

PREDICCIÓN DEL PIB DE COLOMBIA: TÉCNICAS CLÁSICAS VS MACHINE LEARNING.

PROYECTO INTEGRADOR II

MAESTRÍA EN CIENCIA DE LOS DATOS Y ANALÍTICA

JONATHAN MONTES, VICTOR URIBE, MARIA CAMILA
LOPERA, CRISTIAN CASTRO, KARLA TORRES

ÍNDICE DEL DOCUMENTO

- ENTENDIMIENTO DEL NEGOCIO
Idea del Proyecto
- ENTENDIMIENTO DE LOS DATOS
¿Con que vamos a trabajar?
- PREPARACION DE LOS DATOS
¿Como tratamos los datos para el objetivo?
- MODELADO
¿Qué modelos vamos a usar?
- EVALUACIÓN
¿Como evaluamos el exito de los modelos?

1 — ENTENDIMIENTO DEL NEGOCIO

IDEA DEL PROYECTO

INTRODUCCIÓN

NUESTRA METODOLOGÍA



PIB

El PIB representa el valor monetario de todos los bienes y servicios finales producidos en un país.



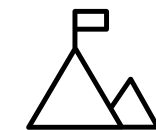
PIB REAL

Utiliza los precios de un año base para eliminar el efecto de las variaciones de precios



CALCULO

- Ingreso
- Gasto
- Valor agregado



OBJETIVO

- Comparar Técnicas Clásicas y de Machine learning para la predicción del PIB.
- Mejor desempeño con MSE.

MARCO TEÓRICO

ARIMA:

Series de tiempo:
Demanda - Fattah,
Ezzine, Aman et al.
(2018)

Índice de precios -
Wahyudi (2017)

ARX, SARIMAX

ARX:
Tasa de desempleo a
partir de comentarios
- Kwon, K. Cho, W. and
Na, J. (2016)

SARIMAX
Venta de alimentos
perecederos - Arunraj,
N.S., Ahrens, D.,
Fernandes, M. (2016)

KNN/ KNN MULTIVARIANTE

Cross-Validation

Predicción de
producción -
Tajmouati, Wahbi,
Bedoui et al. (2021)

Multivariante: Índice
de precios - Ban, ,
Zhang, Pang et al.
(2013)

REGRESIÓN LINEAL, RIDGE Y LASSO

Fracaso Corporativo -
Pereira, Basto and
Silva (2016)

2 — ENTENDIMIENTO DE LOS DATOS

¿CON QUE VAMOS A TRABAJAR?

ENTENDIMIENTO DE LOS DATOS

Datos trimestrales desde 2007-1 a
2023-1.

65 Observaciones

Variables Macroeconomicas

- DESEMPLEO
- IPC
- VARIACIÓN DEL IPC
- CAPACIDAD DE UTILIZACIÓN
DE MANUFACTURAS
- TASA DE INTERVENCIÓN BR

Renta Fija

- TES 1 AÑO
- TES 5 AÑOS
- TES 10 AÑOS

Curva Cero Cupon

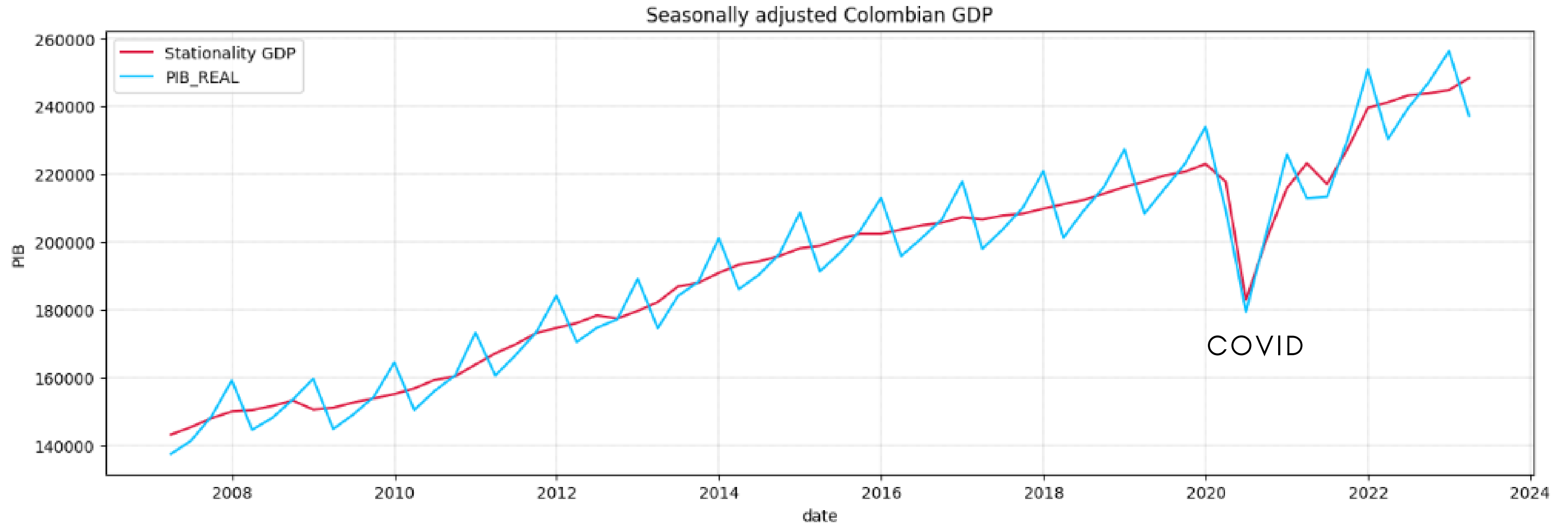
- NIVEL YIELD
- INCLINACIÓN YIELD
- CURVATURA YIELD

Variable Independiente

PIB Real
PIB Real Estacionario

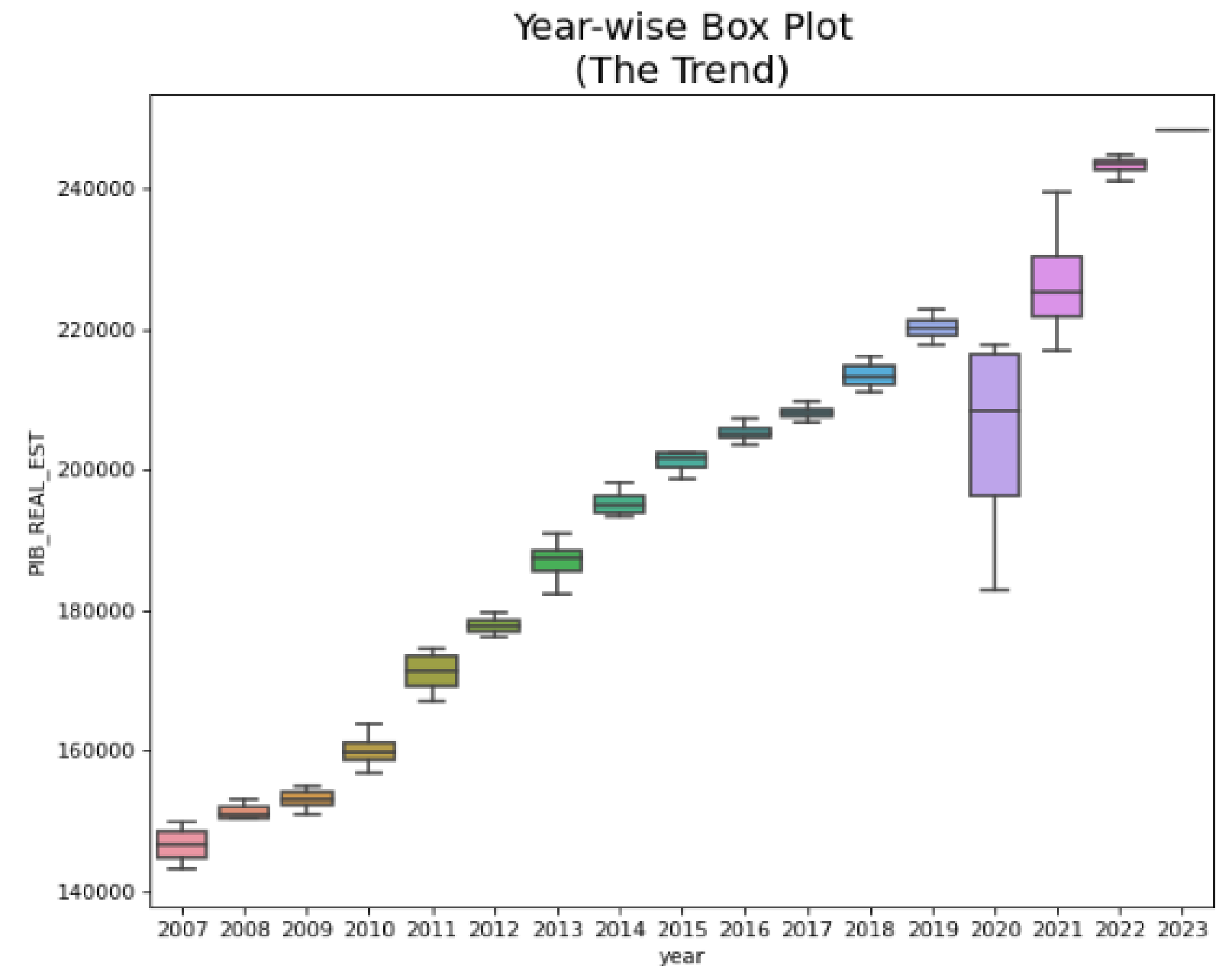
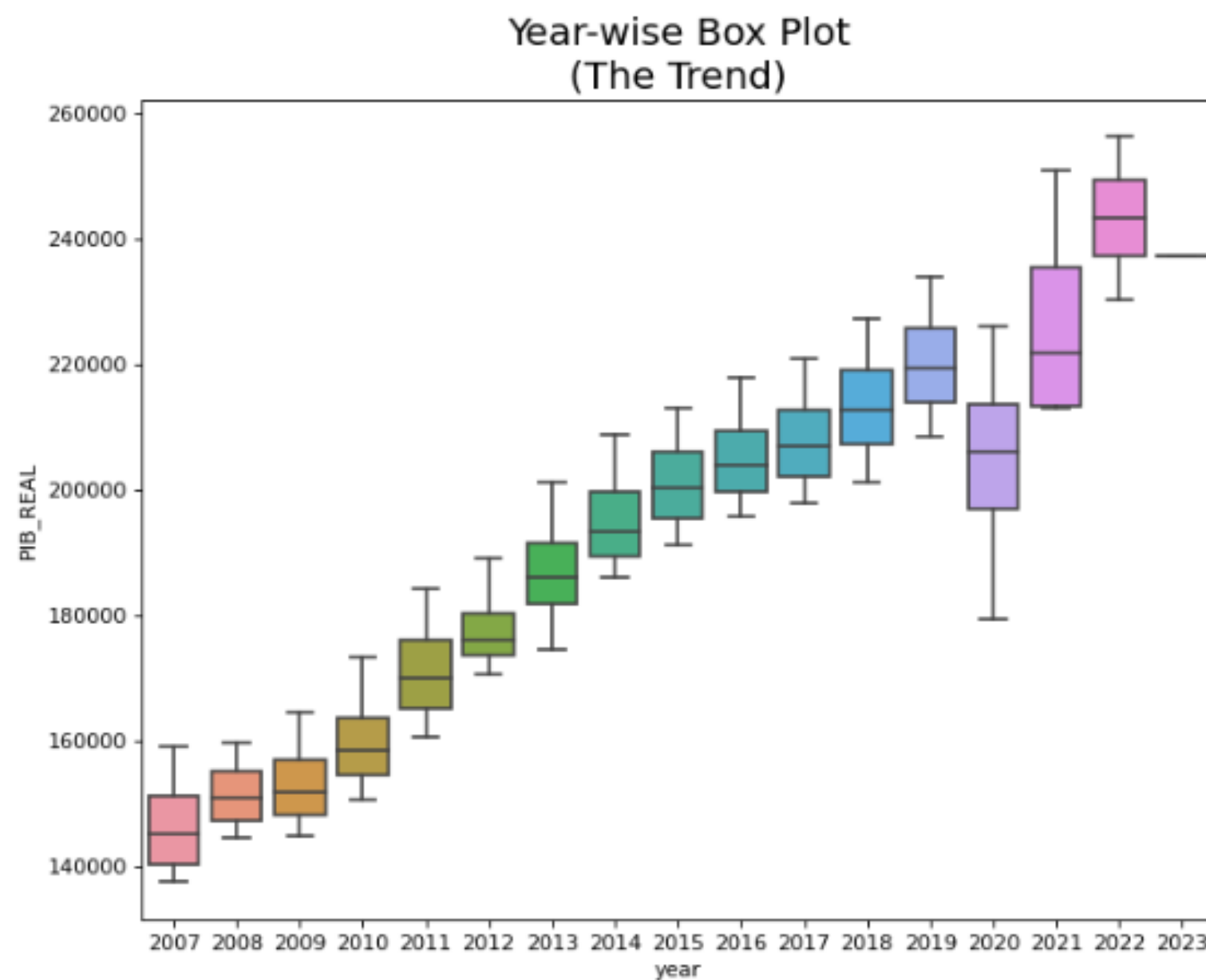
COMPORTAMIENTO DEL PIB REAL

SERIE DE TIEMPO CON TENDENCIA Y ESTACIONAL



COMPORTAMIENTO DEL PIB REAL

SERIE DE TIEMPO CON TENDENCIA Y ESTACIONAL

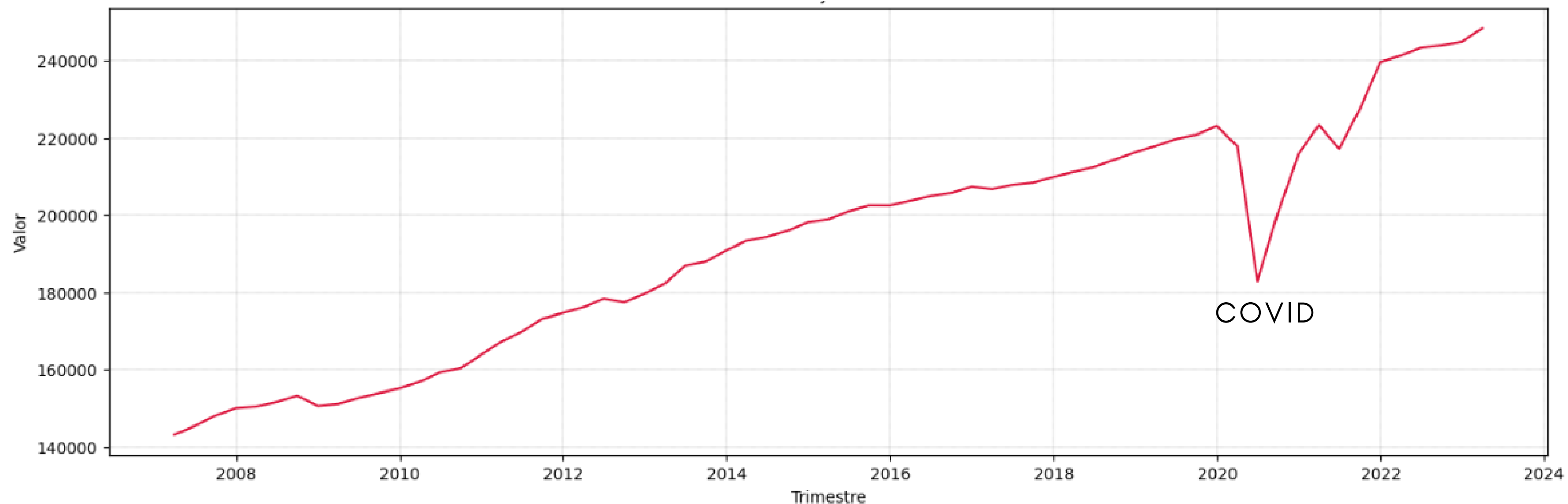


- Se presenta tendencia tanto en la serie Real y ajustada.
- No se presentan datos atípicos dentro del contexto estadístico.

COMPORTAMIENTO DEL PIB REAL

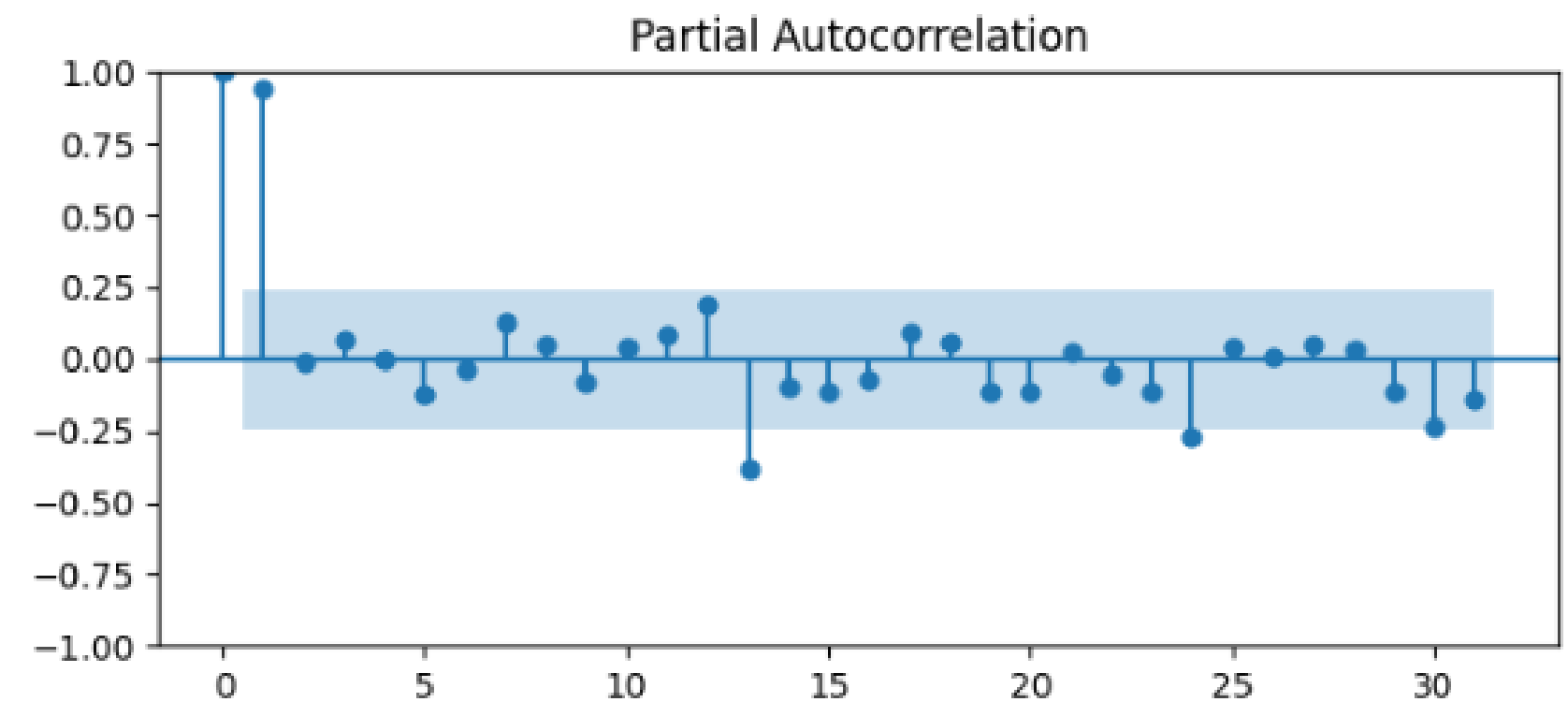
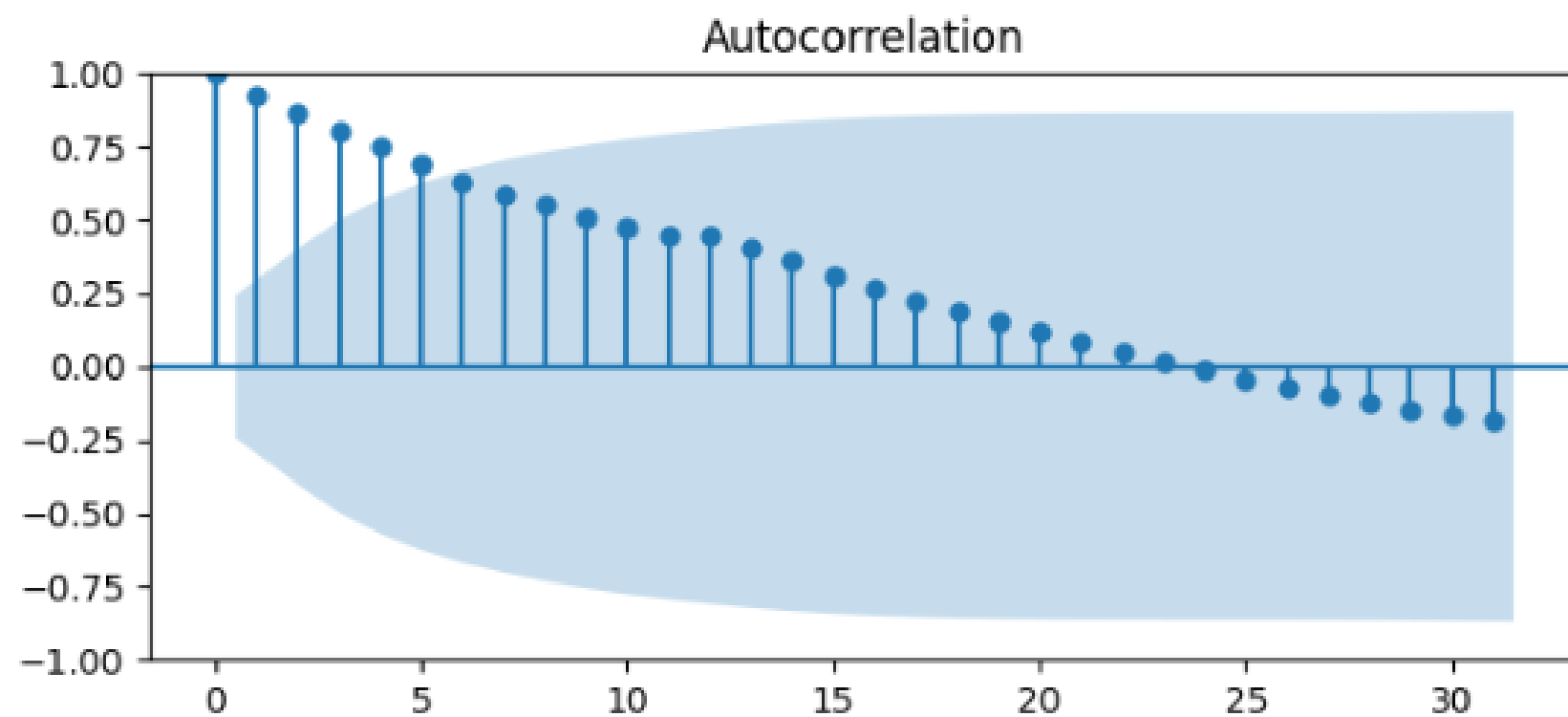
CON AJUSTE ESTACIONAL

VERIFICACIÓN POR INTERPOLACIÓN DE LA SERIE
CON CORREGIMIENTO ESTACIONAL



COMPORTAMIENTO DEL PIB REAL

AUTOCORRELACIÓN Y AUTOCORRELACION PARCIAL



3 – TRATAMIENTO DE LOS DATOS

¿CÓMO TRATAMOS LOS DATOS PARA EL OBJETIVO?

TRATAMIENTOS DE LOS DATOS

ORIGEN DE LOS DATOS

- Se recopiló información del gobierno y del banco de la república, abarcando el periodo comprendido entre 2007 y el primer trimestre de 2023. (65 observaciones)
- La variable objetivo fue el PIB real con ajuste estacional.

TRATAMIENTOS DE LOS DATOS

ESTACIONARIZACIÓN DE LAS VARIABLES PIB REAL Y PIB REAL AJUSTADO
(CON AJUSTE ESTACIONAL)

Una de las partes mas importantes para trabajar con series de tiempo es estacionarizar la serie ya que de esta manera se busca que la media y la varianza de la serie sean constantes en el tiempo.

Por medio de las siguientes técnicas se busca estacionarizar las variables objetivo : PIB Real y PIB Real Ajustado

DIFERENCIAS

Diferencia de entre cada elemento n y $n-1$

DIFERENCIAS LOGARITMICAS

Diferencia de entre cada elemento n y $n-1$ de la serie logarítmica

LOGARITMOS

Logaritmo natural de la serie obeitivo

RAIZ CUADRADA

Raiz cuadrada de la serie objetivo

TRATAMIENTOS DE LOS DATOS

PRUEBA DE DICKEY-FULLER

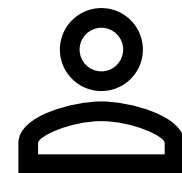
- Se decidió aplicar una transformación logarítmica a los datos y calcular la diferencia entre observaciones consecutivas.
- Se estimó el primer dato de la serie de las diferencias logarítmicas como la media de las siguientes 3 observaciones dada la similitud en los valores trimestrales cercanos. No se podía dropear la fila.
- Sin estas modificaciones de la serie se el test indicaba que la serie no era estacionaria.

4 – MODELADO

MODELOS

¿QUÉ VAMOS A HACER PARA CONSEGUIR NUESTRO OBJETIVO?

UNI-VARIADO



MODELOS AUTOREGRESIVOS DE LA VARIABLE OBJETIVO

Métodos Clásicos

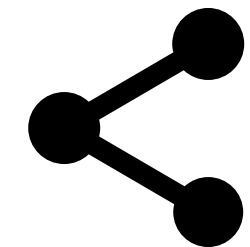
- AR (2 y 4 rezagos)
- ARIMA

Metodos ML

- KNN

MÉTRICA: MSE.

MULTI-VARIADO



MODELOS EVALUADOS POR ESCENARIOS

Métodos Clásicos

- ARX
- SARIMAX
- Regresiones (Lineal, Lasso, Ridge)
- Ajuste polinomial grado n.

Metodos ML

- SVM
- KNN

MÉTRICA: MSE.

MODELOS

INGENIERÍA DE CARACTERÍSTICAS

UNI-VARIADO

- Limpieza NaN
- Identificación de Outliers.

MULTI-VARIADO

BACKWARD/FORWARD STEPWISE SELECTION

- Escenario 1: Curva cero coupon
- Escenario 2: Curva cero coupon, Macro-variables
- Escenario 3: Macro-Variables
- Escenario 4: Proxies
- Escenario 5: Macrovariables, Proxies
- Escenario 6: Curva cero coupon, Macro-Variables, Proxies

PARA TODOS LOS MODELOS;

- Multistep-ahead forecasting:
 - Train: 2007-1 ~ 2020-1
 - Test: 2020-2 ~ 2023-1
 - No se incluyen los datos de la pandemia COVID-19 en los datos de entrenamiento
- Métrica: MSE

5 – EVALUACIÓN

¿COMO EVALUAMOS EL EXITO DE LOS MODELOS?

UNI-VARIABLES

Métrica

MSE

El MSE es considerado una de las mejores métricas para comparar modelos debido a su capacidad para penalizar los errores más grandes de manera significativa

Resultados

ARIMA Y KNN

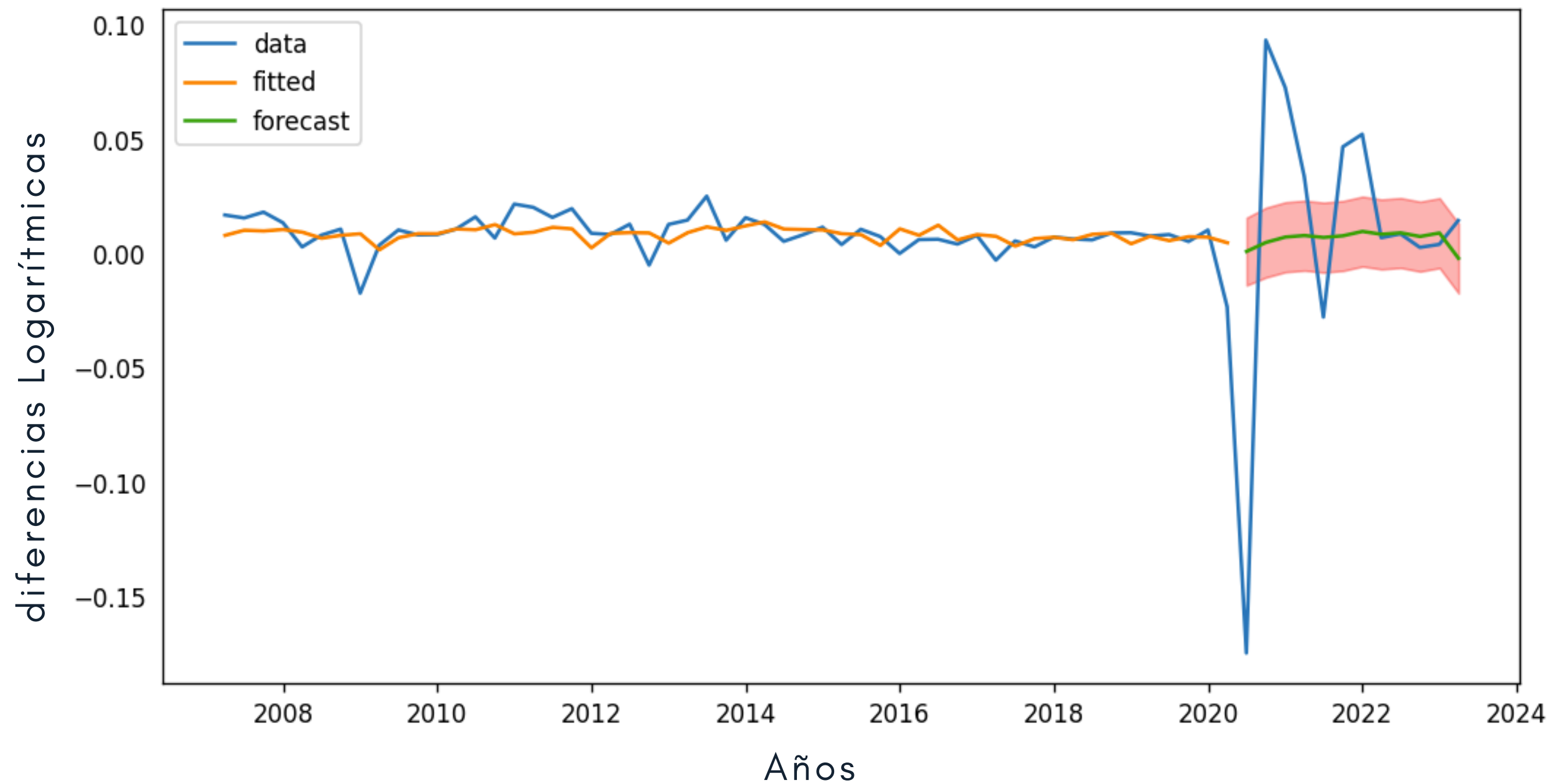
- Los modelos tradicionales presentan un MSE muy cercano
- El modelo ARIMA es el mejor modelo tradicional
- KNN presenta el mejor rendimiento de los univariables

MODELO	MSE
AR (2)	4.22x10 ⁻³
AR (4)	4.23x10 ⁻³
ARIMA	4.03x10 ⁻³
KNN	2.7x10 ⁻³

UNI-VARIADO

MEJOR DE LOS MODELOS CON METODOS DE TRADICIONALES

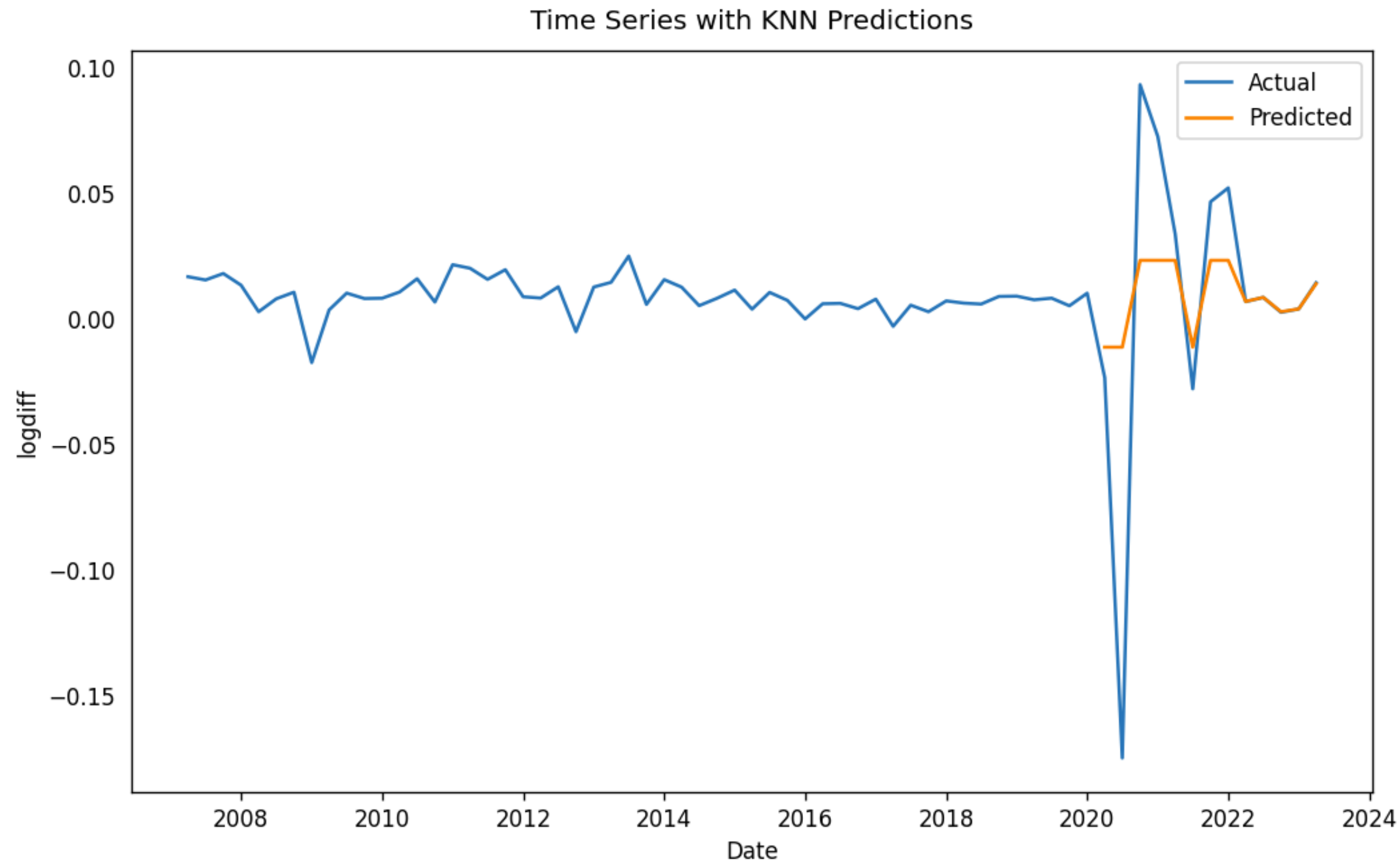
ARIMA



- MSE: 4.03×10^{-3}
- ARIMA (1,0,0)(0,0,1)

UNI-VARIADO

MEJOR DE LOS MODELOS CON METODOS DE ML



- MSE: 2.7×10^{-3}
- Se observa un muy buen ajuste en la predicción a partir de 2022

ESTUDIO DE REFERENCIA

GDP FORECASTING:
MACHINE LEARNING, LINEAR
OR AUTOREGRESSION?

Maccarrone G, Morelli G and
Spadaccini S (2021)

Resultados Paper:

Strategy	SARIMA	KNN	AR
One-step	2,87e-03	1,73e-03	3,47e-03
Multi-step	3,84e-03	3,02e-03	4,18e-03

Resultados Nuestros:

MODELO	MSE
AR (2)	4.22x10 ⁻³
AR (4)	4.23x10 ⁻³
ARIMA	4.03x10 ⁻³
KNN	2.7x10 ⁻³

- Los modelos tradicionales presentan un MSE muy cercano
- El modelo ARIMA es el mejor modelo tradicional
- KNN presenta el mejor rendimiento de los univariables

MULTI-VARIABLES

Métrica

MSE

El MSE es considerado una de las mejores métricas para comparar modelos debido a su capacidad para penalizar los errores más grandes de manera significativa

Resultados

ARIMAX Y KNN

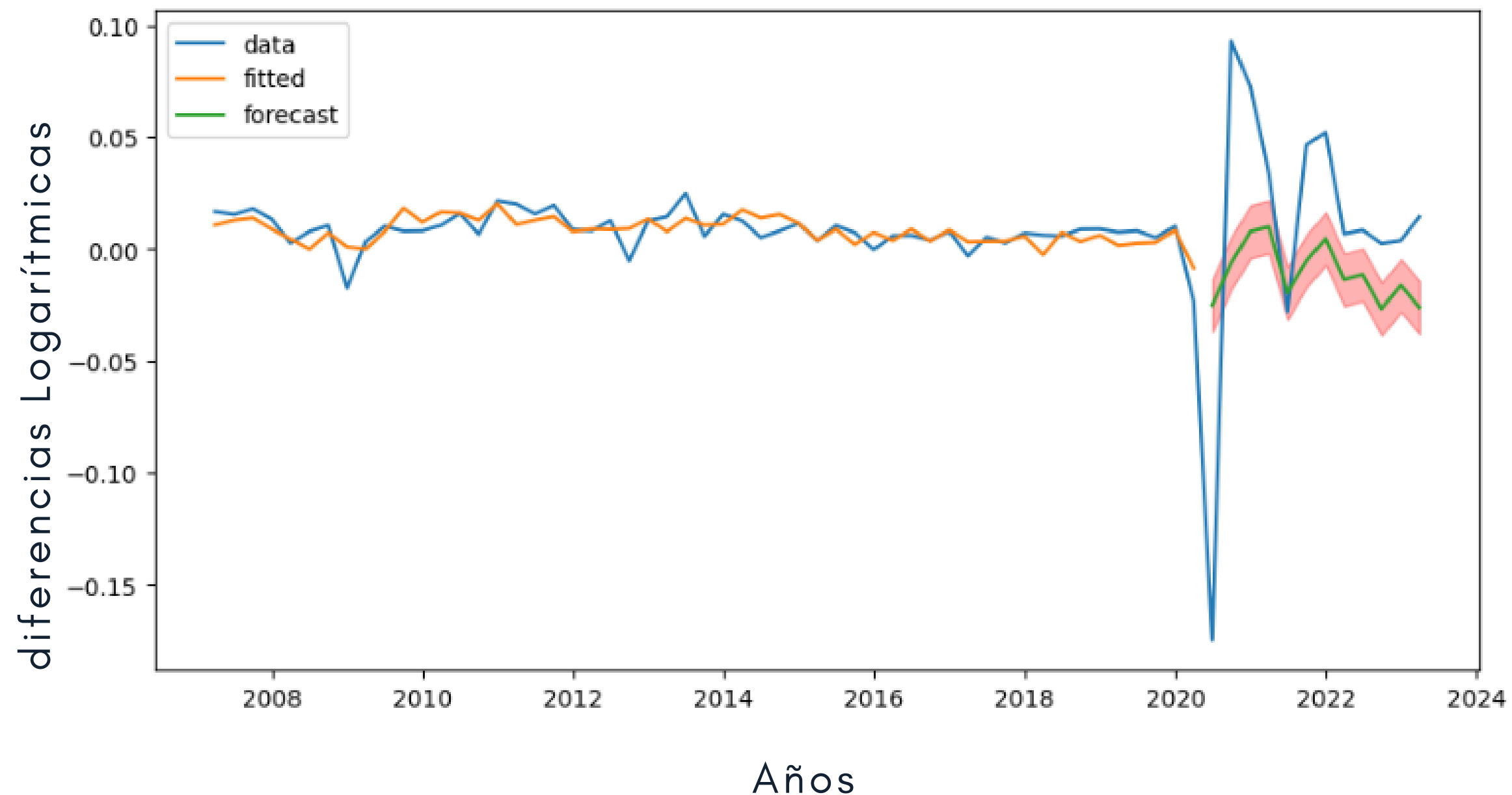
- Dentro e los tradicionales se obtiene como mejor modelo el modelo ARIMAX en el escenario 5
- El escenario 1 el mas favorable para los modelos e ML y escenario 5 para modelos tradicionales

ESCENARIO	REGRESIÓN	ARX	ARIMAX	AJUSTE POLINOMIAL	SVM	KNN
Escenario 1	Lineal: 3.947×10^{-3}	4.04×10^{-3}	4.04×10^{-3}	5.85×10^{-2}	4.2×10^{-3}	3.958×10^{-3}
Escenario 2	Lasso: 4.189×10^{-3}	4.30×10^{-3}	4.27×10^{-3}	22×10^{-2}	4.294×10^{-3}	4.290×10^{-3}
Escenario 3	Lasso: 4.189×10^{-3}	4.23×10^{-3}	4.26×10^{-3}	4.143×10^{-2}	4.286×10^{-3}	4.311×10^{-3}
Escenario 4	Ridge: 3.90×10^{-3}	3.96×10^{-3}	3.99×10^{-3}	4.131×10^{-2}	4.286×10^{-3}	4.118×10^{-3}
Escenario 5	Ridge: 3.84×10^{-3}	3.83×10^{-3}	3.81×10^{-3}	5.012×10^{-2}	4.286×10^{-3}	4.053×10^{-3}
Escenario 6	Ridge: 3.908×10^{-3}	3.84×10^{-3}	3.94×10^{-3}	3.63×10^{-2}	4.286×10^{-3}	4.049×10^{-3}

MULTI-VARIADO

MEJOR DE LOS MODELOS CON METODOS TRADICIONALES

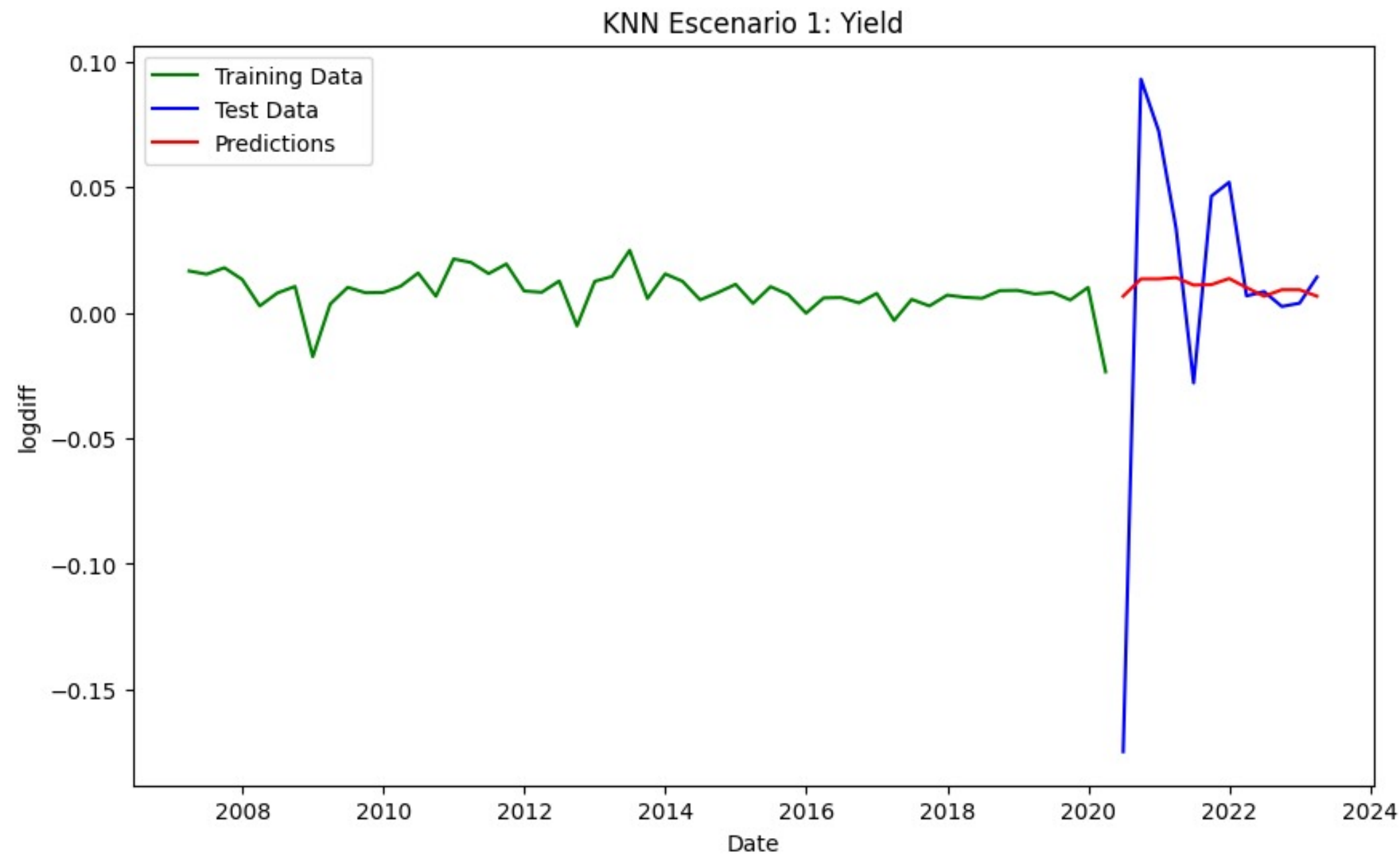
ARIMAX DIFERENCIA DEL PIB REAL CON AJUSTE ESCENARIO 5



- MSE: 3.81×10^{-3}
- ARIMA(0,0,0)(1,0,0)[12]
- Predictores escenario5:
Desempleo, IPC, CUM,
IBR, Nivel Yield,
Inclinación Yield y
Curvatura Yield

MULTI-VARIADO

MEJOR DE LOS MODELOS CON METODOS DE ML



- MSE: 3.958×10^{-3}
- $K = 2$
- Predictores escenario 1:
Tasa TES 1, 5 y 10 años

ESTUDIO DE REFERENCIA

GDP FORECASTING:
MACHINE LEARNING, LINEAR
OR AUTOREGRESSION?

Maccarrone G, Morelli G and
Spadaccini S (2021)

Resultados Paper:

	SARIMAX	LR	Mean LR-SARIMAX	ARX
Multi step				
Scenario 1	3,66e-03	2,50e-03	3,01e-03	4,29e-03
Scenario 2	3,99e-03	2,48e-03	3,13e-03	5,24e-03
Scenario 3	4,05e-03	2,38e-03	3,08e-03	4,69e-03
Scenario 4	3,97e-03	2,40e-03	3,06e-03	5,24e-03
Scenario 5	4,57e-03	2,37e-03	3,26e-03	4,84e-03
Scenario 6	3,97e-03	2,48e-03	3,13e-03	4,86e-03

- Dentro e los tradicionales se obtiene como mejor modelo el modelo ARIMAX en el escenario 5
- El escenario 1 el mas favorable para los modelos e ML y escenario 5 para modelos tradicionales

Resultados Nuestros:

ESCENARIO	REGRESIÓN	ARX	ARIMAX	AJUSTE POLINOMIAL	SVM	KNN
Escenario 1	Lineal: 3.947×10^{-3}	4.04×10^{-3}	4.04×10^{-3}	5.85×10^{-2}	4.2×10^{-3}	3.958×10^{-3}
Escenario 2	Lasso: 4.189×10^{-3}	4.30×10^{-3}	4.27×10^{-3}	22×10^{-2}	4.294×10^{-3}	4.290×10^{-3}
Escenario 3	Lasso: 4.189×10^{-3}	4.23×10^{-3}	4.26×10^{-3}	4.143×10^{-2}	4.286×10^{-3}	4.311×10^{-3}
Escenario 4	Ridge: 3.90×10^{-3}	3.96×10^{-3}	3.99×10^{-3}	4.131×10^{-2}	4.286×10^{-3}	4.118×10^{-3}
Escenario 5	Ridge: 3.84×10^{-3}	3.83×10^{-3}	3.81×10^{-3}	5.012×10^{-2}	4.286×10^{-3}	4.053×10^{-3}
Escenario 6	Ridge: 3.908×10^{-3}	3.84×10^{-3}	3.94×10^{-3}	3.63×10^{-2}	4.286×10^{-3}	4.049×10^{-3}

CONCLUSIONES

1

El mejor de todos los modelos estudiados es el KNN en el método univariado

MSE: 2.7×10^{-3}

2

Similitud de los resultados obtenidos para este estudio y el de referencia

El presente trabajo se desarrolla con el PIB Colombiano mientras que el documento de referencia estudia el PIB de EEUU

3

Se sugiere el uso de KNN para Modelos autorregresivos de la variable objetivo y ARIMAX para modelos multivariantes donde se incorpore información macroeconómica

**MUCHAS
GRACIAS POR
SU ATENCIÓN**