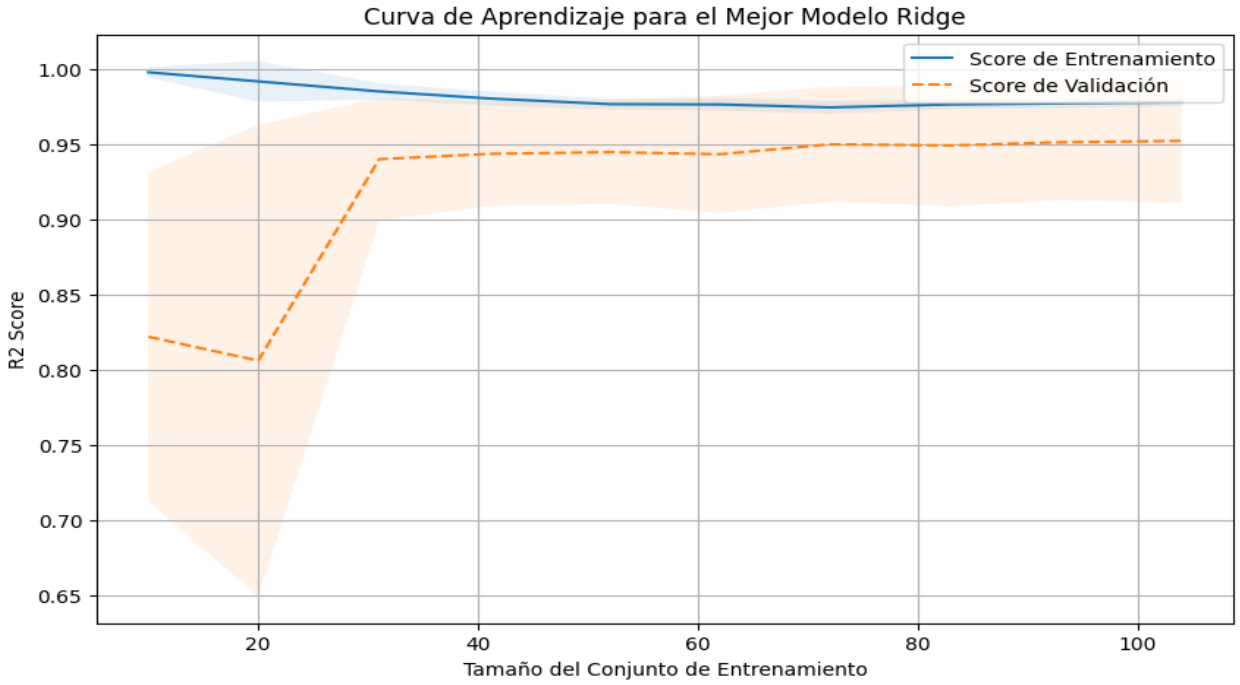

Ilustración 22. *Curva aprendizaje Modelo Ridge*

Análisis de sensibilidad

```
=== Análisis de Sensibilidad para el Hiperparámetro Alpha ===
    Alpha       MSE       MAE      MAPE        R²
0    0.01  0.328612  0.440417  0.111568  0.944055
1    0.10  0.323887  0.446680  0.111503  0.944860
2    1.00  0.353150  0.469437  0.112630  0.939878
3   10.00  0.488151  0.549666  0.133802  0.916895
4  100.00  1.457325  0.985873  0.264501  0.751897
```

Ilustración 23. *Análisis sensibilidad modelo Ridge*

- α óptimo: Se identificó que el valor de α = 0.1 ofreció el mejor balance entre precisión y generalización, con el menor MSE (0.3239) y el mayor R^2 (0.9449). Estos resultados confirman que una regularización moderada es adecuada para este modelo.
- Impacto de una baja regularización (α = 0.01): Aunque el R^2 sigue siendo alto (0.9441), el modelo presenta un leve aumento en el MSE y el MAE en comparación con el α = 0.1. Esto sugiere una tendencia hacia un sobreajuste.
- Impacto de una alta regularización (α > 1.0): A medida que α aumenta, se observa un deterioro significativo en las métricas. Con α =100, el MSE se incrementa a 1.4573 y el R^2 disminuye drásticamente a 0.7519, lo que indica una subestimación de los coeficientes y pérdida de información relevante.

El análisis revela que el hiperparámetro α tiene un impacto significativo en el desempeño del modelo. Valores bajos de α conducen a una menor regularización, mientras que valores altos penalizan excesivamente la complejidad, comprometiendo la capacidad predictiva del modelo $\alpha = 0.1$ resulta ser el mejor compromiso entre precisión y robustez.

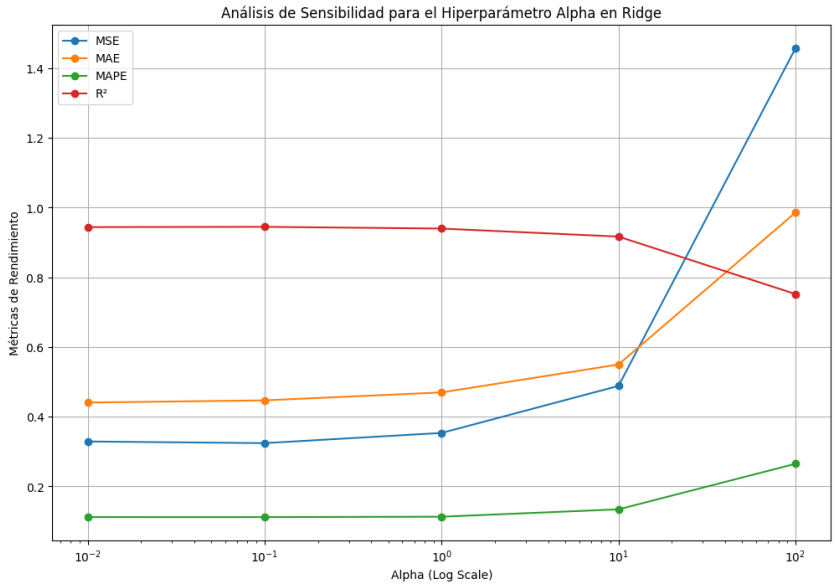

Ilustración 24. *Análisis sensibilidad hiperparámetro modelo Ridge*

Análisis de residuos

Gráfico Q-Q:

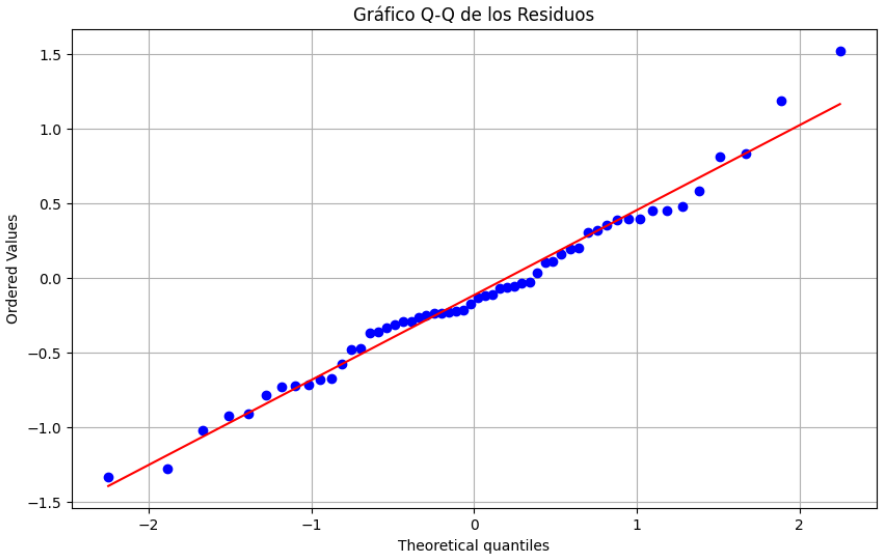

Ilustración 25*. Análisis residuos modelo Ridge*

La mayoría de los puntos en el gráfico Q-Q están alineados con la línea roja que representa los cuantiles teóricos de una distribución normal, lo que evidencia que los residuos del modelo tienen un comportamiento cercano a la normalidad, un aspecto deseable para la validez de los supuestos del modelo. Si bien los puntos centrales muestran una alineación precisa con la línea de referencia, se observan ligeras desviaciones en los cuantiles extremos, tanto en la cola inferior como en la superior, posiblemente relacionadas con valores atípicos o un ajuste imperfecto del modelo en casos extremos. La simetría en el gráfico, con los puntos distribuidos uniformemente alrededor de la línea de referencia, refuerza la hipótesis de que los residuos no presentan un sesgo significativo hacia valores positivos o negativos.

Análisis de robustez

- Un aumento del 1% en "$indice_precios_importaciones_lag1_lag2_lag3$" genera un incremento de aproximadamente 0.0085 unidades en la predicción, mientras que un decremento del 1% causa una disminución similar.
- Para "$futuros_petroleo_wti_lag1_lag2_lag3$", un incremento del 1% reduce la predicción en 0.0115 unidades, mientras que una disminución genera un cambio positivo equivalente.
- Las variables "$TRM_lag1_lag2_lag3$", "$IPP_produccion_nacional_lag1_lag2_lag3$", e "$inflacion_total_lag1_lag2_lag3$" también muestran sensibilidad moderada, con variaciones menores pero relevantes en las predicciones.
- Variables como "$gasto_social_lag1_lag2_lag3$" y "$EFFR_FED_lag1_lag2_lag3$" tienen impactos mínimos, con cambios cercanos a 0 en las predicciones, lo que sugiere una menor influencia directa en el modelo.
- El análisis revela una simetría en la magnitud de los efectos de incremento y decremento, lo que indica que el modelo responde de manera proporcional a cambios en ambas direcciones. Esto es consistente con la naturaleza lineal del modelo Ridge.
- Variables relacionadas con precios de importación, producción nacional, y tasa representativa del mercado (TRM) son determinantes clave en las predicciones del modelo, lo cual es congruente con la dinámica económica subyacente en la determinación de tasas de intervención monetaria.

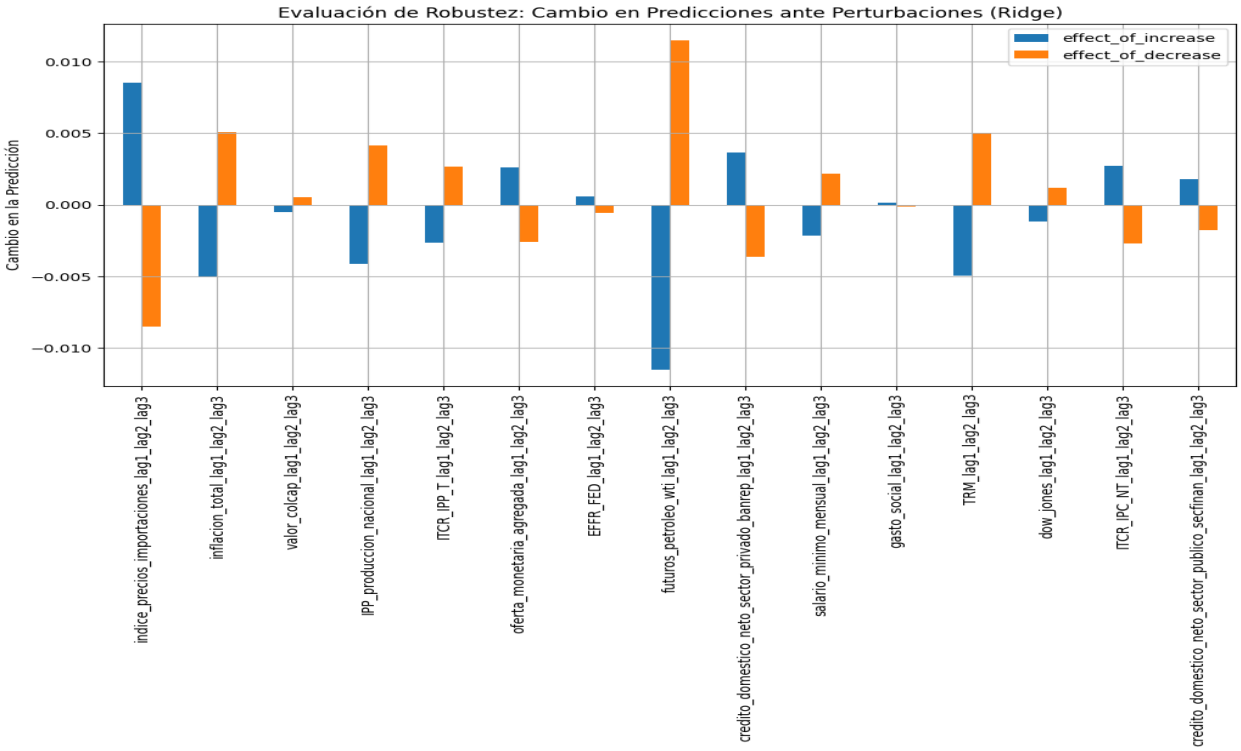

Ilustración 26. *Análisis de robustez modelo Ridge*

Análisis de interpretabilidad SHAP

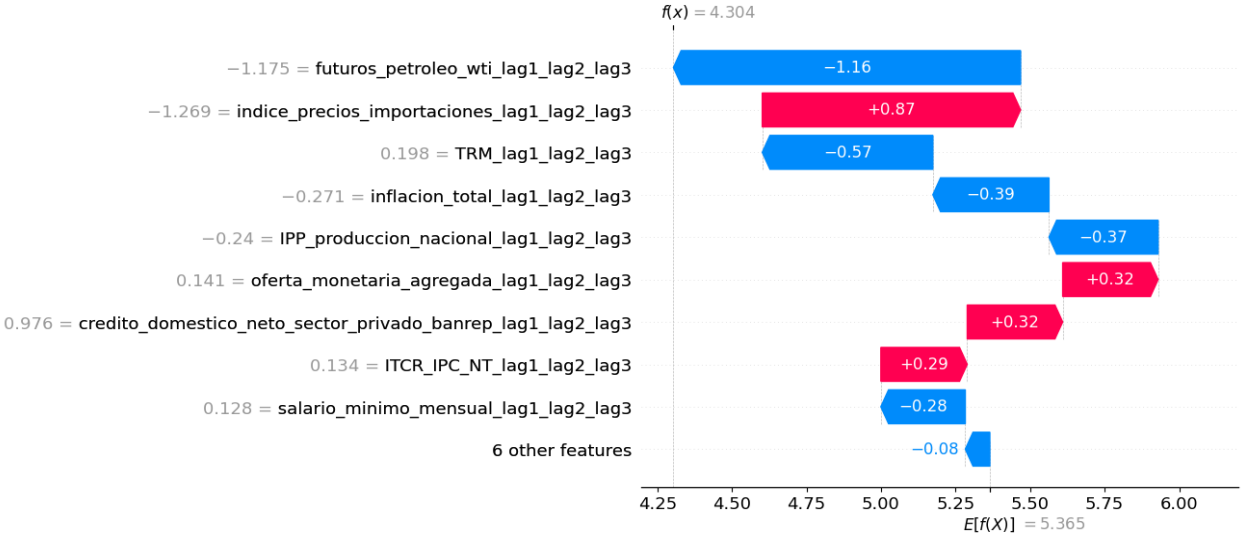

Ilustración 27. *Análisis de interpretabilidad SHAP modelo Ridge*

Mediante el gráfico tipo "waterfall" generado para una observación específica, se evidencia cómo cada característica contribuye al valor predicho individual del modelo para dicho caso:

- Las variables $indice_precios_importaciones_lag1_lag2_lag3$ (-1.269) y $futuros_petroleo_wti_lag1_lag2_lag3$ (-1.175) tienen un impacto considerable y negativo sobre la predicción, evidenciado por valores SHAP negativos significativos. Esto indica que estas características tienden a reducir el valor predicho por el modelo en las observaciones seleccionadas.
- Variables como $credito_domestico_neto_sector_privado_banrep_lag1_lag2_lag3$ (+0.976) presentan una contribución positiva significativa, lo que evidencia que incrementan el valor predicho por el modelo.
- Factores como $TRM_lag1_lag2_lag3$ e $inflacion_total_lag1_lag2_lag3$ tienen impactos menores pero relevantes, indicando que sus variaciones influyen en menor medida, pero siguen contribuyendo al comportamiento global del modelo.
- $futuros_petroleo_wti_lag1_lag2_lag3$ y $indice_precios_importaciones_lag1_lag2_lag3$ son las principales responsables de reducir el valor predicho. Esto podría interpretarse como un impacto directo de estas variables sobre la dinámica económica modelada, posiblemente reflejando efectos adversos en la variable objetivo (por ejemplo, un aumento de precios internacionales o una caída en el petróleo que impacta negativamente el resultado).

Conclusiones

El modelo Ridge ajustado con la técnica de selección de características RFE demuestra un desempeño sobresaliente en términos de precisión, generalización, robustez e interpretabilidad. Las evaluaciones realizadas a través de métricas de desempeño, validaciones cruzadas, análisis de residuos, sensibilidad e interpretabilidad confirman que el modelo cumple con los estándares más rigurosos para problemas de regresión en contextos de alta dimensionalidad y multicolinealidad.

Desempeño predictivo:

- La optimización del hiperparámetro α = 0.1 permitió reducir el sobreajuste observado en la evaluación inicial, mejorando la generalización del modelo.
- El desempeño consistente entre el conjunto de entrenamiento y el conjunto de prueba, así como en la validación cruzada (R^2 promedio = 0.9663) respalda la calidad predictiva y la estabilidad del modelo.

Generalización y Robustez:

- La validación cruzada y la curva de aprendizaje evidenciaron un rendimiento estable incluso con tamaños de entrenamiento reducidos, lo que refuerza su capacidad para adaptarse a datos nuevos sin pérdida significativa de precisión.
- El análisis de sensibilidad muestra que el modelo es robusto a perturbaciones en las variables predictoras, con respuestas proporcionales y simétricas ante cambios incrementales en las características.
- El ajuste regularizado (α) contribuyó a un balance ideal entre la reducción de la varianza y el control del sesgo, evitando tanto el sobreajuste como la subestimación de los coeficientes.

Validez estadística:

- Distribución de residuos cercana a la normalidad: Los residuos son simétricos alrededor de cero y están distribuidos mayoritariamente en un rango estrecho, confirmando la ausencia de sesgo sistemático.
- Aleatoriedad de los residuos: No se observaron patrones claros en los residuos respecto a las predicciones, lo que sugiere que los errores son independientes y aleatorios.
- Homocedasticidad razonable: La variabilidad de los residuos es uniforme en el rango de valores predichos, reforzando la confiabilidad del modelo.

Interpretabilidad:

- Variables como "*índice de precios de importaciones*" y "*futuros del petróleo WTI*" son identificadas como factores determinantes clave, con impactos significativos y coherentes con la dinámica económica.
- El análisis local, mediante gráficos tipo *waterfall*, permite interpretar individualmente cada predicción, lo que hace que el modelo sea más transparente y explicable para su uso en la toma de decisiones.
- Esta capacidad de interpretar los resultados no solo valida la solidez teórica del modelo, sino que lo hace práctico y accesible en contextos aplicados.

Limitaciones observadas:

- Residuos extremos: Se observaron ligeros sesgos en los cuantiles extremos del gráfico Q-Q y algunos puntos residuales fuera del rango principal. Esto podría deberse a la incapacidad del modelo lineal para capturar relaciones no lineales complejas.
- Impacto moderado de algunas variables: Características como "*gasto social*" y "*EFFR_FED*" mostraron impactos limitados, lo que podría implicar que su relevancia en el modelo es secundaria.

7.5. Modelo Lasso RFE

Selección de características

La técnica RFE permitió identificar 7 características relevantes de entre un conjunto más amplio, seleccionadas por su importancia en el ajuste del modelo Lasso. Estas características abarcan las siguientes variables económicas:

```
Características seleccionadas por RFE: Index(['ftse_col_lag1_lag2_lag3',
       'credito_domestico_neto_sector_privado_banrep_lag1_lag2_lag3',
       'EFFR_FED_lag1_lag2_lag3', 'inflacion_total_lag1_lag2_lag3',
       'futuros_cafeC_lag1_lag2_lag3', 'futuros_petroleo_wti_lag1_lag2_lag3',
       'tasa_desempleo_lag1_lag2_lag3'],
      dtype='object')
```

***Ilustración 28**. Selección de características RFE modelo Lasso*

Entrenamiento inicial

En la implementación del modelo Lasso, la división de los datos se realizó mediante la función $train_test_split$ de $sklearn.model_selection$, con un 70% de los datos destinados al entrenamiento y un 30% a la prueba ($test_size = 0.3$). Este porcentaje fue seleccionado para garantizar un balance adecuado entre la capacidad del modelo para aprender patrones relevantes y su generalización sobre datos no vistos.

Además, se estableció un $random_state = 42$ para asegurar la reproducibilidad de los resultados y evitar sesgos asociados a una partición aleatoria no controlada. Esto permite que los experimentos sean replicables por otros investigadores bajo las mismas condiciones.

El modelo fue ajustado utilizando únicamente las características seleccionadas por RFE. Las métricas de desempeño iniciales reflejan un ajuste sólido, pero con indicios de sobreajuste:

- Conjunto de entrenamiento: El modelo alcanzó un coeficiente de determinación $R^2 = 0.9517$ lo que indica que explicó el 95.17% de la varianza en los datos de entrenamiento. Además, presentó un error cuadrático medio (MSE) de 0.4310, un error absoluto medio (MAE) de 0.4524 y un porcentaje de error absoluto medio (MAPE) de 9.68%. Estos valores reflejan errores bajos en el conjunto de entrenamiento, evidenciando un buen desempeño del modelo en la captura de la relación entre las variables predictoras y la variable objetivo.
- Conjunto de prueba: el modelo alcanzó un coeficiente de determinación $R^2 = 0.9237$ lo que indica una capacidad explicativa ligeramente menor en comparación con los datos de entrenamiento. Los valores del error cuadrático medio (MSE) fueron de 0.4484, del error absoluto medio (MAE) de 0.4817 y del porcentaje de error absoluto medio (MAPE) de 10.84%, evidenciando un ligero incremento en los errores respecto al conjunto de entrenamiento. Estos resultados son indicativos de un buen desempeño general del modelo, aunque con una leve disminución en precisión al evaluar nuevos datos.

Optimización de hiperparámetros Grid Search

La optimización de los hiperparámetros en el modelo Lasso se llevó a cabo mediante un GridSearchCV, utilizando una validación cruzada estratificada de $10\ folds\ (cv = 10)$. Esta estrategia permite una evaluación más robusta del rendimiento del modelo, mitigando posibles problemas de sobreajuste o subajuste.

El criterio de selección del mejor modelo se basó en la métrica R^2 $(scoring = 'r2')$, que evalúa la capacidad predictiva del modelo en diferentes subconjuntos de entrenamiento y validación. La grilla de hiperparámetros incluyó valores de regularización (alpha), iteraciones máximas (max_iter) y tolerancia de convergencia (tol), lo que garantizó una exploración exhaustiva del espacio de búsqueda.

Los mejores valores encontrados fueron:

- $\alpha = 0.01$: Un coeficiente de regularización moderado, que permite conservar la mayor parte de las características relevantes.
- $\max\ _iter\ = 1000$: Iteraciones suficientes para garantizar la convergencia del modelo.
- $tol = 0.0001$: Un nivel de tolerancia bajo para mejorar la precisión.

Para los resultados después de la optimización:

- Conjunto entrenamiento: El modelo mostró un incremento en su capacidad explicativa, alcanzando un coeficiente de determinación $R^2 = 0.9613$, lo que representa una mayor proporción de varianza explicada. Además, se observó una reducción en los errores absolutos y relativos, con un error cuadrático medio (MSE) de 0.3458, un error absoluto medio (MAE) de 0.4327 y un porcentaje de error absoluto medio (MAPE) de 9.31%.
- Conjunto prueba: el modelo alcanzó un coeficiente de determinación $R^2 = 0.9266$, indicando un incremento en su capacidad explicativa. Se observó una leve reducción en el error cuadrático medio (MSE), que fue de 0.4311, mientras que el error absoluto medio (MAE) y el porcentaje de error absoluto medio (MAPE) presentaron incrementos ligeros, alcanzando valores de 0.5126 y 12.36%, respectivamente. Estos cambios reflejan ajustes en los coeficientes de regularización, que, si bien mejoraron la capacidad explicativa, impactaron marginalmente en las métricas de error absoluto y relativo.

La optimización de hiperparámetros permitió mejorar la generalización del modelo, logrando un balance adecuado entre precisión y robustez.

Validación cruzada K-Fold

Una vez determinado el mejor alpha, se realizó una validación cruzada adicional con $K - Fold\ (5\ folds, cv = 5)$ para evaluar la estabilidad del modelo. Este proceso garantiza que el desempeño del modelo no dependa de una sola partición de datos.

Se utilizó validación cruzada con 5 particiones para evaluar la estabilidad y la capacidad generalizadora del modelo en distintos subconjuntos de datos. Las métricas promedio obtenidas fueron:

- $R^2 = 0.9266$: Alta consistencia en la varianza explicada.
- $MSE = 0.4288$, $MSE = 0.4754$: Errores promedio similares a los obtenidos en el conjunto de prueba.

Los resultados refuerzan la robustez del modelo optimizado, destacando su capacidad de generalizar a nuevos datos.

Curva de aprendizaje

El análisis de la curva de aprendizaje evidenció:

- Una alta capacidad predictiva del modelo ($R^2 > 0.9$) incluso con tamaños de entrenamiento reducidos.
- Una convergencia entre los scores de entrenamiento y validación a medida que aumenta el tamaño del conjunto de entrenamiento, lo que indica un balance adecuado entre sesgo y varianza.
- Una baja dispersión en el score de validación, lo que demuestra la estabilidad del modelo.

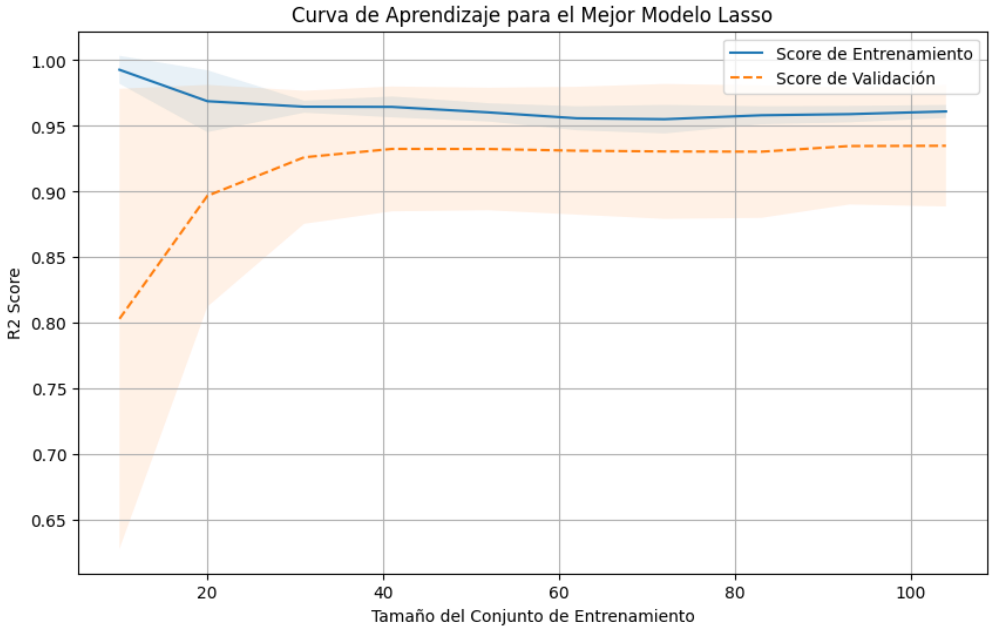

Ilustración 29. *Curva aprendizaje modelo Lasso*

Análisis de sensibilidad

```
=== Análisis de Sensibilidad para el Hiperparámetro Alpha ===
    Alpha       MSE       MAE      MAPE        R²
0    0.01  0.431052  0.512633  0.123559  0.926615
1    0.10  0.448403  0.481722  0.108367  0.923662
2    1.00  1.666146  0.926494  0.261086  0.716347
3   10.00  6.216713  1.991990  0.525037 -0.058365
4  100.00  6.216713  1.991990  0.525037 -0.058365
```

Ilustración 30. *Análisis sensibilidad modelo Lasso*

- $\alpha = 0.01$: Este valor ofreció el mejor desempeño en términos de precisión predictiva, con el menor error cuadrático medio ($MSE = 0.4311$) y el mayor valor $R^2 = 0.9266$. Esto indica que una regularización ligera es adecuada para este modelo, permitiendo capturar relaciones relevantes sin comprometer la precisión.
- $\alpha = 0.1$: Aunque el desempeño sigue siendo aceptable ($R^2 = 0.9237$), se observa un ligero incremento en MSE y una reducción en el error absoluto medio ($MAE = 0.4817$). Este valor tiende a penalizar más los coeficientes, lo que reduce la complejidad del modelo.
- $\alpha = 1.0$: Aumentar la regularización provoca una caída significativa en R^2 ($R^2 = 0.7163$) y un incremento considerable en MSE y $MAPE$. Esto indica una subestimación de los coeficientes, lo que afecta negativamente la capacidad predictiva.
- $\alpha = 10.0$ y $\alpha = 100.0$: En estos valores, el modelo presenta un desempeño deficiente ($R^2 < 0$), indicando que la penalización excesiva anula prácticamente la información relevante, resultando en predicciones inapropiadas.

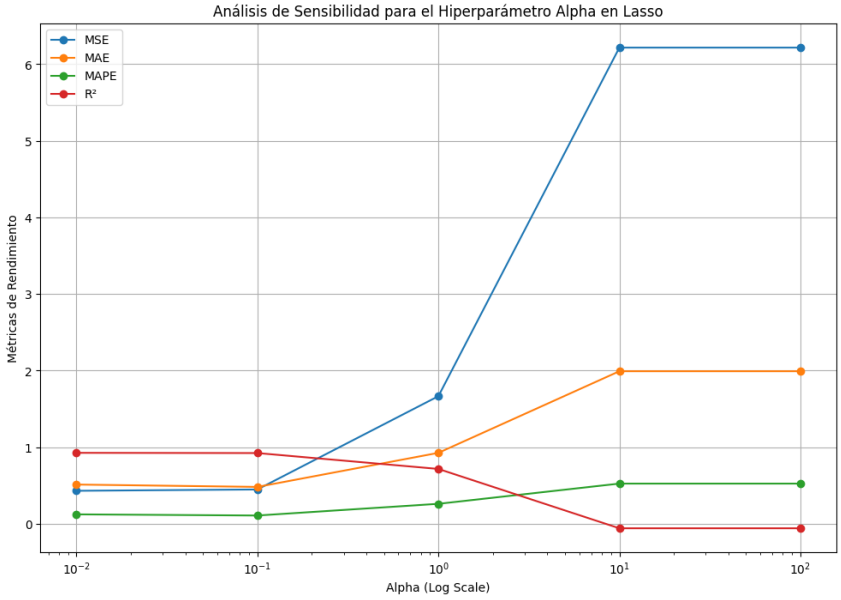

Ilustración 31. *Análisis sensibilidad hiperparámetro modelo Lasso*

El valor óptimo de regularización se encontró en $\alpha = 0.01$, donde se logró un balance ideal entre precisión y robustez. Valores mayores de α penalizan excesivamente los coeficientes, reduciendo la capacidad explicativa del modelo.

Análisis de residuos

Gráfico Q-Q:

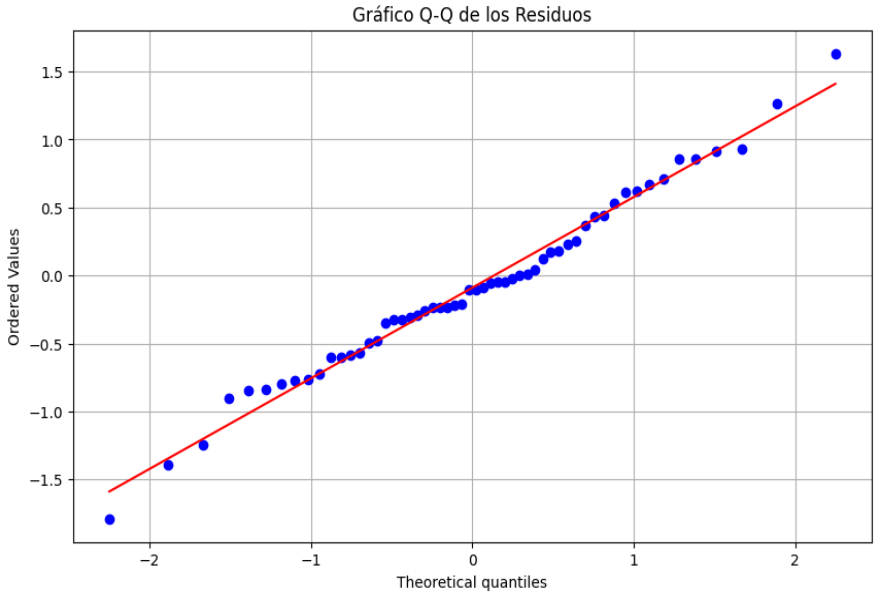

Ilustración 32. *Análisis de residuos modelo Lasso*

- La mayoría de los puntos del gráfico Q-Q se encuentran alineados con la línea de referencia roja, que representa los cuantiles teóricos de una distribución normal estándar.
- Esto sugiere que los residuos tienen un comportamiento que se aproxima a la normalidad, un atributo deseable para validar los supuestos estadísticos del modelo.
- Se observan ligeras desviaciones en los extremos (cuantiles bajos y altos). Las colas de la distribución presentan valores residuales algo más extremos de lo esperado bajo una distribución normal ideal.
- Estas desviaciones pueden deberse a la presencia de valores atípicos en los datos o a una ligera falta de ajuste del modelo en los casos extremos.
- Los puntos centrales, que corresponden a los valores residuales más frecuentes, muestran una excelente alineación con la línea de referencia. Esto indica que el modelo Lasso captura correctamente la relación entre las características seleccionadas y la variable objetivo para la mayoría de las observaciones.

Análisis de robustez

- Para $crédito_domestico_neto_sector_privado_banrep_lag1_lag2_lag3$ el incremento

del 1% en esta variable resultó en un aumento de 0.003668 unidades en la predicción. Este resultado muestra que esta variable tiene un impacto positivo significativo en la predicción de la tasa de intervención política monetaria.

- La $inflación_total_lag1_lag2_lag3$ un aumento del 1% en esta variable redujo la predicción en 0.006445 unidades, indicando un impacto negativo considerable.
- La variable $futuros_petroleo_wti_lag1_lag_2_lag3$ mostró un efecto negativo prominente, donde un incremento del 1% disminuyó la predicción en 0.006932 unidades.
- $Fste_col_lag1_lag2_lag3$ el impacto de un incremento del 1% fue mínimo, reduciendo la predicción en solo 0.000279 unidades, lo que sugiere que esta variable tiene una influencia limitada.

Los efectos de los decrementos fueron simétricos en magnitud y opuestos en dirección a los efectos de los incrementos. Esto demuestra que el modelo responde de manera lineal y consistente a las variaciones en las características. Por consiguiente, El modelo Lasso presenta una robustez adecuada, con predicciones estables frente a perturbaciones pequeñas.

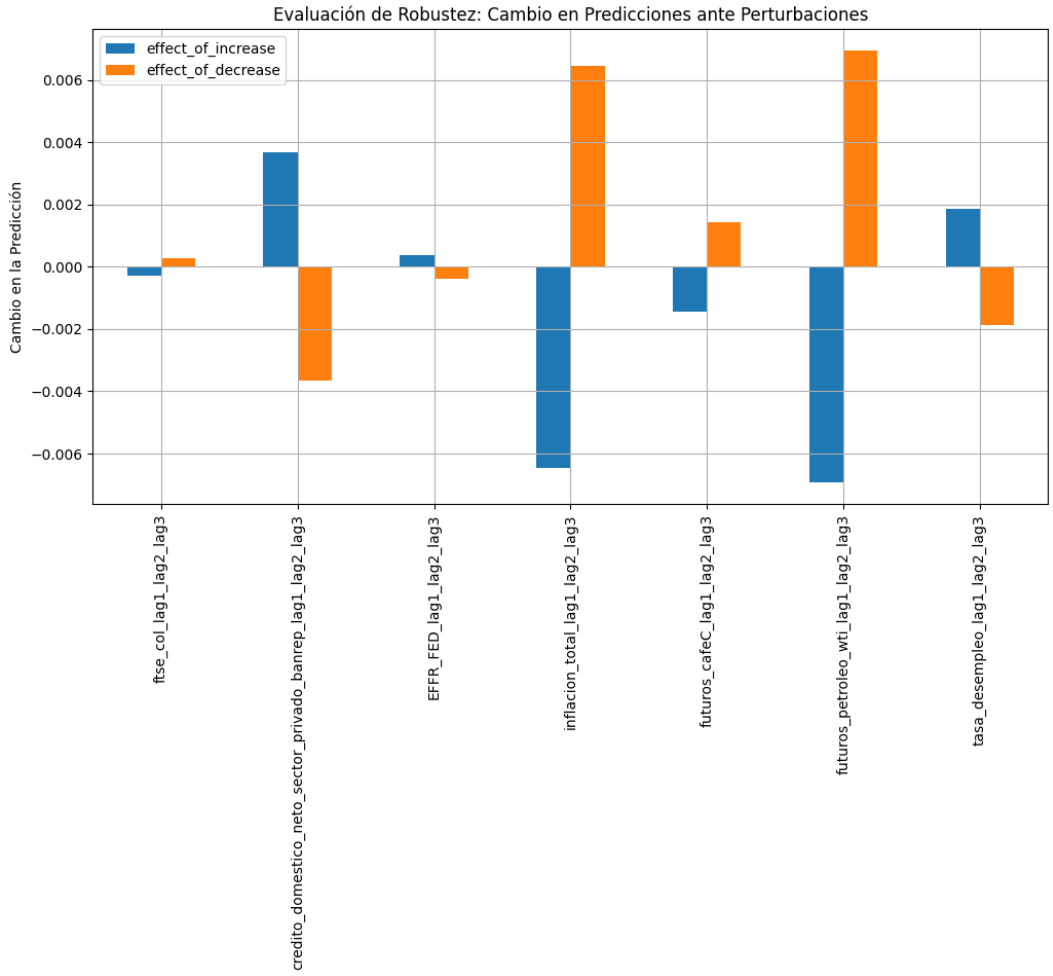

Ilustración 33. *Análisis robustez modelo Lasso*

Análisis de interpretabilidad SHAP

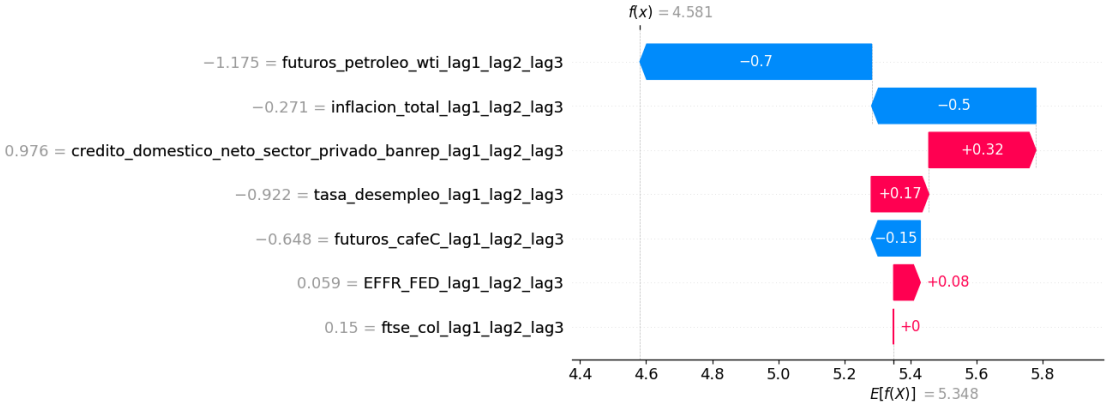

Ilustración 34. *Análisis interpretabilidad SHAP modelo Lasso*

- La variable $futuros_petroleo_wti_lag1_lag2_lag3$ con la contribución más fuerte y negativa en esta instancia específica tiene un valor SHAP de -1.175, lo que reduce significativamente el valor predicho. Por su parte, las variables $tasa_desempleo_lag1_lag2_lag3$ y $futuros_cafeC_lag1_lag2_lag3$ también contribuyen de manera negativa, aunque con una magnitud menor en comparación con la influencia del petróleo. Estas contribuciones negativas destacan el impacto de estas características en la disminución de la predicción para el caso analizado.
- Entre las contribuciones positivas más significativas, la variable $credito_domestico_neto_sector_privado_banrep_lag1_lag2_lag3$ destaca con un valor SHAP de +0.976, aumentando considerablemente la predicción del modelo en esta instancia específica. Por otro lado, la variable $EFFR_FED_lag1_lag2_lag3$ también presenta una contribución positiva, aunque de menor magnitud, manteniéndose consistente con su comportamiento global observado en el modelo. Estas variables resaltan como factores clave en el incremento de la predicción.
- La predicción para esta observación fue de 4.581 unidades, generada como la suma de las contribuciones SHAP de todas las características más el valor base del modelo (5.348 unidades).

Conclusiones

Desempeño predictivo:

Tras la optimización de hiperparámetros, el modelo alcanzó métricas mejoradas en términos de R^2 (0.9613 en entrenamiento y 0.9266 en prueba) y una reducción del error cuadrático medio, aunque con ligeros incrementos en MAE y MAPE en el conjunto de prueba. Esto demuestra que la regularización optimizada balanceó la precisión y la complejidad del modelo.

Generalización y Robustez:

- Validación cruzada: Con un R^2 promedio de 0.9266, las métricas fueron consistentes en múltiples particiones del conjunto de datos. Esto refuerza la estabilidad del modelo y su capacidad para generalizar a diferentes subconjuntos de datos.
- Curva de aprendizaje: La convergencia entre los scores de entrenamiento y validación, junto con la baja dispersión en el score de validación, indica un balance adecuado entre sesgo y varianza. Esto demuestra que el modelo es estable incluso con tamaños reducidos del conjunto de entrenamiento.
- En cuanto a robustez, el análisis de sensibilidad frente a perturbaciones pequeñas en las características muestra que el modelo responde de manera proporcional y lineal, con cambios simétricos en las predicciones. Esto valida que el modelo es consistente y no es excesivamente sensible a pequeñas variaciones en las variables de entrada

Validez estadística:

- Homocedasticidad: La dispersión de los residuos es uniforme a lo largo de las predicciones, indicando una varianza constante de los errores.
- Independencia de los errores: La aleatoriedad en la distribución de los residuos refuerza la suposición de independencia.
- Normalidad de los errores: El gráfico Q-Q muestra que los residuos siguen una distribución aproximadamente normal, con ligeras desviaciones en las colas, probablemente asociadas con valores atípicos. Este comportamiento es aceptable para el análisis.

Interpretabilidad:

- Variables como "*futuros del petróleo wti*" y "*crédito domestico neto sector privado*" emergieron como los determinantes principales del modelo, con contribuciones significativas y consistentes tanto en análisis globales como locales.
- La suma de las contribuciones SHAP (positivas y negativas) reflejó con precisión el valor predicho, lo que confirma la consistencia del modelo y la validez de las inferencias. La capacidad del modelo para identificar factores relevantes en la dinámica económica subyacente lo posiciona como una herramienta interpretable y útil para la toma de decisiones.

Limitaciones observadas:

- Residuos extremos: Ligeras desviaciones en los residuos extremos.
- Si bien el ajuste de α mejoró las métricas globales, se observaron incrementos en los errores absolutos (MAE y MAPE) en el conjunto de prueba tras la optimización. Esto puede deberse a una ligera subestimación en ciertas predicciones, particularmente en los extremos del rango de valores.
- Variables como el petróleo y la inflación tuvieron un impacto desproporcionado en las predicciones, lo que podría reflejar cierta vulnerabilidad del modelo a cambios significativos en estas características.

7.6. Modelo Random Forest RFE

Selección de características

La técnica de selección recursiva de características (RFE), utilizando el modelo Random Forest Regressor como estimador base, permitió identificar las 7 características más relevantes entre el conjunto completo de variables predictoras. Estas características incluyen:

```
Características seleccionadas por RFE (Random Forest): Index(['activos_netos_exterior_banrep_lag1_lag2_lag3',
       'IPP_produccion_ofertainterna_lag1_lag2_lag3',
       'dtf_efectiva_anual_lag1_lag2_lag3', 'EFFR_FED_lag1_lag2_lag3',
       'inflacion_total_lag1_lag2_lag3',
       'oferta_monetaria_agregada_lag1_lag2_lag3',
       'PIB_nominal_lag1_lag2_lag3'],
      dtype='object')
```

Ilustración 35. *Selección características RFE modelo Random Forest*

Entrenamiento inicial

- En el conjunto de entrenamiento, el modelo alcanzó un coeficiente de determinación $R^2 = 0.9948$ lo que indica que explicó el 99.48% de la varianza en los datos, reflejando un ajuste casi perfecto. Además, presentó un error cuadrático medio (MSE) de 0.0468, un error absoluto medio (MAE) de 0.1182 y un porcentaje de error absoluto medio (MAPE) de 2.51%. Este bajo valor de MAPE evidencia que el modelo realiza predicciones muy precisas en promedio, con errores relativos mínimos, lo que refuerza su eficacia en el ajuste a los datos de entrenamiento.
- En el conjunto de prueba, el modelo obtuvo un coeficiente de determinación $R^2 = 0.9948$, ligeramente inferior al del conjunto de entrenamiento, pero aún indicando que explicó el 97.09% de la varianza en los datos. Los errores calculados reflejan un buen desempeño, con un error cuadrático medio (MSE) de 0.1707, un error absoluto medio (MAE) de 0.2478 y un porcentaje de error absoluto medio (MAPE) de 6.02%. Este bajo valor de MAPE evidencia que el modelo mantiene un error relativo reducido incluso al evaluar datos no vistos, confirmando su capacidad de generalización.
- En el código proporcionado, la división de los datos en conjuntos de entrenamiento y prueba se realiza utilizando la función $train_test_split$ de la biblioteca scikit-learn. Este procedimiento es estándar en la práctica de machine learning y se implementa de la siguiente manera:
    - $test_size = 0.3$: Esto indica que el 30% de los datos se reservan para pruebas, mientras que el 70% restante se utiliza para entrenamiento. Este porcentaje es común en la literatura y proporciona un equilibrio adecuado entre la cantidad de datos disponibles para entrenar el modelo y la evaluación del mismo.
    - $random_state = 42$: Este parámetro asegura que la división de los datos sea reproducible. Al fijar una semilla aleatoria, se garantiza que los resultados sean consistentes en diferentes ejecuciones del código.

Optimización de hiperparámetros Grid Search

Se utilizó Grid Search con Validación Cruzada (GridSearchCV) para optimizar los hiperparámetros del modelo de Random Forest. Este procedimiento se implementa de la siguiente manera:

- $cv = 10$: Esto indica que se utiliza una validación cruzada de 10 folds. Cada fold divide los datos de entrenamiento en 10 partes, utilizando 9 para entrenar y 1 para validar, repitiendo el proceso 10 veces. Esto garantiza que todos los datos se utilicen tanto para entrenamiento como para validación, proporcionando una estimación robusta del rendimiento del modelo.
- $scoring = 'r2'$: La métrica utilizada para evaluar el rendimiento del modelo es el coeficiente de determinación (R^2), que es una métrica común para problemas de regresión.
- $n_jobs = -1$: Esto permite paralelizar el proceso de búsqueda de hiperparámetros, utilizando todos los núcleos disponibles en la máquina para acelerar el cálculo.

El mejor conjunto de hiperparámetros encontrado fue:

- $n_estimators$: 50.
- $\max\ _depth$: *None* (sin restricciones en la profundidad de los árboles).
- $\min\ _samples_split$: 2.
- $\min\ _samples_leaf$: 1.

Resultados del modelo optimizado:

- Conjunto de entrenamiento: R^2 0.9949, con un MSE de 0.0451, MAE de 0.1284, y un MAPE de 2.79%.
- Conjunto de prueba: R^2 0.9721, con un MSE de 0.1640, MAE de 0.2446, y un MAPE de 6.07%.

La optimización logró una ligera mejora en R^2 y una reducción en el MSE, lo que evidencia un balance óptimo entre ajuste y generalización. Los cambios en las métricas son mínimos, lo que sugiere que el modelo ya tenía una configuración adecuada antes de la optimización.

Validación cruzada K-Fold

Además, después de encontrar el mejor modelo con Grid Search, se realiza una validación cruzada adicional con K-Fold para evaluar el rendimiento del modelo de manera independiente:

- $n_splits = 5$: Se utiliza una validación cruzada de 5 folds para evaluar el modelo final.
- $shuffle = True$: Los datos se mezclan antes de dividirlos en folds, lo que ayuda a garantizar que la evaluación no esté sesgada por el orden de los datos.
- $random_state = 42$: Nuevamente, se fija una semilla para garantizar reproducibilidad.

La validación cruzada, realizada con 5 particiones (K-Fold), evaluó la estabilidad y robustez del modelo optimizado en diferentes subconjuntos de datos. Los resultados promedio obtenidos reflejan un desempeño sólido, con un coeficiente de determinación $R^2 = 0.9549$, lo que confirma una alta capacidad explicativa, un error cuadrático medio (MSE) de 0.3648 y un error

absoluto medio (MAE) de 0.3342. La consistencia de estas métricas a lo largo de las particiones valida la robustez del modelo y su capacidad de generalización en diferentes muestras.

Se utilizó validación cruzada externa (mediante GridSearchCV y KFold) en lugar de confiar únicamente en el método OOB (Out of Bag). Esta decisión se basó en mayor robustez de evaluación, ajuste de hiperparámetros con el uso de Grid Search que permitió explorar de manera sistemática el espacio de hiperparámetros y seleccionar la combinación óptima basada en el rendimiento promedio en los folds de validación. Asimismo, comparabilidad con otros modelos garantizando que la evaluación fuera consistente con los demás modelos.

Curva de aprendizaje

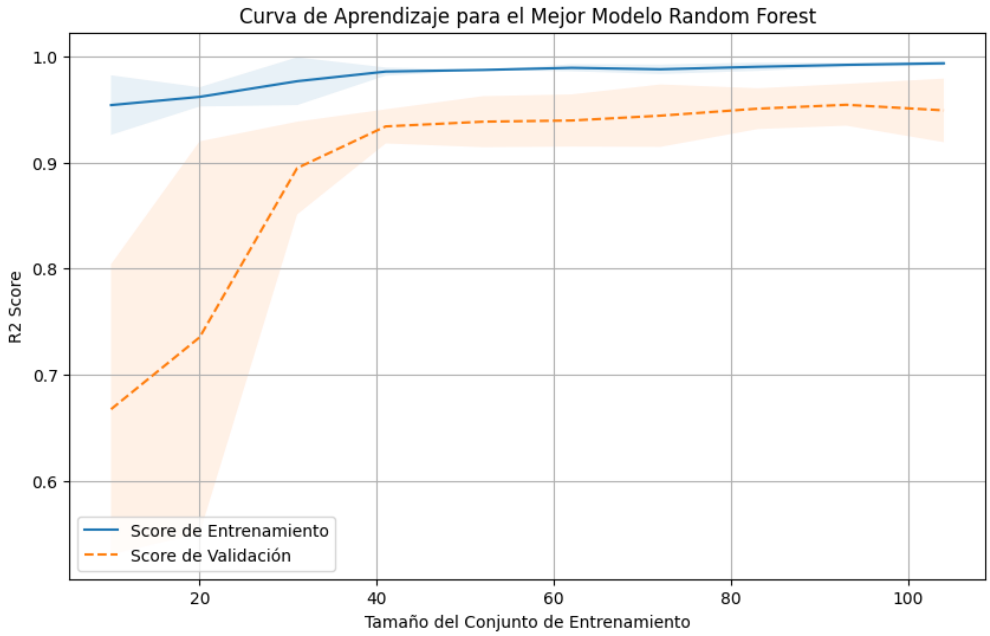

***Ilustración 36**. Curva aprendizaje modelo Random Forest*

- Alta capacidad predictiva desde tamaños de entrenamiento pequeños: Con solo el 20% de los datos de entrenamiento, el modelo ya alcanzó un $R^2 > 0.90$ en el conjunto de validación.
- Convergencia entre los puntajes de entrenamiento y validación: A medida que aumenta el tamaño del conjunto de entrenamiento, la diferencia entre ambos puntajes se reduce, indicando un balance adecuado entre sesgo y varianza.
- Estabilidad del modelo: La baja dispersión en los puntajes de validación evidencia que el modelo es consistente y robusto.

Análisis de sensibilidad

Para el hiperparámetro $n_estimators$ las métricas de desempeño del modelo son los siguientes:

- Con $n_estimators = 10$, el modelo obtuvo un $R^2 = 0.9608$, indicando una capacidad explicativa alta. Sin embargo, el Error Cuadrático Medio (MSE) de 0.2305 y el Error Absoluto Medio (MAE) de 0.2930 reflejan un menor ajuste en comparación con configuraciones con más árboles, mientras que el Porcentaje de Error Absoluto Medio (MAPE) fue del 7.02%, el mayor registrado. Estos resultados sugieren que un número reducido de árboles limita la capacidad del modelo para capturar patrones complejos en los datos, aumentando los errores en las predicciones.
- Con $n_estimators = 50$, el modelo alcanzó su mejor desempeño, logrando un coeficiente de determinación $R^2 = 0.9721$, el valor más alto registrado en el análisis. Además, presentó un error cuadrático medio (MSE) de 0.1640, significativamente menor que con 10 árboles, un error absoluto medio (MAE) de 0.2446 y un porcentaje de error absoluto medio (MAPE) de 6.07%, indicando una mejora en la precisión relativa. Estos resultados reflejan el mejor balance entre capacidad predictiva y complejidad computacional, capturando de manera efectiva las relaciones subyacentes en los datos.
- Con $n_estimators = 100$, el modelo obtuvo un coeficiente de determinación $R^2 = 0.9709$, levemente menor que con 50 árboles. Además, presentó un error cuadrático medio (MSE) de 0.1707, ligeramente superior, un error absoluto medio (MAE) de 0.2478 y un porcentaje de error absoluto medio (MAPE) de 6.02%, comparable al caso de 50 árboles. Aunque este desempeño sigue siendo competitivo, la diferencia marginal respecto a n_estimators = 50 no justifica un incremento adicional en el número de árboles desde la perspectiva de costo computacional.
- Con $n_estimators = 200$, el modelo obtuvo un coeficiente de determinación $R^2 = 0.9683$, mostrando una reducción progresiva en su capacidad explicativa. Además, presentó un error cuadrático medio (MSE) de 0.1860, el segundo valor más alto registrado, un error absoluto medio (MAE) de 0.2532 y un porcentaje de error absoluto medio (MAPE) de 6.10%, sin mejoras evidentes respecto a configuraciones más simples. Estos resultados sugieren que, a medida que aumenta el número de árboles, los beneficios marginales en precisión disminuyen, probablemente debido a un potencial sobreajuste del modelo.
- Con $n_estimators = 500$, el modelo alcanzó un coeficiente de determinación $R^2 = 0.9700$, un valor similar al observado con 100 árboles. El error cuadrático medio (MSE) fue de 0.1759, indicando un rendimiento intermedio, mientras que el error absoluto medio (MAE) fue de 0.2477 y el porcentaje de error absoluto medio (MAPE) de 5.94%, el más bajo registrado en términos de error relativo, aunque marginalmente mejor que configuraciones con menos árboles. Estos resultados sugieren que aumentar el número de árboles mejora ligeramente el desempeño predictivo, pero a costa de un incremento significativo en los costos computacionales, lo que limita su justificación práctica.

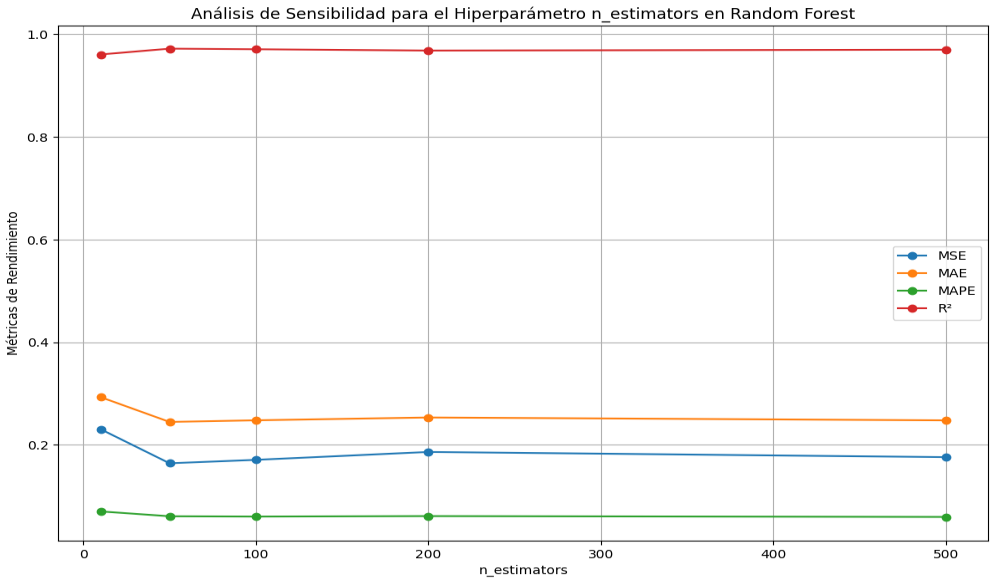

Ilustración 37. *Análisis sensibilidad hiperparámetro modelo Random Forest*

El valor óptimo de $n_estimators$ se encuentra en 50, donde el modelo logra el mayor R^2 y los menores errores absolutos relativos. Incrementar el número de árboles más allá de este valor no aporta mejoras sustanciales en las métricas, indicando que el modelo alcanza una saturación en su capacidad predictiva.

Análisis de residuos

Gráfico Q-Q:

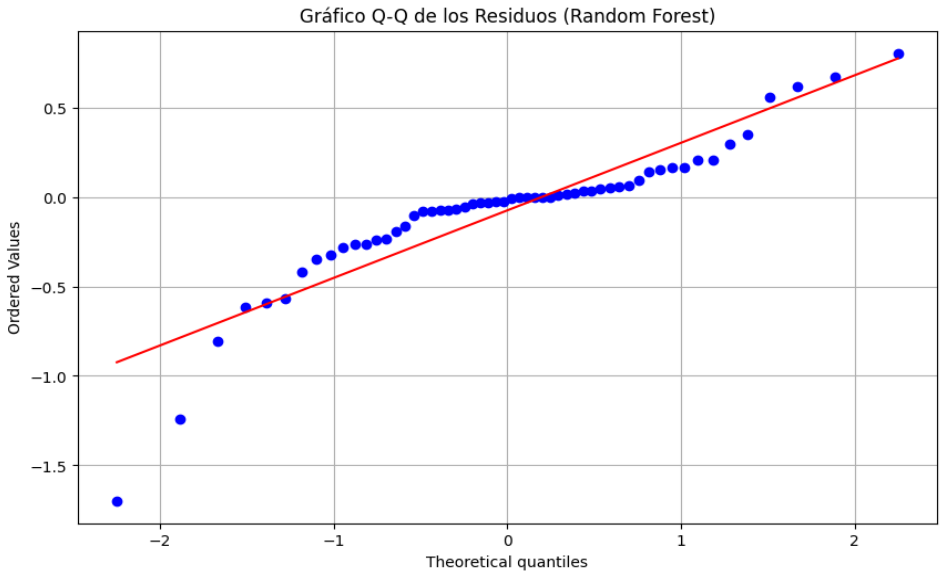

Ilustración 38. *Análisis residuos modelo Random Forest*

- Alineación central con la línea teórica: Los residuos más frecuentes se alinean bien con la línea de referencia roja, confirmando que, para la mayoría de las observaciones, los errores son aproximadamente normales.
- Desviaciones en las colas: Se observan ligeras desviaciones en los extremos (cuantiles bajos y altos). Las colas presentan valores residuales más extremos de lo esperado bajo una distribución normal.
- Simetría general: A pesar de las desviaciones en las colas, la simetría del gráfico confirma la ausencia de sesgos significativos en la distribución de los residuos.

Análisis de robustez

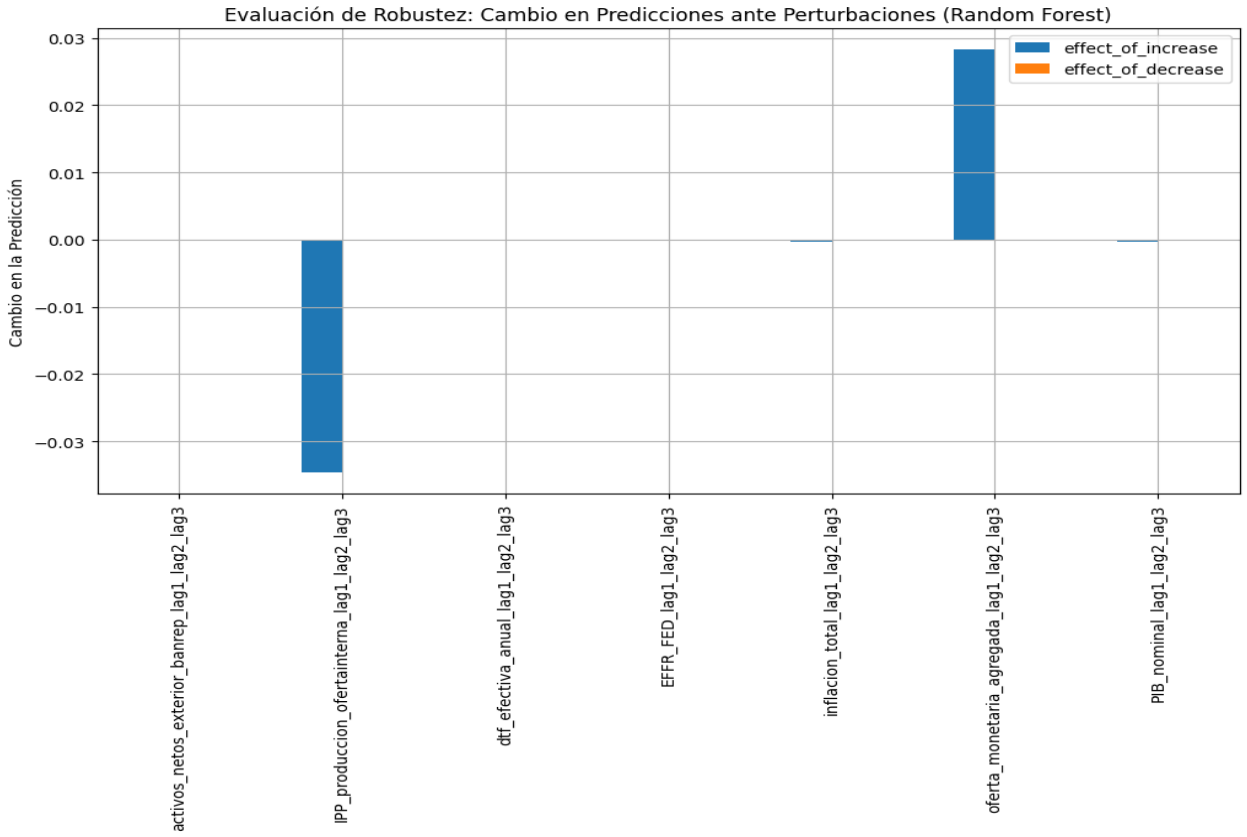

Ilustración 39. *Análisis robustez modelo Random Forest*

- $IPP_produccion_ofertainterna_lag1_lag2_lag3$*:* Esta variable presentó una disminución notable en la predicción con el incremento del 1%, específicamente un cambio de −0.0347.
- $oferta_monetaria_agregada_lag1_lag2_lag3$: En contraste, el incremento de esta variable resultó en un aumento positivo en la predicción (+0.0282), lo que destaca su relevancia dentro de las predicciones del modelo.
- Variables como $activos_netos_exterior_banrep_lag1_lag2_lag3$, $dtf_efectiva_anual_lag1_lag2_lag3$ y $EFFR_FED_lag1_lag2_lag3$ no mostraron

cambios en la predicción del modelo ante perturbaciones. Esto indica que estas características tienen una menor importancia o que el modelo ha aprendido a considerarlas redundantes en la predicción.

- Si el modelo depende en gran medida de dos variables, existe el riesgo de que los cambios estructurales o errores en estas variables afecten desproporcionadamente las predicciones del modelo.

Análisis de interpretabilidad SHAP

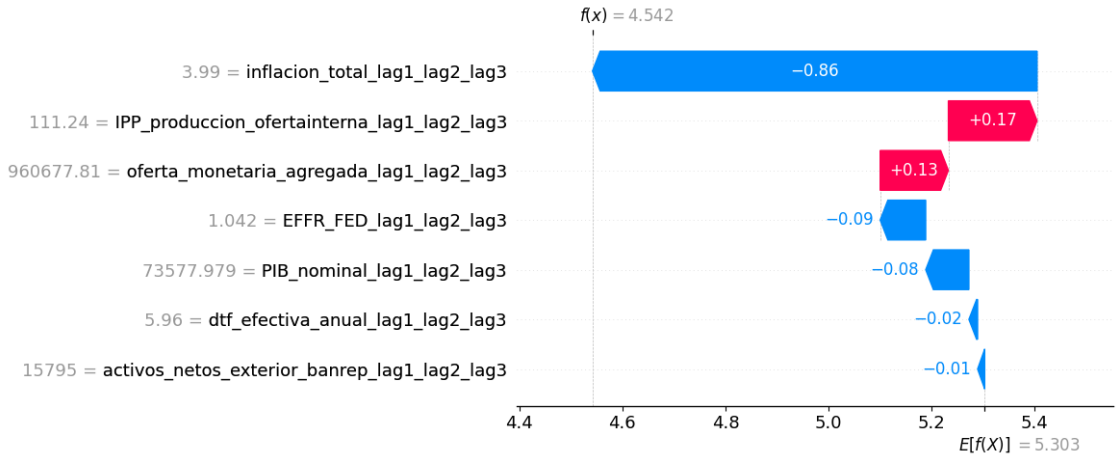

***Ilustración 40**. Análisis interpretabilidad SHAP modelo Random Forest*

- La $inflación_total_lag1_lag2_lag3$ fue la característica más influyente, con una contribución promedio negativa significativa. Esto indica que mayores niveles de inflación tienden a reducir la predicción de la tasa objetivo.
- Variables como $IPP_produccion_ofertainterna_lag1_lag2_lag3$ y $oferta_monetaria_agregada_lag1_lag2_lag3$ también desempeñaron un papel importante, aunque sus contribuciones fueron positivas, sugiriendo que incrementos en estas variables están asociados con valores más altos de la tasa objetivo.
- Otras características, como $EFFR_FED_lag1_lag2_lag3$ y $PIB_nominal_lag1_lag2_lag3$, presentaron contribuciones más moderadas, pero no despreciables.
- Valores altos de $inflación_total_lag1_lag2_lag3$ están consistentemente asociados con contribuciones negativas, destacando su relación inversa con la variable objetivo.
- En contraste, valores altos de $IPP_produccion_ofertainterna_lag1_lag2_lag3$ generan contribuciones positivas, lo que refuerza su impacto positivo en la predicción.
- Para la observación seleccionada, el modelo predijo una tasa de intervención de 4.542 unidades, resultado de la suma de las contribuciones SHAP de las características y el valor base del modelo, que corresponde a la media de las predicciones $(E[f(x)] = 5.303)$. Este

desglose permite interpretar cómo las características específicas de la observación influyeron en la predicción, ajustando el valor base para reflejar el resultado final del modelo.

Conclusiones

- **Desempeño predictivo:** El modelo Random Forest alcanzó un desempeño predictivo sobresaliente tanto en el conjunto de entrenamiento como en el conjunto de prueba. Los valores del coeficiente de determinación superiores al 97% reflejan una capacidad casi perfecta para explicar la varianza de los datos. Esto se acompaña de errores absolutos y relativos muy bajos, con un MAPE del 6.07% en el conjunto de prueba después de la optimización de hiperparámetros, lo que indica una precisión consistente en las predicciones. El bajo MSE y MAE refuerzan la capacidad del modelo para capturar patrones complejos en los datos, minimizando errores significativos.
- **Capacidad de generalización:** La evaluación en el conjunto de prueba y mediante validación cruzada confirma una alta capacidad de generalización del modelo. Aunque el desempeño en validación cruzada (R^2 = 0.9549) es ligeramente inferior al del conjunto de prueba, este resultado se encuentra dentro de un rango aceptable, demostrando que el modelo no está sobreajustado a los datos de entrenamiento. La convergencia observada entre las métricas de entrenamiento y validación en la curva de aprendizaje sugiere un adecuado balance entre sesgo y varianza, con mínima pérdida de precisión al evaluar datos no vistos.
- **Robustez:** El análisis de sensibilidad del modelo frente a perturbaciones en las variables clave reveló que ciertas características, como $IPP_produccion_ofertainterna_lag1_lag2_lag3$ y $oferta_monetaria_agregada_lag1_lag2_lag3$, tienen un impacto significativo en las predicciones. En contraste, variables como $activos_netos_exterior_banrep_lag1_lag2_lag3$ y $dtf_efectiva_anual_lag1_lag2_lag3$ mostraron cambios insignificantes ante perturbaciones, lo que sugiere que el modelo ha aprendido a depender principalmente de las variables más relevantes. Si bien esto refuerza la robustez del modelo frente a ruido en las características secundarias, también resalta una posible vulnerabilidad ante cambios estructurales en las variables dominantes.
- El análisis de sensibilidad del hiperparámetro $n_estimators$ identificó que 50 árboles ofrecieron el mejor equilibrio entre precisión y eficiencia computacional. Incrementar el número de árboles más allá de este valor proporcionó mejoras marginales, lo que evidencia una saturación en el rendimiento del modelo.
- **Validez estadística:** El análisis de residuos respalda la validez estadística del modelo. Los residuos se distribuyen de manera simétrica alrededor de cero y muestran independencia respecto a las predicciones, cumpliendo con los supuestos fundamentales de independencia y homocedasticidad. Aunque se observaron ligeras desviaciones en las colas del gráfico Q-Q, estas no afectan significativamente el ajuste global del modelo. La concentración de residuos

en un rango estrecho sugiere que el modelo mantiene un error bajo en la mayoría de las instancias, con valores atípicos limitados.

- **Interpretabilidad:** El análisis SHAP proporcionó información valiosa sobre cómo las características individuales contribuyen a las predicciones del modelo. A nivel global, variables como $inflación_total_lag1_lag2_lag3$, $IPP_produccion_ofertainterna_lag1_lag2_lag3$ y $oferta_monetaria_agregada_lag1_lag2_lag3$ fueron las más influyentes, alineándose con expectativas económicas. A nivel local, los gráficos waterfall permitieron desglosar las predicciones individuales, mostrando cómo cada característica ajusta el valor base para reflejar el resultado final del modelo.
- **Limitaciones:** La alta influencia de unas pocas características, como la inflación y la producción interna, sugiere que el modelo podría ser vulnerable a cambios estructurales en estas variables. Aunque esto refuerza su interpretabilidad, también indica un riesgo si estas características se ven afectadas por factores externos o ruido.

7.7. Modelo Gradient Boosting – XGBoost RFE

Selección de características

El proceso de selección recursiva de características (RFE) aplicado al modelo XGBoost identificó las 7 características más relevantes para predecir la tasa de intervención política monetaria. Estas características son:

```
Características seleccionadas por RFE: Index(['dow_jones_lag1_lag2_lag3', 'tasa_desempleo_lag1_lag2_lag3',
       'indice _terminos_intercambio_lag1_lag2_lag3', 'nasdaq_lag1_lag2_lag3',
       'EFFR_FED_lag1_lag2_lag3',
       'IPP_produccion_ofertainterna_lag1_lag2_lag3',
       'inflacion_total_lag1_lag2_lag3'],
      dtype='object')
```

***Ilustración 41**. Selección características modelo XGBoost*

Entrenamiento inicial

- En el conjunto de entrenamiento, el modelo alcanzó un coeficiente de determinación (R^2) de 0.9999, lo que indica que explicó el 99.99% de la varianza en los datos, reflejando un ajuste prácticamente perfecto. Además, presentó un error cuadrático medio (MSE) de 8.61e-07 y un error absoluto medio (MAE) de 0.00065, confirmando una precisión predictiva excepcional. El porcentaje de error absoluto medio (MAPE) fue de apenas 0.015%, evidenciando una desviación insignificante entre las predicciones y los valores reales, lo que demuestra la eficacia del modelo en capturar las relaciones subyacentes en los datos de entrenamiento.

- En el estudio, la división de los datos en conjuntos de entrenamiento y prueba se realizó utilizando el método $train_test_split$ de la librería scikit-learn. En este caso, se asignó el 80% de los datos para entrenamiento (X_train, y_train) y el 20% para prueba (X_test, y_test). El parámetro $random_state = 42$ asegura que la división sea reproducible, lo cual es crucial para la consistencia de los resultados.
- En el conjunto de prueba, el modelo alcanzó un coeficiente de determinación R^2 de 0.9806, indicando que explicó el 98.06% de la varianza en los datos no vistos. Presentó un error cuadrático medio (MSE) de 0.1139, reflejando un ligero incremento respecto al conjunto de entrenamiento, y un error absoluto medio (MAE) de 0.2048, evidenciando errores bajos incluso en datos nuevos. El porcentaje de error absoluto medio (MAPE) fue de 4.93%, lo que confirma un error relativo reducido y la capacidad del modelo para generalizar de manera efectiva a datos no utilizados durante su entrenamiento.

El modelo inicial presentó un ajuste extremadamente alto en los datos de entrenamiento, reflejando potencial sobreajuste. Sin embargo, el desempeño en el conjunto de prueba fue excelente, confirmando una generalización adecuada.

Optimización de hiperparámetros Grid Search

Para ajustar los hiperparámetros del modelo XGBoost, se utilizó la técnica de validación cruzada ($cross - validation$) en combinación con Grid Search. Este enfoque permite evaluar el rendimiento del modelo en diferentes subconjuntos de los datos de entrenamiento, asegurando que el modelo no esté sobreajustado y que los hiperparámetros seleccionados sean robustos.

En este caso, se utilizó una validación cruzada de 5 folds (cv=5), lo que significa que los datos de entrenamiento se dividieron en 5 subconjuntos, y el modelo se entrenó y validó 5 veces, utilizando cada vez un subconjunto diferente como conjunto de validación y el resto como conjunto de entrenamiento. Esto permite una evaluación más robusta de los hiperparámetros.

La optimización mediante Grid Search ajustó los hiperparámetros clave del modelo para mejorar su desempeño, logrando los siguientes valores óptimos:

- $learning_rate$: 0.1
- $n_estimators$: 200
- $\max\ _depth$: 3
- En el conjunto de entrenamiento, el modelo alcanzó un coeficiente de determinación (R^2) de 0.9998, manteniendo un ajuste extremadamente alto. El error cuadrático medio (MSE) fue de 0.00158, reflejando un ajuste ligeramente más conservador en comparación con el modelo inicial, mientras que el error absoluto medio (MAE) fue de 0.0267. El porcentaje de error absoluto medio (MAPE) se ubicó en 0.65%, mostrando un pequeño incremento en el error

relativo, aunque dentro de un rango aceptable, lo que confirma que el modelo sigue capturando de manera eficiente las relaciones en los datos de entrenamiento.

- En el conjunto de prueba, el modelo alcanzó un coeficiente de determinación (R^2) de 0.9812, reflejando un ligero incremento respecto al modelo inicial y una mejora en su capacidad explicativa. El error cuadrático medio (MSE) fue de 0.1106, mostrando una reducción en el error absoluto, mientras que el error absoluto medio (MAE) se situó en 0.2415. El porcentaje de error absoluto medio (MAPE) fue de 5.24%, confirmando la precisión relativa del modelo incluso tras la optimización y su capacidad para generalizar de manera efectiva en datos no vistos.

El proceso de optimización logró un mejor balance entre capacidad explicativa y generalización. Aunque los cambios fueron sutiles, las métricas indican un desempeño consistente y robusto.

Validación cruzada K-Fold

El modelo XGBoost incluye mecanismos internos de validación, como la validación $out - of - bag$ (OOB) y la validación cruzada interna. Sin embargo, en este estudio, se optó por utilizar la validación cruzada externa mediante GridSearchCV para tener un control más explícito sobre el proceso de ajuste de hiperparámetros y para asegurar que los resultados sean comparables con otros modelos que no tienen mecanismos de validación interna.

Se realizó validación cruzada con 5 particiones, evaluando la estabilidad y robustez del modelo en diferentes subconjuntos de datos. Los resultados promedio fueron:

- R^2: 0.9474, confirmando una alta capacidad explicativa en diferentes particiones.
- MSE: 0.4007, un incremento esperado debido a la variabilidad inherente en la validación cruzada.
- MAE: 0.3824, mostrando un desempeño predecible en diferentes subconjuntos.

Los resultados confirman que el modelo es robusto y mantiene un desempeño alto en distintas muestras de datos, reduciendo el riesgo de sobreajuste.

Curva de aprendizaje

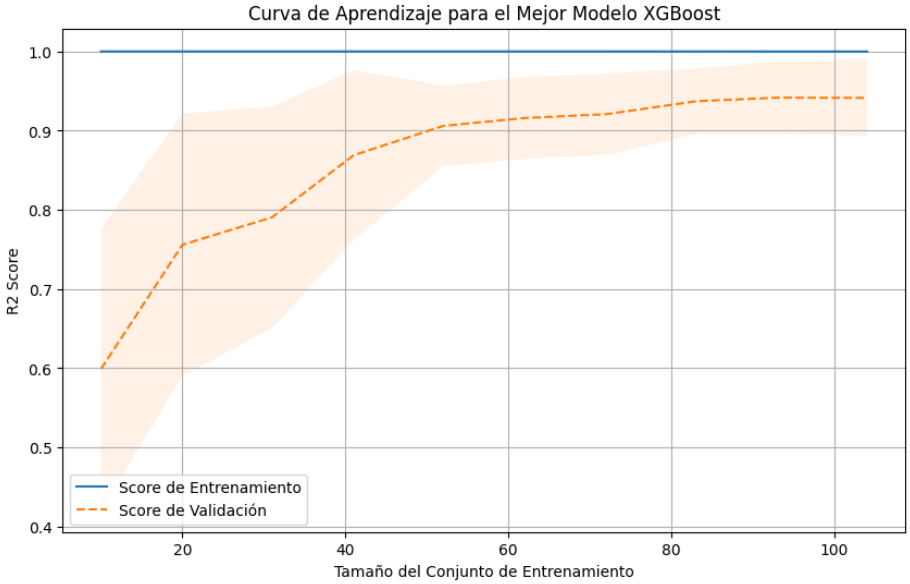

Ilustración 42. *Curva aprendizaje modelo XGBoost*

- Conforme aumenta el tamaño del conjunto de entrenamiento, los puntajes de R^2 en ambos conjuntos convergen, indicando un balance adecuado entre sesgo y varianza.
- Esto evidencia que el modelo no está sobreajustando a los datos de entrenamiento y puede generalizar a nuevos datos.
- Incluso con solo el 20% de los datos de entrenamiento, el modelo alcanzó un $R^2 > 0.90$ en el conjunto de validación.

Análisis de sensibilidad

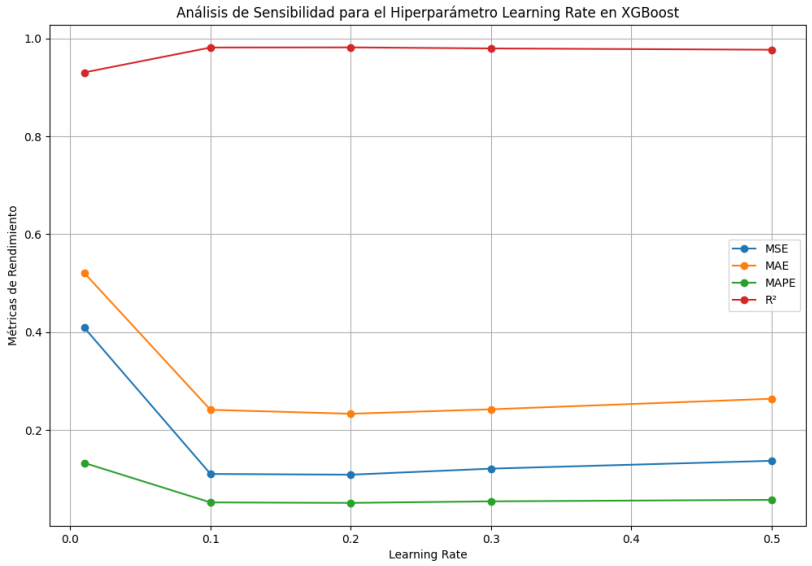

***Ilustración 43**. Análisis sensibilidad hiperparámetro modelo XGBoost*

- Con $learning_rate = 0.01$, el modelo muestra un desempeño significativamente inferior, con un MSE de 0.409, MAE de 0.521 y MAPE de 13.31%, acompañado de un R^2 de 0.9303. Esto sugiere una baja capacidad de ajuste debido a un aprendizaje excesivamente lento.
- A medida que *learning_rate* aumenta a 0.1 y 0.2, las métricas de rendimiento mejoran notablemente. Para $learning_rate = 0.2$, se obtuvo el mejor desempeño, con un MSE de 0.109, MAE de 0.233, MAPE de 5.15%, y un R^2 de 0.9814. Estos resultados reflejan un equilibrio óptimo entre velocidad de aprendizaje y ajuste.
- Con valores más altos de $learning_rate$ (0.3 y 0.5), el desempeño se deteriora gradualmente. A $learning_rate = 0.5$, el modelo muestra un incremento en los errores (MSE: 0.137, MAE: 0.264) y una disminución del R^2 a 0.9766, indicando que la mayor tasa de aprendizaje induce sobreajuste y posibles oscilaciones en el aprendizaje.

El valor óptimo de $learning_rate$ se encuentra en el rango 0.1–0.2, donde el modelo logra un ajuste preciso con un bajo error absoluto y cuadrático, así como un alto coeficiente de determinación. Asimismo, Tasas de aprendizaje demasiado bajas (0.01) o excesivamente altas (0.5) generan un aprendizaje insuficiente o inestable, comprometiendo el rendimiento del modelo.

Análisis de residuos

Gráfico Q-Q:

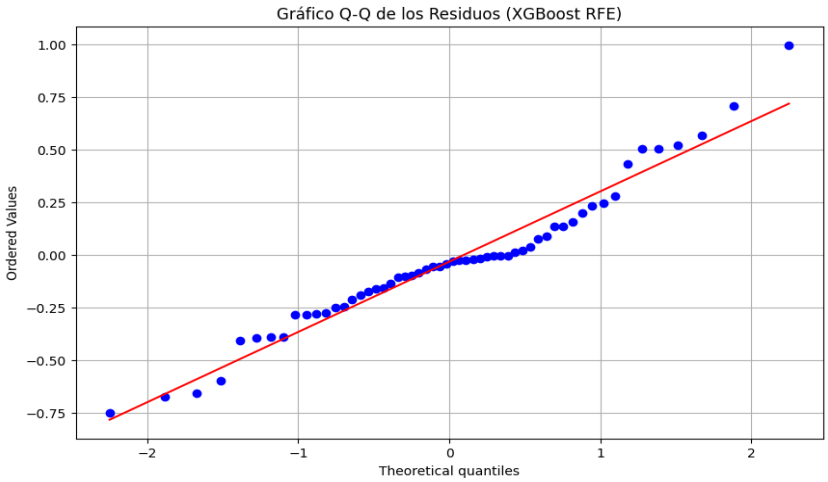

***Ilustración 44**. Análisis residuos modelo XGBoost*

- Alineación central con la línea teórica: La mayoría de los puntos en el gráfico Q-Q se alinean adecuadamente con la línea teórica de distribución normal, lo que refuerza la hipótesis de que los residuos están distribuidos normalmente en gran medida.
- Desviaciones en las colas: Se identifican ligeras desviaciones en los extremos de la distribución. Esto indica que algunos residuos extremos son más grandes o pequeños de lo que se esperaría bajo una distribución normal estándar, lo que podría deberse a la influencia de valores atípicos en el conjunto de datos.
- Simetría general: A pesar de las desviaciones en las colas, la simetría del gráfico Q-Q confirma que el modelo no presenta sesgos significativos en su capacidad predictiva

Análisis de robustez

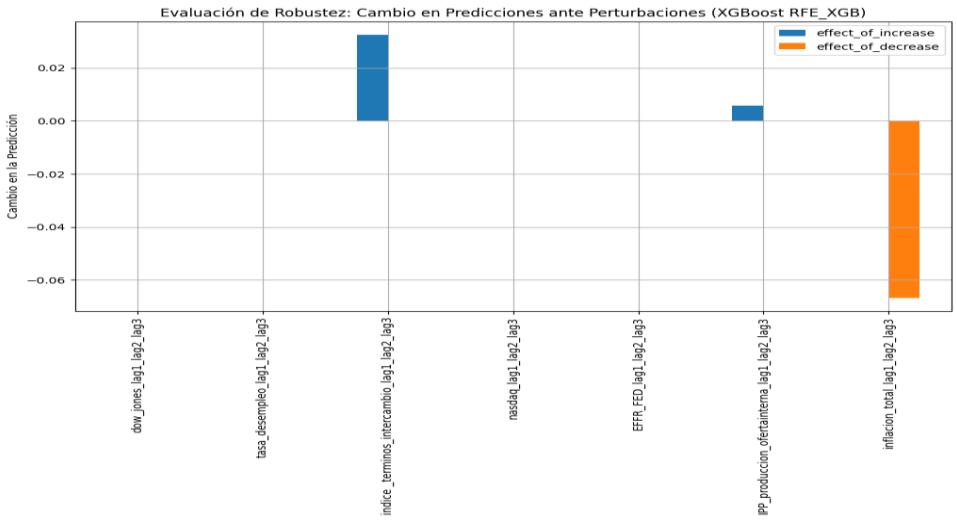

***Ilustración 45**. Análisis robustez modelo XGBoost*

- $indice_terminos_intercambio_lag1_lag2_lag3$: Un incremento del 1% en esta variable produjo un cambio positivo en la predicción de +0.0325, lo que indica su relevancia en la capacidad predictiva del modelo. Esta sensibilidad sugiere que esta variable tiene una relación directa y significativa con la variable objetivo.
- $IPP_produccion_ofertainterna_lag1_lag2_lag3$: Un incremento de esta variable resultó en un cambio positivo menor de +0.0058, reflejando una influencia moderada en las predicciones del modelo. Aunque el impacto no es tan alto como en la variable anterior, sigue siendo relevante.
- $inflacion_total_lag1_lag2_lag3$: Al reducir esta variable en un 1%, el modelo mostró un cambio negativo significativo de −0.0667, resaltando su fuerte relación inversa con la variable objetivo. Este comportamiento subraya que el modelo depende en gran medida de esta variable, y cambios negativos en la inflación total tienden a disminuir las predicciones de la tasa de intervención política monetaria.
- Las variables $dow_jones_lag1_lag2_lag3$, $tasa_desempleo_lag1_lag2_lag3$, $nasdaq_lag1_lag2_lag3$ y $EFFR_FED_lag1_lag2_lag3$ no presentaron cambios perceptibles en las predicciones bajo las perturbaciones. Esto indica que estas características tienen un impacto limitado o que el modelo ha aprendido a tratarlas como redundantes o de baja relevancia.

El modelo XGBoost con selección de características mediante RFE demuestra ser robusto frente a cambios marginales en la mayoría de las variables predictoras. Sin embargo, su alta dependencia de variables como $inflacion_total_lag1_lag2_lag3$ y $indice_terminos_intercambio_lag1_lag2_lag3$ podría ser una vulnerabilidad si estas variables sufren cambios estructurales significativos en contextos futuros.

Análisis de interpretabilidad SHAP

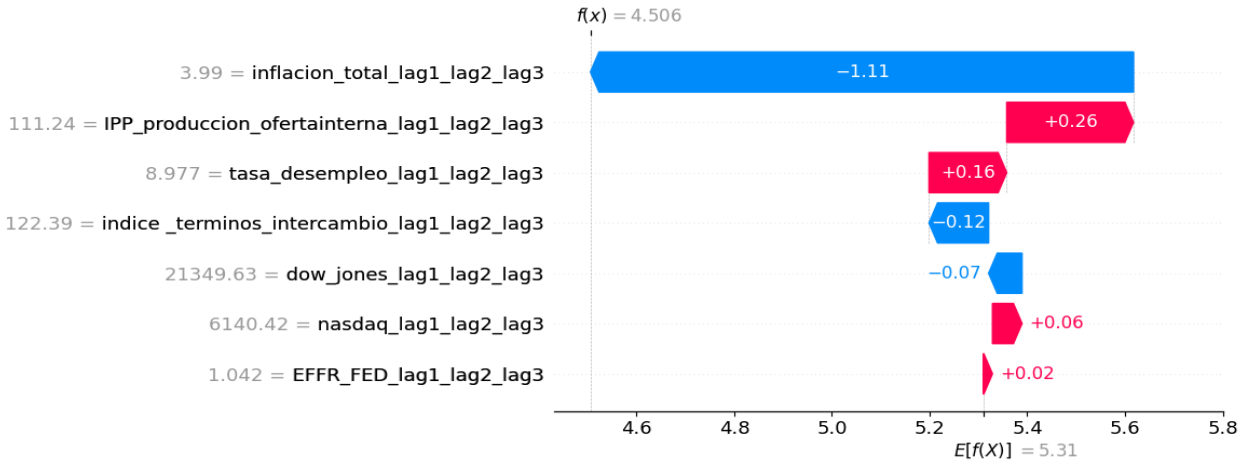

Ilustración 46. *Análisis interpretabilidad SHAP modelo XGBoost*

- Inflación total ($inflacion_total_lag1_lag2_lag3$): Esta variable emergió como la característica más influyente, con un impacto promedio negativo significativo en las predicciones. Esto sugiere que un aumento en los valores de inflación está asociado con una reducción de la tasa objetivo predicha, probablemente indicando una relación inversa entre inflación y la tasa de intervención política monetaria.
- Producción interna ($IPP_produccion_ofertainterna_lag1_lag2_lag3$): Fue la segunda variable más influyente, pero con un impacto positivo en las predicciones. Un incremento en esta variable tiende a aumentar la tasa objetivo estimada, lo que puede reflejar la relevancia de la actividad económica interna en las decisiones de política monetaria.
- Tasa de desempleo ($tasa_desempleo_lag1_lag2_lag3$): Contribuyó de manera moderada pero consistente. El impacto negativo sugiere que mayores tasas de desempleo podrían estar asociadas con menores tasas de intervención, reflejando un posible enfoque expansivo en contextos de mayor desempleo.
- Índice de términos de intercambio ($indice_terminos_intercambio_lag1_lag2_lag3$): Mostró contribuciones positivas, aunque menos pronunciadas, indicando que mayores valores en este índice están asociados con incrementos leves en la tasa predicha.
- Otros indicadores ($nasdaq_lag1_lag2_lag3$, $dow_jones_lag1_lag2_lag3$, $EFFR_FED_lag1_lag2_lag3$): Estas variables presentaron impactos menores en promedio, aunque no despreciables, lo que sugiere que su influencia es contextual y posiblemente más relevante en situaciones específicas.
- La predicción de la tasa objetivo para la instancia fue de 4.506 unidades, derivada de la combinación de las contribuciones individuales de las características y el valor base del modelo (E[f(x)] = 5.31). El valor base representa la media de las predicciones del modelo.

Conclusiones

- **Desempeño predictivo:** El modelo mostró un desempeño predictivo excepcional en el conjunto de entrenamiento, alcanzando métricas sobresalientes, como un R^2 de 0.9999, un MSE de 8.61e-07 y un MAPE de 0.015%, lo que demuestra que es altamente eficaz en capturar las relaciones entre las variables independientes y la variable objetivo. Asimismo, el conjunto de prueba, el modelo alcanzó un R^2 de 0.9806, indicando que explica el 98.06% de la varianza en datos no vistos. Las métricas de error, como un MSE de 0.1139 y un MAPE de 4.93%, reflejan una alta precisión en predicciones fuera de muestra. Esto confirma que el modelo tiene una notable capacidad predictiva, logrando mantener errores bajos tanto en datos vistos como en no vistos.
- **Capacidad de generalización:** La optimización de hiperparámetros mediante Grid Search permitió mejorar el equilibrio entre el ajuste del modelo y su capacidad para generalizar a nuevos datos. Con un R^2 de 0.9812 y un MSE de 0.1106 en el conjunto de prueba después de la optimización, el modelo demostró ser consistente y eficaz en predecir valores fuera de muestra. La validación cruzada (K-Fold) reforzó esta conclusión al mostrar un R^2 promedio de

0.9474 en diferentes particiones del conjunto de datos, lo que evidencia estabilidad en diferentes subconjuntos y una baja probabilidad de sobreajuste.

- **Robustez:** Características como $inflacion_total_lag1_lag2_lag3$ y $indice_terminos_intercambio_lag1_lag2_lag3$ demostraron una sensibilidad significativa. Por ejemplo, una reducción del 1% en la inflación total produjo un cambio negativo de −0.0667 en la predicción, reflejando la fuerte relación inversa entre esta variable y la tasa de intervención. Algunas variables, como $dow_jones_lag1_lag2_lag3$, $nasdaq_lag1_lag2_lag3$ y $EFFR_FED_lag1_lag2_lag3$, mostraron un impacto limitado o incluso nulo frente a perturbaciones, lo que indica que estas características tienen menor relevancia en el modelo final.
- **Validez estadística:** Los residuos del modelo presentan una distribución simétrica, centrados en torno a cero, lo que indica la ausencia de sesgos significativos en las predicciones. Además, muestran homocedasticidad, con una variabilidad constante a lo largo del rango de predicciones, aunque con ligeros incrementos en los valores más extremos. Por otro lado, el análisis del gráfico Q-Q reveló que la mayoría de los residuos siguen una distribución normal, aunque se observaron pequeñas desviaciones en las colas, lo que sugiere un ajuste global adecuado con leves excepciones en los valores extremos.
- **Interpretabilidad:** El análisis identificó a $inflacion_total_lag1_lag2_lag3$ como la característica más influyente, con un impacto promedio negativo significativo en las predicciones, seguida de $IPP_produccion_ofertainterna_lag1_lag2_lag3$, que mostró un impacto positivo consistente. El análisis local, representado en el gráfico Waterfall, destacó cómo las contribuciones individuales de estas características ajustan el valor base del modelo para generar la predicción final. Además, las influencias detectadas son coherentes con relaciones económicas plausibles, como la relación inversa entre la inflación y las tasas de intervención, así como el impacto positivo de la actividad económica interna sobre las decisiones de política monetaria.
- **Limitaciones:** La alta influencia de unas pocas características, como la inflación y la producción interna, sugiere que el modelo podría ser vulnerable a cambios estructurales en estas variables. Aunque esto refuerza su interpretabilidad, también indica un riesgo si estas características se ven afectadas por factores externos o ruido.

7.8. Modelo Support Vector Regressor RFE

Selección de características

El proceso de selección recursiva de características (RFE) aplicado al modelo SVR identificó las 7 características más relevantes para predecir la tasa de intervención política monetaria. Este proceso garantiza que las variables seleccionadas sean las más influyentes en el modelo, reduciendo la dimensionalidad y mejorando la interpretabilidad.

Estas características son:

```
Características seleccionadas por RFE: Index(['s&p_500_lag1_lag2_lag3', 'gasto_social_lag1_lag2_lag3',
       'salario_minimo_mensual_lag1_lag2_lag3',
       'futuros_petroleo_wti_lag1_lag2_lag3',
       'oferta_monetaria_agregada_lag1_lag2_lag3',
       'inflacion_total_lag1_lag2_lag3', 'ingresos_pgn_lag1_lag2_lag3'],
      dtype='object')
```

Ilustración 47. *Selección características modelo SVR*

Entrenamiento inicial

- En el conjunto de entrenamiento, el modelo alcanzó un coeficiente de determinación (R^2) de 0.9414, indicando que explicó el 94.14% de la varianza en los datos, lo que refleja un ajuste sólido. Sin embargo, las métricas de error, con un error cuadrático medio (MSE) de 0.5233, un error absoluto medio (MAE) de 0.4799 y un porcentaje de error absoluto medio (MAPE) de 10.25%, evidencian que aún existe margen para optimización en términos de precisión y reducción de errores.
- En el conjunto de prueba, el modelo alcanzó un coeficiente de determinación R^2 de 0.9074, explicando el 90.74% de la varianza en los datos no vistos, lo que confirma una capacidad aceptable de generalización. Las métricas de error obtenidas incluyen un error cuadrático medio (MSE) de 0.5437, un error absoluto medio (MAE) de 0.5166 y un porcentaje de error absoluto medio (MAPE) de 11.94%, lo que refleja un desempeño consistente, aunque con espacio para mejorar la precisión en datos nuevos.

El desempeño inicial del modelo indicó un buen ajuste general, pero con margen de mejora en la precisión y capacidad predictiva.

Optimización de hiperparámetros Grid Search

La optimización mediante Grid Search ajustó los hiperparámetros clave del modelo SVR para mejorar su desempeño. Los valores óptimos encontrados fueron:

- C: 10
- $Epsilon$: 0.2
- $Kernel$: *rbf*
- En el conjunto de entrenamiento, el modelo alcanzó un coeficiente de determinación (R^2) de 0.9850, lo que indica una capacidad explicativa muy elevada. Además, presentó un error cuadrático medio (MSE) de 0.1338, un error absoluto medio (MAE) de 0.2528 y un porcentaje de error absoluto medio (MAPE) de 5.42%. Estas métricas reflejan una notable reducción en los errores y un ajuste más preciso, evidenciando el impacto positivo de la optimización en la capacidad predictiva del modelo.
- En el conjunto de prueba, el modelo alcanzó un coeficiente de determinación (R^2) de 0.9772, demostrando una alta capacidad explicativa en datos no vistos. Presentó un error cuadrático

medio (MSE) de 0.1340, un error absoluto medio (MAE) de 0.2601 y un porcentaje de error absoluto medio (MAPE) de 6.01%. Estas métricas evidencian un desempeño sólido y una buena generalización, confirmando que la optimización mejoró significativamente la precisión del modelo al evaluar datos nuevos.

La optimización logró un balance adecuado entre capacidad explicativa y generalización, reduciendo los errores absolutos y mejorando la capacidad predictiva en datos no vistos.

Validación cruzada K-Fold

Se implementó validación cruzada con 5 particiones para evaluar la estabilidad del modelo. Los resultados promedio fueron:

- R^2: 0.9608, confirmando una alta capacidad explicativa y consistencia en las particiones.
- MSE: 0.2797, reflejando la variabilidad inherente en los subconjuntos.
- MAE: 0.3502, mostrando un desempeño predecible en diferentes particiones.

La validación cruzada reafirma la robustez del modelo frente a la variabilidad de los datos y reduce el riesgo de sobreajuste.

Curva de aprendizaje

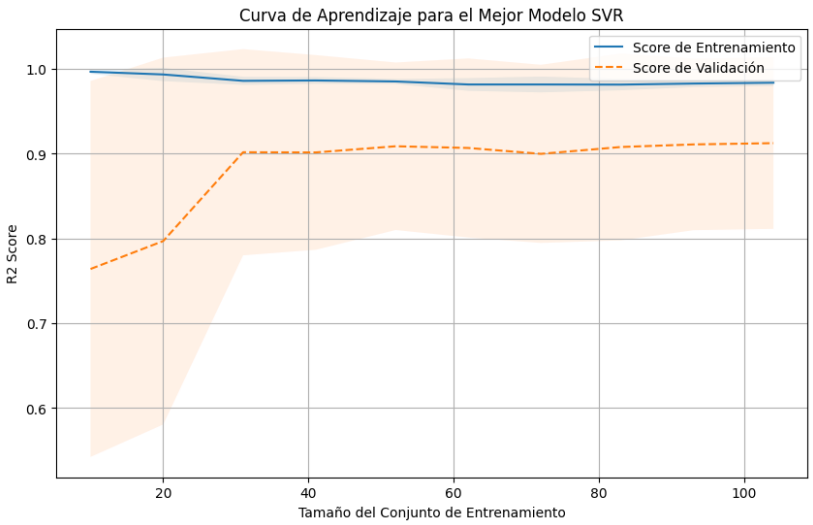

Ilustración 48. *Curva aprendizaje modelo SVR*

- Conforme aumenta el tamaño del conjunto de entrenamiento, los puntajes de R^2 en ambos conjuntos convergen, indicando un balance adecuado entre sesgo y varianza.
- Incluso con solo el 20% de los datos de entrenamiento, el modelo alcanzó un $R^2 > 0.90$ en el conjunto de validación.

Análisis de sensibilidad

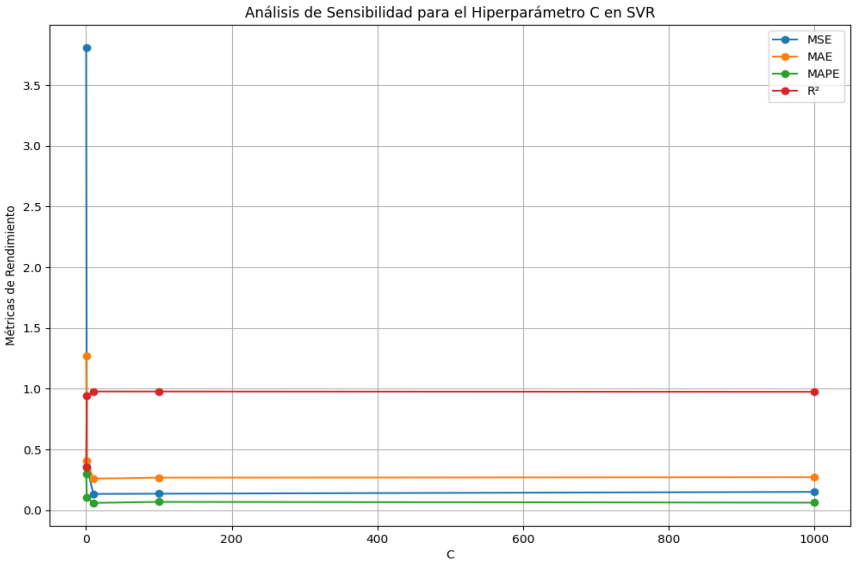

Ilustración 49. *Análisis sensibilidad hiperparámetro modelo SVR*

- Con $C = 0.1$, el modelo muestra un desempeño significativamente inferior, con un MSE de 3.809, MAE de 1.268, y MAPE de 30.26%, acompañado de un R^2 de 0.3515. Estos resultados indican una capacidad de ajuste limitada, atribuible a un margen de tolerancia amplio para errores, lo cual reduce la penalización a desviaciones en las predicciones.
- A medida que C aumenta a 1.0, las métricas de rendimiento mejoran notablemente. El modelo alcanza un MSE de 0.358, MAE de 0.406, y MAPE de 10.47%, con un incremento en el R^2 a 0.9391, lo que refleja un mejor equilibrio entre ajuste y capacidad generalizadora.
- Con $C = 10.0$, el modelo logra su mejor desempeño, alcanzando un MSE de 0.134, MAE de 0.260, y MAPE de 6.01%, acompañado de un R^2 de 0.9772. Este valor de C representa un equilibrio óptimo, donde el modelo puede ajustar las predicciones a los datos sin incurrir en sobreajuste ni subajuste.
- A valores mayores de C $(como\ 100.0\ y\ 1000.0)$, el desempeño del modelo comienza a deteriorarse ligeramente. Para $C = 100.0$, se observa un incremento en los errores (MSE: 0.136, MAE: 0.268, MAPE: 6.84%) y una leve disminución en el R^2 a 0.9769. Con C = 1000.0, los errores aumentan aún más (MSE: 0.151, MAE: 0.271, MAPE: 6.22%), y el R^2 cae a 0.9744, sugiriendo que un valor de C demasiado alto conduce a un sobreajuste, donde el modelo otorga excesiva importancia a minimizar los errores, comprometiendo la capacidad de generalización.

El análisis confirma que el valor óptimo de C se encuentra en 10.0, donde el modelo logra un equilibrio ideal entre precisión predictiva, error bajo, y robustez en la capacidad explicativa. Valores de C demasiado bajos (como 0.1) conducen a un aprendizaje insuficiente, mientras que

valores excesivamente altos (1000.0) resultan en un sobreajuste y degradación del desempeño global del modelo.

Análisis de residuos

Gráfico Q-Q:

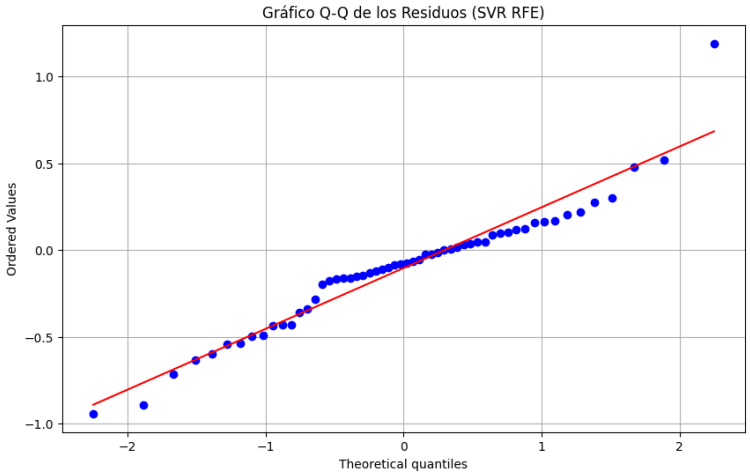

Ilustración 50. *Análisis residuos modelo SVR*

- Alineación central con la línea teórica: La mayoría de los puntos en el gráfico Q-Q se alinean adecuadamente con la línea teórica de distribución normal, lo que refuerza la hipótesis de que los residuos están distribuidos normalmente en gran medida.
- Desviaciones en las colas: Se identifican ligeras desviaciones en los extremos de la distribución. Esto indica que algunos residuos extremos son más grandes o pequeños de lo que se esperaría bajo una distribución normal estándar, lo que podría deberse a la influencia de valores atípicos en el conjunto de datos.
- Simetría general: A pesar de las desviaciones en las colas, la simetría del gráfico Q-Q confirma que el modelo no presenta sesgos significativos en su capacidad predictiva

Análisis de robustez

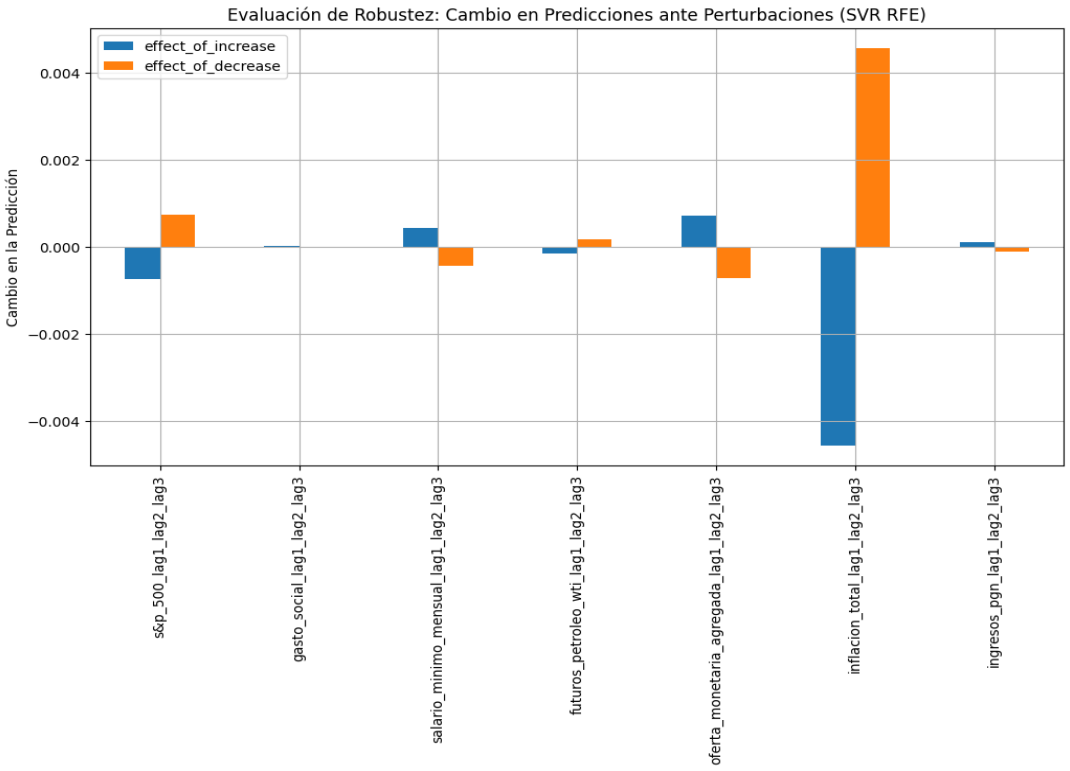

Ilustración 51. *Análisis robustez modelo SVR*

- $inflacion_total_lag1_lag2_lag3$*:* Al incrementar esta variable en un 1%, se observó un cambio negativo significativo en la predicción de -0.0046, mientras que al disminuirla en un 1%, el cambio fue positivo con $+0.0046$. Este resultado resalta su fuerte relación inversa con la variable objetivo, lo que subraya la dependencia del modelo de esta variable clave.
- $s\&p_500_lag1_lag2_lag3$*:* Un incremento del 1% en esta variable produjo un cambio negativo de -0.0007, mientras que una disminución resultó en un cambio positivo simétrico de $+0.0007$. Aunque el impacto es pequeño, su efecto bidireccional indica una relación moderada con la tasa de intervención política monetaria.
- $oferta_monetaria_agregada_lag1_lag2_lag3$*:* Incrementos del 1% en esta variable mostraron un impacto positivo de $+0.0007$, mientras que reducciones resultaron en un cambio negativo de -0.0007. Este comportamiento refleja su contribución moderada pero constante a las predicciones.
- $salario_minimo_mensual_lag1_lag2_lag3$: Al aumentar esta variable en un 1%, el modelo registró un cambio positivo de $+0.0004$, y al reducirla, un impacto negativo de -0.0004. Aunque el efecto es menor, su consistencia indica su importancia relativa en el modelo.
- $futuros_petroleo_wti_lag1_lag2_lag3$: Cambios del 1% en esta variable mostraron impactos menores en la predicción (-0.0001 al incrementarse y $+0.0001$ al disminuirse), sugiriendo una influencia limitada en el modelo.

- $gasto_social_lag1_lag2_lag3$ e $ingresos_pgn_lag1_lag2_lag3$: Estas variables presentaron efectos marginales en las predicciones, con cambios de magnitud muy pequeña tanto en incrementos como en decrementos, lo que indica una contribución limitada al modelo.

El modelo SVR con selección de características mediante RFE demuestra robustez frente a perturbaciones marginales en la mayoría de las variables predictoras. Sin embargo, su alta dependencia de variables como $inflacion_total_lag1_lag2_lag3$ representa una posible vulnerabilidad ante cambios estructurales significativos en dicha variable.

Análisis de interpretabilidad SHAP

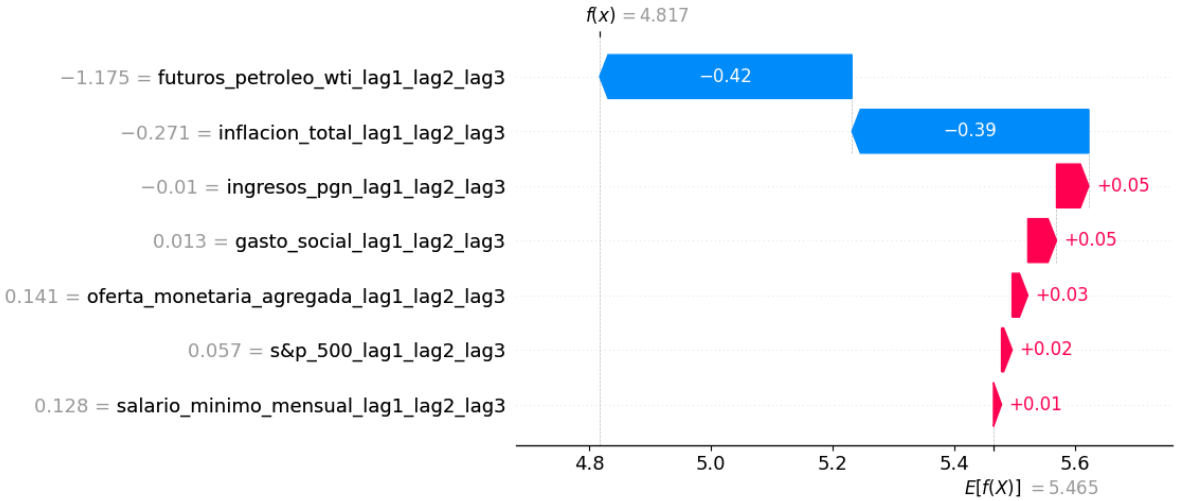

Ilustración 52. *Análisis interpretabilidad SHAP modelo SVR*

- Futuros del petróleo WTI ($futuros_petroleo_wti_lag1_lag2_lag3$): Esta variable surgió como la más influyente, con un impacto promedio negativo significativo en las predicciones. Esto indica que un aumento en los valores de los futuros del petróleo WTI está asociado con una disminución de la tasa objetivo predicha, reflejando una relación inversa entre los precios del petróleo y las decisiones de intervención monetaria.
- Inflación total ($inflacion_total_lag1_lag2_lag3$): También tuvo un impacto negativo significativo. Un aumento en esta variable está relacionado con una reducción en la predicción de la tasa objetivo, reforzando su papel crucial como factor determinante en las políticas monetarias.
- Ingresos del gobierno ($ingresos_pgn_lag1_lag2_lag3$): Su impacto fue pequeño pero negativo. Esto indica que incrementos en los ingresos del gobierno están relacionados, en menor medida, con una reducción de la tasa de intervención monetaria.
- Gasto social ($gasto_social_lag1_lag2_lag3$): Esta variable presentó una influencia positiva leve en las predicciones. Esto podría implicar que mayores gastos sociales están asociados con

incrementos en la tasa objetivo, reflejando un posible impacto expansivo en la política monetaria.

- Oferta monetaria agregada ($oferta_monetaria_agregada_lag1_lag2_lag3$): Contribuyó de manera positiva moderada, sugiriendo que incrementos en esta variable están relacionados con aumentos en la predicción de la tasa objetivo, reflejando su impacto en la liquidez del mercado.
- S&P 500 ($s\&p_500_lag1_lag2_lag3$): Su contribución fue menor pero positiva, indicando que mayores valores en este índice bursátil pueden estar asociados con incrementos marginales en la tasa predicha.
- Salario mínimo mensual ($salario_minimo_mensual_lag1_lag2_lag3$): Mostró un impacto positivo moderado en las predicciones, lo que podría reflejar su relevancia como indicador del nivel de bienestar económico y su influencia en las decisiones de política monetaria.

La predicción de la tasa objetivo para la instancia analizada fue de 4.817 unidades, derivada de la combinación de las contribuciones individuales de las características y el valor base del modelo (E[f(x)] = 5.465). Este valor base representa el promedio de las predicciones realizadas por el modelo. La descomposición de los valores SHAP permitió identificar que los futuros del petróleo WTI y la inflación total tienen los efectos más pronunciados, reflejando la sensibilidad del modelo a estas variables críticas.

Conclusiones

- **Desempeño predictivo:** El modelo SVR mostró un desempeño predictivo robusto tanto en el conjunto de entrenamiento como en el de prueba. En el conjunto de entrenamiento, alcanzó un R^2 de 0.9850, un MSE de 0.1338, y un MAPE de 5.42%, destacando su capacidad para capturar relaciones subyacentes en los datos. En el conjunto de prueba, los resultados fueron igualmente sólidos, con un R^2 de 0.9772, un MSE de 0.1340, y un MAPE de 6.01%, confirmando su capacidad para realizar predicciones precisas en datos no vistos. Este desempeño refleja un equilibrio logrado mediante la optimización de hiperparámetros, lo que permite minimizar errores sin sacrificar capacidad explicativa.
- **Capacidad de generalización:** La implementación de Grid Search y la validación cruzada K-Fold consolidaron la estabilidad y la capacidad de generalización del modelo. Tras la optimización, el modelo logró un promedio de R^2 de 0.9608 y un MSE de 0.2797 en la validación cruzada, lo que evidencia su consistencia al enfrentarse a diferentes particiones de los datos. Estos resultados minimizan el riesgo de sobreajuste, lo que indica que el modelo mantiene un desempeño predecible y estable en nuevos conjuntos de datos.
- **Robustez:** El análisis de sensibilidad destacó la relevancia de variables como $inflacion_total_lag1_lag2_lag3$ y $oferta_monetaria_agregada_lag1_lag2_lag3$. Por ejemplo, un incremento del 1% en la inflación total resultó en una disminución significativa de -0.0046 en la predicción, subrayando su fuerte relación inversa con la tasa objetivo. Variables como $s\&p_500_lag1_lag2_lag3$ y $salario_minimo_mensual_lag1_lag2_lag3$

mostraron impactos positivos, aunque menos pronunciados, reflejando su contribución moderada en el modelo. Por otro lado, características como $futuros_petroleo_wti_lag1_lag2_lag3$ tuvieron impactos menores, indicando que su relevancia en la predicción es limitada.

- **Validez estadística:** Los análisis de residuos confirmaron que el modelo cumple con los supuestos fundamentales de independencia y homocedasticidad. Los residuos se distribuyen simétricamente alrededor de cero y presentan una variabilidad constante a lo largo de las predicciones, aunque se observan leves incrementos en las colas. El gráfico Q-Q mostró que los residuos siguen en gran medida una distribución normal, con pequeñas desviaciones en los extremos, lo que refuerza la validez del modelo con excepciones en valores residuales extremos.
- **Interpretabilidad:** El análisis SHAP identificó a $inflacion_total_lag1_lag2_lag3$ como la variable más influyente, con un impacto promedio negativo destacado en las predicciones, seguido de $oferta_monetaria_agregada_lag1_lag2_lag3$, que tuvo un efecto positivo. El análisis local con el gráfico Waterfall demostró cómo estas características ajustan el valor base del modelo (E[f(x)] = 5.465) para obtener la predicción final, que en el caso específico evaluado fue de 4.817. Estas relaciones son consistentes con fundamentos económicos: por ejemplo, la relación inversa entre la inflación y la tasa de intervención monetaria, y la influencia positiva de la liquidez en el mercado.
- **Limitaciones:** La alta dependencia de variables como $inflacion_total_lag1_lag2_lag3$ representa una posible vulnerabilidad del modelo. Cambios estructurales en estas características, ya sea por choques externos o ruido en los datos, podrían comprometer la precisión de las predicciones. Adicionalmente, aunque variables como $s\&p_500_lag1_lag2_lag3$ tienen impactos consistentes, sus efectos son relativamente pequeños, lo que podría limitar su relevancia en diferentes contextos económicos

7.9. Comparación general modelos

Tabla 3. *Comparación general modelos optimizados*

Modelo	R^2	MSE	MAE	MAPE
Ridge	0.944860	0.323887	0.446680	11.15%
Lasso	0.92662	0.43105	0.51263	12.36%
Random Forest	0.97208	0.16399	0.244575	6%
XGBoost	0.98118	0.11056	0.241506	5.2%
SVR	0.97719	0.13400	0.260131	6.01%

Para identificar el mejor modelo entre Ridge, Lasso, Random Forest, XGBoost, y SVR, considerando los criterios de desempeño predictivo, generalización y robustez, validez estadística, interpretabilidad, y limitaciones observadas, se realizó un análisis integral:

1. **Desempeño predictivo:**

- XGBoost: Presentó el mejor desempeño predictivo en términos de R^2 en prueba (0.9812), con un MSE bajo (0.1106) y el menor MAPE (5.24%), lo que refleja predicciones precisas y consistentes.
- Random Forest: Alto R^2 (0.9721), MSE (0.1640), y un MAPE ligeramente mayor que XGBoost (6.07%), indicando un desempeño casi comparable.
- SVR: Alto R^2 (0.9772) y MSE (0.1340), pero con un MAPE mayor (6.01%) que XGBoost y Random Forest.
- Ridge: Buen desempeño general $(R^2 = 0.9449, MAPE = 11.15\%)$, pero inferior a los modelos no lineales.
- Lasso: Menor desempeño $(R^2 = 0.9266, MAPE = 12.36\%)$ y menor capacidad para manejar relaciones complejas.

Mejor modelo: XGBoost, seguido por Random Forest y SVR.

2. **Generalización y robustez:**

- XGBoost: Mostró una excelente generalización, confirmado por su R^2 alto (0.9812) y la estabilidad observada en validación cruzada $(R^2\ promedio = 0.9474)$. Además, su rendimiento estable en análisis de sensibilidad refuerza su robustez.
- Random Forest: Generalización robusta, con un R^2 en prueba de 0.9721 y validación cruzada $(R^2\ promedio = 0.9549)$. Sin embargo, presenta una mayor sensibilidad a perturbaciones en variables clave.
- SVR: Buena generalización $(R^2 = 0.9772)$ y estabilidad en validación cruzada $(R^2\ promedio = 0.9608)$, lo que sugiere consistencia en predicciones.
- Ridge y Lasso: Aunque lineales, ambos modelos son menos capaces de generalizar relaciones no lineales complejas presentes en los datos.

Mejor modelo: XGBoost, seguido por Random Forest y SVR.

3. **Validez estadística:**

- XGBoost y Random Forest: Ambos cumplen con supuestos clave de independencia y homocedasticidad en los residuos. Aunque se observaron ligeras desviaciones en las colas, estas no afectan significativamente la calidad de ajuste.
- SVR: Residuos simétricos y distribuidos normalmente, pero también presenta pequeñas desviaciones en los extremos, similares a los modelos anteriores.
- Ridge y Lasso: Ambos presentan residuos simétricos y normalidad aproximada, pero tienen más dificultades en capturar patrones en valores extremos.

Mejor modelo: Empate entre XGBoost, Random Forest, y SVR.

4. **Interpretabilidad:**

- Ridge: Alta interpretabilidad debido a su naturaleza lineal, lo que permite identificar fácilmente el impacto de cada variable en las predicciones.
- Lasso: Similar a Ridge, pero selecciona automáticamente un subconjunto más pequeño de características, mejorando la interpretabilidad.
- XGBoost, Random Forest, y SVR: Utilizando análisis SHAP, estos modelos ofrecen interpretabilidad razonable. Sin embargo, la naturaleza no lineal y compleja de estos modelos los hace menos intuitivos en comparación con los métodos lineales.

Mejor modelo: Ridge y Lasso, seguidos por XGBoost (gracias a SHAP).

5. **Limitaciones observadas:**

- XGBoost: Alta dependencia en algunas variables críticas como inflación total y producción interna, lo que podría ser una vulnerabilidad ante cambios estructurales. Sin embargo, su regularización evita sobreajustes graves.
- Random Forest: Presenta limitaciones en sensibilidad a perturbaciones, lo que podría reducir la robustez si las variables importantes están sujetas a ruido o errores.
- SVR: Dependencia alta de algunas variables clave y menor desempeño en presencia de relaciones muy no lineales complejas.
- Ridge y Lasso: Su naturaleza lineal limita su capacidad para capturar relaciones complejas, afectando la precisión.

Mejor modelo: XGBoost, seguido por Random Forest.

Aunque el modelo XGBoost demostró el mejor desempeño en términos de precisión predictiva, este proyecto prioriza la interpretabilidad de los resultados para comprender y comunicar cómo las variables económicas influyen en la tasa de interés de política monetaria. En este contexto, Ridge y Lasso fueron seleccionados como los modelos más adecuados, ya que su enfoque lineal permite una identificación clara y directa del impacto de cada variable en la predicción, asegurando coherencia con la teoría económica.

Si bien esta decisión implica sacrificar algo de precisión predictiva frente a XGBoost, Ridge y Lasso mantienen una capacidad predictiva robusta, con métricas consistentes como R^2 superiores a 0.92 y errores absolutos bajos tanto en los conjuntos de entrenamiento como de prueba. Además, estos modelos ofrecen validez estadística, con residuos distribuidos de forma normal y homocedasticidad razonable, lo que refuerza su fiabilidad y estabilidad en diferentes particiones de los datos. Finalmente, Ridge destaca por su capacidad de regularizar todas las variables relevantes, mientras que Lasso simplifica el modelo al seleccionar únicamente los predictores más influyentes. Ambos ofrecen un balance ideal entre precisión, robustez e interpretabilidad, satisfaciendo los objetivos de esta investigación y facilitando la generación de conclusiones transparentes y útiles para la toma de decisiones económicas.

8. MODELO ECONOMETRICO Y COMPARACIÓN CON MODELOS DE MACHINE LEARNING

8.1. Introducción

En este capítulo, se lleva a cabo una comparación entre los enfoques tradicionales de econometría y los modelos avanzados de Machine Learning en el contexto de la predicción de la Tasa de Intervención de Política Monetaria. El objetivo principal es evaluar las ventajas, limitaciones y diferencias de rendimiento entre ambos paradigmas, utilizando métricas de evalución y un conjunto común de variables económicas seleccionadas.

Se inicia el capítulo con la implementación de un modelo de regresión lineal múltiple, diseñado para establecer una línea base y proporcionar un enfoque explicativo mediante la interpretación de coeficientes. A continuación, se compara el desempeño de este modelo con los algoritmos de Machine Learning Ridge, Lasso, Random Forest, SVR y XGBoost.

Finalmente, el capítulo concluye con un análisis de los resultados obtenidos, destacando las fortalezas y limitaciones de cada enfoque. Se interpreta la relevancia de las variables económicas en el modelo econométrico y se examina cómo estas contribuyen a las predicciones de los modelos de Machine Learning. Este análisis ofrece una visión integral sobre la utilidad práctica y las implicaciones teóricas de las diferentes metodologías, proporcionando una base sólida para futuras aplicaciones en la toma de decisiones económicas y financieras

8.2. Regresión lineal múltiple

8.2.1. Implementación inicial

El modelo inicial se construyó utilizando las siguientes variables predictoras sin estandarizar, seleccionadas con base a su relevancia de la teoría económica y correlación con la variable dependiente:

- $Dtf_efectiva_anual$
- $Inflación_total$
- $Meta_inflacion$
- TRM
- PIB_real
- $PIB_nominal$
- $tasa_interes_prime_rate$
- $oferta_monetaria_agregada$
- $reserva_monetaria$
- $balance_fiscal$
- $EFFR_FED$

Prueba de Estacionariedad (ADF)

Se aplicó la prueba de Dickey-Fuller aumentada (ADF) para identificar la estacionaridad de las series temporales. Los resultados mostraron que las varias variables inflación total, PIB real, PIB

nominal, tasa de interés prime rate, oferta monetaria agregada, reserva monetaria, TRM, balance fiscal y EFFR, no eran estacionarias $(p - value > 0.05)$. Por lo tanto, se aplicaron transformaciones de diferenciación para alcanzar estacionaridad. Después de aplicar la primera diferencia, fue necesario aplicar una segunda diferencia para PIB nominal, oferta monetaria agregada y reserva monetaria.

Multicolinealidad y selección de variables

VIF para cada variable seleccionada:

	Variable	VIF
0	const	43.587851
1	inflacion_total	1.183868
2	meta_inflacion	1.180783
3	PIB_real	2.026127
4	PIB_nominal	1.518136
5	tasa_interes_prime_rate	13.206217
6	dtf_efectiva_anual	1.145619
7	oferta_monetaria_agregada	2.150578
8	reserva_monetaria	1.049474
9	TRM	1.084105
10	balance_fiscal	1.751975
11	EFFR_FED	12.825236

***Ilustración 53**. Multicolinealidad modelo econométrico inicial*

Se utilizó un análisis de correlación para priorizar las variables más relacionadas con la tasa de intervención. Posteriormente, se evaluó la multicolinealidad mediante el Factor de Inflación de la Varianza (VIF), seleccionando únicamente aquellas variables con valores aceptables (VIF < 10), lo que redujo el riesgo de relaciones espurias entre predictores.

Estimación del modelo

La estimación se realizó mediante mínimos cuadrados ordinarios (OLS), garantizando la obtención de coeficientes que minimizan la suma de los errores cuadráticos.

Resultados modelo inicial

```
Métricas de rendimiento en el conjunto de prueba:
MSE: 0.18042541579599422
MAE: 0.355867916428712
MAPE: 0.07722279238864026
R^2: 0.9674106680340653
```

***Ilustración 54**. Métricas modelo econométrico inicial*

El modelo mostró un desempeño predictivo notable, evidenciado por las siguientes métricas:

- R^2 $(Entrenamiento)$: 0.975, indicando que el 97.5% de la varianza en la tasa de intervención política monetaria se explicó por las variables seleccionadas.
- R^2 $(Prueba)$: 0.967, confirmando una alta capacidad predictiva al evaluar datos no vistos.
- Errores predictivos: error cuadrático medio (MSE) 0.18 y error absoluto medio (MAE) 0.35.

- Significancia de los coeficientes: Las pruebas estadísticas mostraron que las variables "inflación total" $(p = 0.01)$ y "DTF efectiva anual" $(p < 0.001)$ fueron altamente significativas, mientras que "PIB nominal" (p = 0.47) y "TRM" $(p = 0.38)$, no alcanzaron significancia estadística.

Asimismo, se realizaron diagnósticos adicionales:

- Autocorrelación: La estadística de Durbin-Watson $(DW = 0.323)$ indicó la presencia de autocorrelación positiva en los residuos, lo que indicó relaciones temporales no capturadas completamente.
- Heterocedasticidad: La prueba de Breusch-Pagan $(p < 0.05)$ confirmó la presencia de heterocedasticidad, señalando que la varianza de los errores no es constante.

Limitaciones

Si bien el modelo mostró un desempeño predictivo sólido, presentó varias limitaciones que deben ser abordadas para mejorar su robustez y validez:

- Autocorrelación: La baja estadística de Durbin-Watson indica que el modelo no capturó completamente las relaciones temporales en los datos, lo que podría afectar las predicciones futuras.
- Heterocedasticidad: La variabilidad inconsistente en los residuos podría generar errores estándar poco confiables, afectando la precisión en la estimación de los coeficientes.
- Significancia de variables: Varias variables con relevancia teórica, como el PIB nominal y la TRM, no fueron estadísticamente significativas. Esto podría deberse a la multicolinealidad residual o a relaciones no lineales no capturadas por el modelo.

8.2.2. Modelo ajustado

Tras los hallazgos del modelo inicial, se realizó un ajuste para refinar la estimación de la tasa de intervención política monetaria, empleando únicamente las variables estadísticamente significativas identificadas previamente: "*DTF efectiva anual*" e "*inflación total*". El proceso de ajuste del modelo se desarrolló en las siguientes etapas:

```
VIF para cada variable seleccionada:
             Variable       VIF
0               const  4.963858
1  dtf_efectiva_anual  4.258370
2     inflacion_total  4.258370
```

Ilustración 55. *Multicolinealidad modelo econométrico ajustado*

- Prueba de estacionaridad: La serie de "inflación total" fue no estacionaria según la prueba de Dickey-Fuller Aumentada (ADF) $(p > 0.05)$. Para abordar este problema, se aplicó una diferenciación de primer orden que logró estacionaridad en la serie (p < 0.05). La serie de "DTF efectiva anual" fue estacionaria desde el inicio, por lo que no requirió transformación.
- Selección de variables: El modelo ajustado incluyó las dos variables significativas

seleccionadas, con valores aceptables de multicolinealidad ($VIF \approx 4.26$ para ambas variables), lo que indica una baja redundancia entre predictores.

- Estimación del modelo: El modelo fue estimado mediante mínimos cuadrados ordinarios (OLS), maximizando la capacidad predictiva al reducir la dimensionalidad y mantener las variables clave para la estimación.

Resultados modelo ajustado

```
Métricas de rendimiento en el conjunto de prueba:
MSE: 0.1485312925394861
MAE: 0.2920175980526438
R^2: 0.970159899532633
```

***Ilustración 56**. Métricas modelo econométrico ajustado*

El modelo ajustado presentó un desempeño predictivo sólido, evidenciado por las siguientes métricas y análisis:

- R^2 $(Entrenamiento)$: 0.971, lo que indica que el modelo explicó el 97.1% de la varianza en la tasa de intervención política monetaria utilizando las dos variables seleccionadas.
- R^2 $(Prueba)$: 0.970, confirmando una alta capacidad de generalización al evaluar datos no utilizados en el entrenamiento.
- Errores predictivos en el conjunto de prueba: error cuadrático medio (MSE) 0.149 y Error absoluto medio (MAE) 0.292.
- Significancia de los coeficientes: La "DTF efectiva anual" fue altamente significativa ($p < 0.001$), con un coeficiente de 1.0107, reflejando su influencia directa y positiva en la tasa objetivo. Asimismo, La "inflación total" no fue significativa en este modelo ($p = 0.450$), aunque su inclusión respondió a su relevancia teórica y contribución en modelos previos.
- Autocorrelación: La estadística de Durbin-Watson $(DW = 0.169)$ indicó autocorrelación positiva en los residuos, lo que evidenció que algunas dependencias temporales aún no fueron capturadas por el modelo.
- Heterocedasticidad: La prueba de Breusch-Pagan $(p < 0.05)$ confirmó la presencia de heterocedasticidad, indicando que la varianza de los errores no es constante, lo que puede afectar la precisión de los intervalos de confianza.

Limitaciones del modelo ajustado

- Presencia de autocorrelación: La baja estadística de Durbin-Watson sugiere que el modelo no captura completamente la estructura temporal de los datos, afectando la fiabilidad de las predicciones dinámicas.
- Heterocedasticidad: La varianza inconsistente en los residuos puede sesgar las estimaciones estándar.

8.2.3. Comparación con modelos de machine learning

Tabla 4. *Métricas modelos de machine learning y modelo econométrico*

Modelo	MSE	MAE	MAPE	R^2
Regresión lineal múltiple inicial	0.18	0.36	8%	0.97
Regresión lineal múltiple ajustado	0.15	0.29	7%	0.97
Ridge	0.32	0.44	11.15%	0.94
Lasso	0.43	0.51	12.36%	0.92
Random Forest	0.16	0.24	6%	0.97
XGBoost	0.11	0.24	5.2%	0.98
SVR	0.13	0.26	6.01%	0.97

1. Desempeño Predictivo

El desempeño predictivo de los modelos fue evaluado mediante métricas como el Error Cuadrático Medio (MSE), Error Absoluto Medio (MAE), Porcentaje de Error Absoluto Medio (MAPE) y el Coeficiente de Determinación (R^2). Los resultados indican que los modelos de machine learning, particularmente XGBoost y Random Forest, superaron al modelo econométrico ajustado en términos de precisión y capacidad explicativa:

- XGBoost presentó el mejor desempeño, con un MSE de 0.11, MAE de 0.24, y un MAPE de 5.2%. Esto refleja una capacidad notable para capturar relaciones no lineales complejas y minimizar errores.
- Random Forest también mostró resultados sólidos, con un MSE de 0.16, MAE de 0.24, y un MAPE de 6.07%, cercano al desempeño de XGBoost.
- Modelo econométrico ajustado alcanzó un MSE de 0.15, MAE de 0.29, y un MAPE de 7%, indicando una alta capacidad predictiva, aunque ligeramente inferior a los modelos no lineales.

Implicaciones económicas: Los modelos de machine learning ofrecen mayor precisión en la predicción de la tasa de intervención, lo que podría traducirse en mejores herramientas para la toma de decisiones en política monetaria, especialmente en contextos con dinámicas complejas.

2. **Generalización y robustez**

La generalización de los modelos fue evaluada en conjuntos de prueba y mediante validación cruzada. Los resultados mostraron:

- XGBoost y Random Forest demostraron una notable capacidad de generalización, con coeficientes de determinación en validación cruzada superiores al 94%. Esto evidenció que estos modelos pueden adaptarse a nuevos datos sin una pérdida significativa de precisión.
- El modelo econométrico ajustado mantuvo un alto R^2 (0.97) tanto en entrenamiento como en prueba, pero mostró vulnerabilidad a la autocorrelación y heterocedasticidad, lo que puede limitar su aplicación en contextos de datos temporales más complejos.

Implicaciones económicas: La capacidad de generalización de los modelos de machine learning presentados es esencial para prever cambios en la tasa de intervención ante escenarios futuros, mientras que los modelos econométricos, al ser más rígidos, podrían no captar dinámicas emergentes o estructurales en la economía.

3. Interpretabilidad

La interpretabilidad es una fortaleza clave del modelo econométrico:

- Modelo econométrico: Proporciona una interpretación directa y clara de los coeficientes, alineada con la teoría económica. Por ejemplo, la relación positiva y significativa entre la DTF efectiva anual y la tasa de intervención monetaria refuerza fundamentos teóricos.
- Modelos de machine learning: Aunque menos intuitivos, herramientas utilizadas como SHAP permiten identificar las variables más influyentes (identificados en el capítulo 7), ofreciendo una interpretación razonable de los factores clave.

Implicaciones económicas: La claridad interpretativa del modelo econométrico es fundamental para comunicar hallazgos y respaldar decisiones de política monetaria. Sin embargo, los modelos de machine learning compensan con mayor precisión predictiva, lo que puede complementar los análisis teóricos tradicionales.

4**. Limitaciones técnicas y económicas**

- Modelo econométrico: Presentó problemas de autocorrelación y heterocedasticidad, lo que podría afectar la confiabilidad de los intervalos de confianza y las inferencias estadísticas. Además, su capacidad para capturar relaciones lineales y no lineales es limitada.
- Modelos de machine learning: Aunque precisos, su dependencia de pocas variables como en el caso de Random Forest, XGBoost y SVR en las variables de inflación total, podría ser una vulnerabilidad en escenarios de alta volatilidad económica. Además, su complejidad puede dificultar la comunicación de resultados a audiencias no técnicas. Caso contrario al Lasso que no presentan este problema.

Implicaciones económicas: En escenarios donde se requiere un balance entre interpretabilidad y precisión, un enfoque híbrido que combine las fortalezas de ambos paradigmas podría ser ideal. Los modelos econométricos pueden ofrecer una base teórica sólida, mientras que los modelos de machine learning pueden mejorar la precisión en la predicción.

9. CONCLUSIONES Y TRABAJOS FUTUROS

9.1. CONCLUSIONES

1. El uso del modelo Lasso se destaca por su aplicabilidad práctica en la toma de decisiones, ya que no solo permiten identificar las variables más influyentes en la tasa de interés de política monetaria, sino también cuantificar su impacto relativo. Esta capacidad resulta especialmente valiosa para individuos, empresas y formuladores de políticas económicas, ya que proporciona una herramienta predictiva robusta que facilita la anticipación de cambios en variables identificadas para Lasso (futuros petróleo wti, inflación total, crédito domestico neto del sector privado, tasa de desempleo, futuros del café y la tasa de interés de la FED).

   En términos prácticos, esto se traduce en una mejor planificación financiera y estratégica. Por ejemplo, las empresas pueden prever cómo variaciones en la tasa de interés influirán en el costo del crédito, ajustando sus planes de inversión y financiamiento en consecuencia. Asimismo, los hogares pueden tomar decisiones informadas sobre la adquisición de préstamos, mientras que las instituciones financieras y los gobiernos cuentan con una base más sólida para diseñar políticas que promuevan la estabilidad económica. En conjunto, estos modelos no solo mejoran la capacidad de respuesta ante cambios económicos, sino que también contribuyen a un entorno más predecible y eficiente en la gestión de riesgos económicos y financieros.

2. La regresión lineal múltiple, aunque respaldada por fundamentos teóricos sólidos y su capacidad para ofrecer interpretabilidad clara, tiene limitaciones importantes al abordar problemáticas económicas complejas, especialmente cuando se enfrenta a relaciones no lineales, multicolinealidad elevada o datos de alta dimensionalidad. Por otro lado, los modelos de Machine Learning Ridge, Lasso, Random Forest, XGBoost y SVR, ofrecen herramientas más avanzadas para manejar estos desafíos, gracias a su flexibilidad y capacidad de capturar patrones complejos. Herramientas recientes como el SHAP han mejorado considerablemente la interpretabilidad de estos modelos, permitiendo a los usuarios comprender mejor cómo cada variable contribuye a las predicciones.

3. No obstante, la elección entre ambos enfoques no debe considerarse excluyente. Los modelos econométricos, como la regresión lineal múltiple, son ideales para obtener insights iniciales, proporcionar un marco teórico y explorar relaciones lineales claras. Por otro lado, los modelos de Machine Learning son más adecuados cuando el objetivo es maximizar la precisión predictiva en problemas complejos. Una combinación de ambos enfoques puede aprovechar las fortalezas de cada método, utilizando la econometría para comprender la estructura subyacente de los datos y los modelos de Machine Learning para optimizar el rendimiento en aplicaciones prácticas.

9.2. TRABAJOS FUTUROS

- Aunque en este estudio se utilizó Recursive Feature Elimination (RFE) como técnica principal de selección de características, existe la oportunidad de explorar enfoques alternativos que podrían ofrecer ventajas adicionales en términos de desempeño y robustez. Por ejemplo, métodos basados en regularización como Elastic Net, que combina las propiedades de Ridge y Lasso, podrían seleccionar características relevantes mientras gestionan la multicolinealidad de manera más eficiente. Asimismo, técnicas no supervisadas como el análisis de componentes principales (PCA) podrían ser utilizadas para reducir la dimensionalidad del conjunto de datos, capturando patrones subyacentes complejos sin depender de la variable objetivo.
- Una línea de investigación innovadora sería diseñar modelos híbridos que combinen enfoques econométricos y de Machine Learning. Se podrían usar modelos econométricos para identificar relaciones fundamentales entre variables y, posteriormente, integrar estos resultados como insumos en algoritmos de Machine Learning más complejos como XGBoost o Random Forest. Este enfoque podría maximizar tanto la capacidad interpretativa como la predictiva, aprovechando las fortalezas de cada paradigma para ofrecer un análisis más integral y útil en la predicción de tasas de política monetaria.
- Dado que la regresión lineal múltiple tiene limitaciones para capturar dinámicas económicas complejas, futuros trabajos podrían explorar modelos econométricos más avanzados como los modelos autorregresivos con retardos distribuidos (ARDL), modelos de corrección de errores (ECM), o incluso modelos de cointegración para datos macroeconómicos. Estos métodos permiten analizar relaciones dinámicas y de largo plazo entre las variables, lo que enriquecería la comparación con los modelos de Machine Learning y aportaría una perspectiva más profunda sobre las interacciones económicas.
- La implementación de los modelos Ridge o Lasso presentados en este proyecto en un dashboard interactivo sería una aplicación práctica invaluable para la toma de decisiones. Este dashboard podría monitorear en tiempo real las variables más influyentes en la tasa de interés de política monetaria, proporcionando estimaciones actualizadas sobre su impacto relativo. Una interfaz visual y fácil de usar podría ser especialmente útil para analistas financieros, responsables de políticas económicas y empresas, permitiendo que ajusten sus estrategias en función de cambios proyectados en variables clave identificadas.
- El análisis podría enriquecerse incluyendo una mayor cantidad de variables internacionales, como las tasas de interés de otros bancos centrales, indicadores de riesgo geopolítico, o datos económicos de países con vínculos comerciales significativos. En particular, se podrían investigar cómo las decisiones el Banco Central Europeo (BCE), o el Banco Popular de China impactan indirectamente las tasas de política monetaria locales (en el proyecto se implementó la tasa de la reserva federal de EE. UU.).

10. BIBLIOGRAFÍA

[1] Banco de la República de Colombia, «Informe de Política Monetaria,» Noviembre 2023. [En línea]. Available: https://www.banrep.gov.co/es/publicaciones-investigaciones/informe-politica-monetaria/historico-informes. [Último acceso: 08 12 2023].

[2] F. S. Mishkin, de *The economics of money, banking, and financial markets (13th ed.)*, Pearson, 2022, pp. 354-355.

[3] N. G. Mankiw, Principles of Economics, Cengage Learning, 2020.

[4] Subgerencia de Política Monetaria e Información Económica / Gerencia Técnica, «Banco de la República de Colombia,» Noviembre 2019. [En línea]. Available: https://www.banrep.gov.co/sites/default/files/paginas/proceso-de-toma-decisiones-de-politica-monetaria.pdf. [Último acceso: 20 Noviembre 2023].

[5] Banco de la República, «Informe de Política Monetaria,» Banco de la República, Bogotá, Colombia, 2022.

[6] J. Taylor, «The Role of Interest Rate Policies in Stabilizing the Economy,» *Journal of Economic Perspectives,* vol. 24, nº 2, pp. 31-46, 2010.

[7] N. G. Mankiw, de *Macroeconomics (12a ed.)*, Worth Publishers, 2019, p. 345.

[8] P. Krugman, International economics (11a ed.), Pearson, 2023.

[9] N. Gregory Mankiw, de *Principios de Economía*, CENGACE, 2012, pp. 643-664.

[10] M. Cárdenas, «Introducción a la economía colombiana,» Alpha Editorial , 2020, p. 20.

[11] A. S. Blinder, M. Gertler y D. W. Skidelsky, «The relationship between interest rates and economic growth,» *Journal of Economic Perspectives,* vol. 15, nº 4, pp. 3-27, 2001.

[12] B. d. l. R. d. Colombia, «Banco de la República de Colombia,» 1 Febrero 2023. [En línea]. Available: https://www.banrep.gov.co/es/glosario/tasa-de-cambio. [Último acceso: 21 Noviembre 2023].

[13] N. G. Mankiw, de *Macroeconomics*, Worth Publishers, 2019, pp. 388-392.

[14] R. R. Da Guimaraes, «DEEP LEARNING MACROECONOMICS,» *Universidade Federal Do Rio,* 2021.

[15] J. Friedman, R. Tibshirani y T. Hastie, de *The Elements of Statistical Learning*, Springer, 2009, pp. 61-68.

[16] T. Hastie, R. Tibshirani y J. Friedman, de *The Elements of Statistical Learning*, Springer, 2009, pp. 68-69.

[17] L. Breiman, «Random Forests,» *Journal of Machine Learning Research,* vol. 45, pp. 5-32, 2001.

[18] J. H. Friedman, «Greedy Function Approximation: A Gradient Boosting Machine,» *The Annals of Statistics,* vol. 29, nº 5, pp. 1189-1232, 2001.

[19] C. Cortes y V. Vapnik, «Support-vector networks,» *Machine Learning,* vol. 20, nº 3, pp. 273-297, 1995.

[20] I. Goodfellow, Y. Bengio y A. Courville, de *Deep Learning*, London, Massachusetts Institute of Technology, 2016, pp. 161-217.

[21] A. Gaves, de *Supervised Sequence Labelling with Recurrent*, Springer, 2012, pp. 18-23.

[22] Y. LeCun, L. Bottou, Y. Bengio y P. Haffner, «Gradient-Based Learning Applied to Document Recognition,» *Proceedings of the IEEE,* vol. 86, nº 11, p. 2278–2324, 1998.

[23] P. S, «Deep Learning para la predicción de tasas de interés en mercados emergentes,» *Journal of Economic Forecasting,* vol. 11, nº 1, pp. 25-40, 2019.

[24] J. Muller y E. Hansen, «Predicción de tasas de interés mediante modelos híbridos de econometría y Machine Learning en el mercado europeo,» *European Financial Journal,* vol. 28, nº 5, pp. 123-145, 2021.

[25] R. Kim y T. Suzuki, «Impacto de las tasas de interés en la economía global y su predicción mediante redes neuronales artificiales,» *Global Economic Studies,* vol. 5, nº 2, pp. 22-39, 2020.

[26] Banco de la República , «Tasas de interés de política monetaria,» [En línea]. Available: https://www.banrep.gov.co/es/estadisticas/tasas-interes-politica-monetaria#:~:text=La%20tasa%20de%20intervenci%C3%B3n%20de,de%20mercado%20abierto%20(OMA)..

[27] J. J. Taborda Ceballos y J. S. Cerón Castaño, «Efectos de las Tasas de Interés sobre el precio de las acciones en el mercado colombiano,» *Universidad EIA,* 2022.

[28] N. Carmona Restrepo, «Redes neuronales regularizadas para la predicción de la inflación colombiana,» *Universidad Nacional de Colombia,* 2022.

[29] Á. D. Torres Becerra y M. F. Téllez Becerra, «Nowcasting de la tasa de desempleo de Colombia por medio de Machine Learning y pronóstico de la inflación y representación por medio de la curva de Phillips proyectada,» *Universidad Piloto de Colombia,* 2022.

11. ANEXOS

En esta sección se presentan los enlaces a los scripts utilizados en Google Colab para la implementación y análisis de los diferentes modelos abordados en este proyecto. Los scripts incluyen el detalle de las etapas de procesamiento, análisis y evaluación de cada modelo:

Importación y transformación de datos

Este script contiene el proceso completo de importación, consolidación, transformación y limpieza de datos. Incluye análisis descriptivo y exploratorio, identificación y manejo de datos faltantes mediante el modelo Isolation Forest, y la generación de rezagos tanto para datos estandarizados como no estandarizados (https://colab.research.google.com/drive/16ygL3aleooSscugivD9sX-Z9HIHRt0f5?usp=sharing).

Modelo Ridge

En este enlace se documenta el proceso de implementación del modelo Ridge. Incluye el análisis de correlación, selección de características con la técnica de Recursive Feature Elimination (RFE), análisis de sensibilidad, evaluación de residuos y robustez, e interpretabilidad del modelo mediante herramientas SHAP (https://colab.research.google.com/drive/1rci4YmyRHgc4-uFrHBrxmhQ5T1MRsXcJ?usp=sharing)

Modelo Lasso

Este script abarca el desarrollo del modelo Lasso. Incluye análisis de correlación, selección de características con RFE, evaluación de sensibilidad, análisis de residuos, pruebas de robustez, y visualización de la interpretabilidad de las predicciones a través de SHAP (https://colab.research.google.com/drive/1_jRfMvyG6DbVwMS1f43Zh3peqhpZEgh5?usp=sharing).

Modelo Random Forest

Aquí se detalla el proceso para el modelo Random Forest. Contempla el análisis de correlación, selección de características con RFE, análisis de sensibilidad, evaluación de residuos y robustez, además de la interpretación del modelo utilizando SHAP (https://colab.research.google.com/drive/13ytL5i3gRUPXjqLPIuGL3Wo_XVoS_Kpx?usp=sharing).

Modelo Support Vector Regressor (SVR)

En este script se describe la implementación del modelo SVR, incluyendo análisis de correlación, selección de características con RFE, evaluación de sensibilidad, análisis de residuos, robustez, y herramientas de interpretabilidad SHAP (https://colab.research.google.com/drive/1GVrF_EhINPqRiz53WwOAWcjRm4C-

Eh7p?usp=sharing).

Modelo Gradient Boosting - XGBoost

Este enlace abarca todo el proceso del modelo XGBoost, con un enfoque en el análisis de correlación, selección de características con RFE, evaluación de sensibilidad, análisis de residuos, robustez, e interpretación avanzada de las predicciones mediante SHAP (https://colab.research.google.com/drive/1GVrF_EhINPqRiz53WwOAWcjRm4C-Eh7p?usp=sharing).

Modelo econométrico de regresión lineal múltiple

Este script detalla el desarrollo del modelo econométrico de regresión lineal múltiple. Incluye pruebas de estacionaridad, diagnóstico de multicolinealidad, estimación de coeficientes mediante mínimos cuadrados ordinarios, y análisis de autocorrelación y heterocedasticidad en los residuos (https://colab.research.google.com/drive/198TeQGKKt0vqLVjlP2R0AJrzA1FZzdzh?usp=sharing).