Predicción del PIB de Colombia: Técnicas Clásicas vs Machine *Learning*

Karla Victoria Torres Parra, Jonathan Javier Montes Castro, Cristian Castro Arias,
María Camila Lopera Giraldo, Victor Ricardo Uribe Durango.
{mcloperag, vruribed}@eafit.edu.co

Maestría en Ciencias de los Datos y Analítica

EAFIT

Introducción.

El presente estudio se centra en la comprensión y la predicción del comportamiento del Producto Interno Bruto (PIB) real de Colombia para el periodo entre 2007 a 2023. El PIB real, definido como el valor de mercado de todos los bienes y servicios finales producidos utilizando los factores de producción disponibles en un periodo específico, se considera una medida crucial para evaluar la actividad económica de una nación al igual que indicadores económicos como la desigualdad. Su cálculo se basa en el uso de precios constantes, lo que permite medir el cambio en la producción sin la interferencia de las variaciones en los precios.

El objetivo fundamental de esta investigación consiste en analizar en profundidad las tendencias y algunos de los factores que influyen en el PIB real, con el fin de desarrollar modelos predictivos de alta precisión capaces de estimar su comportamiento futuro. El logro de este objetivo es de suma importancia para la toma de decisiones por parte de los *"Policy Makers"* al igual que para el diseño de políticas económicas efectivas.

Para alcanzar dicho propósito, se ha llevado a cabo un exhaustivo análisis exploratorio utilizando una base de datos recolectada del gobierno y del Banco de la República de Colombia, la cual comprende información trimestral del PIB real desde el año 2005 hasta el primer trimestre de 2023. Este análisis ha incluido pruebas rigurosas de estacionariedad, así como la aplicación de diversas técnicas de transformación y reemplazo de valores faltantes, con el objetivo de garantizar la calidad integridad de los datos y determinación de posibles *outliers*, ya que como veremos en la sección de resultados, algunos modelos de *machine learning* son sensibles a este tipo de datos.

El análisis exploratorio ha revelado patrones de tendencia alcista en el PIB real a lo largo de los años, con la excepción notoria del segundo trimestre de 2020, en el cual se registró

una disminución significativa debido al impacto negativo del brote de la enfermedad por coronavirus (COVID-19). Además, se ha identificado la existencia de un componente estacional en los meses de septiembre y diciembre, en los cuales se han observado patrones recurrentes en la serie temporal.

En la siguiente etapa de la investigación, se emplearán diversas técnicas de modelado avanzado, entre las que se incluyen modelos ARIMA, regresiones lineales (incluyendo enfoques de Lasso y Ridge), máquinas de soporte vectorial (SVM) polinomiales, así como modelos KNN univariantes y multivariantes. Además, se utilizarán modelos ARX y SARIMAX, todos ellos considerando diferentes configuraciones de rezagos, para desarrollar modelos predictivos robustos y precisos. Estos modelos serán empleados para la estimación del PIB real en distintos escenarios, considerando variables como la curva cupón cero, macro variables y variables proxies.

Con el fin de mejorar aún más los resultados obtenidos, se llevarán a cabo acciones de mejora que incluirán la optimización de los modelos desarrollados, la incorporación de variables adicionales pertinentes y la implementación de técnicas de validación cruzada para evaluar la precisión y la generalización de los modelos propuestos.

Es importante destacar que el análisis y la predicción del PIB real tienen implicaciones significativas en la toma de decisiones económicas, la planificación estratégica y la evaluación de políticas públicas, permitiéndole al ente regulador modificar su política monetaria y fiscale con el fin de lograr sus objetivos macroeconómicos.

Marco Teórico.

El modelo de media móvil integrado autorregresivo (ARIMA) fue introducido al mundo de la estadística en 1970 por Box y Jenkins (Harvey, 1990). En este modelo, los autores toman

un modelo ARMA, que consta en una serie de tiempo en la que se le introduce la autocorrelación de una variable con ella misma a través del tiempo (modelo autorregresivo AR) y de promedio móvil (MA), y le hacen un ajuste tomando las diferencias de las series para obtener una serie estacionaria. (Chávez Quisbert, 1997). Estos modelos son conocidos como modelos ARIMA (P, D, Q) donde las variables en el paréntesis significan: número de términos autorregresivos, número de diferencias y número de medias móviles.

La utilidad de estos modelos se ha visto muy encadenada a la predicción de series de tiempo en múltiples áreas de conocimiento. Fattah, Ezzine, Aman et al. (2018) utilizaron este tipo de modelo para predecir la demanda de una compañía del sector de alimentos, ayudando a los directivos de la compañía con los lineamientos de producción de esta. Wahyudi (2017) también hizo uso de un modelo ARIMA, pero esta vez para la predicción del valor del índice compuesto de precios de acciones de Indonesia (CSPI) concluyendo que el modelo resulta de mucha utilidad para la predicción de corto plazo. Por otro lado, la aplicación de este modelo también es de utilidad en áreas como la medicina donde autores como Katoch y Sidhu (2021) utilizaron este tipo de modelo para predecir la dinámica del COVID19 en India con el fin de que las instituciones de salud pública del país se beneficiarán planificando la asignación de productos de una manera más adecuada.

Igualmente, este tipo de series de tiempo tienen diferentes variaciones que fueron utilizadas en el presente análisis; ARX (AR con variables exógenas) y SARIMAX (ARIMA estacional con variables exógenas). Este tipo de modelos también han sido ampliamente utilizados para la predicción, ejemplo es el estudio de Kwon, Cho y Na (2016) en el cuál los autores generaron un modelo ARX capaz de predecir la tasa de desempleo a partir de comentarios en redes sociales. De igual manera, el modelo SARIMAX fue bastante efectivo

al tener en cuenta todos los factores externos que influyen en la demanda de alimentos perecederos para prever las ventas diarias de estos alimentos en un comercio minorista en el 2016 cuando Arunraj, Ahrens, y Fernandes (2016) hicieron uso de este modelo.

Con respecto a las regresiones lineales, hay dos métodos de regularización conocidos como la regresión de Ridge y Lasso; estos se utilizan principalmente para resolver uno de los problemas más comunes de las regresiones, el sobreajuste, además que este tipo de técnicas buscan proporcionar un mejor rendimiento cuando un modelo se vuelve muy complicado. La diferencia entre estas es que la regresión Ridge tiene un elemento de penalización en la función de costo de la regresión lineal y reduce el tamaño de los coeficientes, por su lado Lasso, aunque también tiene un término de penalización este es proporcional al total de los valores absolutos de los coeficientes. Esto hace que algunos de los coeficientes se acerquen al 0, lo que hace que algunos aspectos del modelo sean completamente irrelevantes (Sharma, 2023). Este tipo de regresiones han ayudado a la predicción de fracaso corporativo, especialmente en el caso de Pereira, Basto and Silva (2016) que puede ser de interés para inversores, acreedores, empresas prestatarias y gobiernos.

El algoritmo de K vecinos más cercanos (KNN por sus siglas en inglés) es comúnmente conocido por su uso en clasificación de aprendizaje supervisado y para regresión. Este consta de almacenar una colección de ejemplos donde cada uno contiene características descriptivas de cada ejemplo y su clase asociada (clasificación o “label”) o valor numérico (predicción). Para encontrar el número óptimo de K vecinos más cercanos con el objetivo aumentar la precisión de los resultados, se utiliza un método denominado *Cross Validation.* Esto consta de dividir los datos en tres partes, entrenamiento, validación cruzada y prueba; los datos de entrenamiento ayudan a encontrar los vecinos más cercanos y en la etapa de

validación cruzada se encuentra el mejor valor de K que se va a validar con la etapa de prueba. (2020).

El KNN es comúnmente conocido por su uso en clasificación de aprendizaje supervisado, sin embargo, se le ha encontrado utilidad en la predicción de las series de tiempo. Tajmouati, Wahbi, Bedoui et al. (2021) hacen uso de esta metodología para analizar la producción de leche en EE. UU. y en el Reino Unido para demostrar la aplicación y la eficiencia de la metodología con ciertas técnicas adicionales para la escogencia de parámetros. Y de igual manera que el método previamente mencionado, ha sido utilizado en series de tiempo relacionadas a los índices de mercados como lo hizo Ban, Zhang, Pang et al. (2013) con el S&P 500; estos encontraron bastante utilidad en el método ya que sus resultados experimentales mostraron que un enfoque multivariante KNN proporciona una precisión de pronóstico mejorada y encuentran que la regresión KNN univariante no proporciona esto. Observaremos el gran aporte de este algoritmo en este trabajo investigativo y como en el caso univariado y en los escenarios de los multivariantes, los resultados que arroja el KNN aplicado a nuestros datos predictores para predecir el PIB con corregimiento estacional como variable objetivo, son óptimos en términos de las medidas de error entregadas por el MSE y MAE.

Desarrollo Metodológico.

Entendimiento del problema.

El Producto Interno Bruto (PIB) de un país es de gran interés para los agentes del mercado ya que les permite tomar decisiones a nivel socioeconómico con el objetivo de maximizar sus beneficios u objetivos propios, además de que permite dar una idea general de indicadores sociales como la desigualdad; entre los diferentes agentes del mercado están

los gobiernos, los inversionistas (externos e internos) y los ciudadanos. Es por esto por lo que el principal objetico que se quiere resolver en este análisis es la predicción PIB de Colombia (específicamente el Real) teniendo en cuenta diferentes variables macroeconómicas y financieras y, además, determinar cuáles técnicas de modelación predictiva tienen buena precisión para estimar su comportamiento futuro.

Análisis Exploratorio de los datos.

Entendimiento de los datos.

En este análisis, se utilizaron datos recopilados de dos entidades gubernamentales, DANE y el Banco de la República, abarcando el periodo comprendido desde el primer trimestre de 2007 hasta el segundo trimestre de 2023. La variable objetivo es el Producto Interno Bruto (PIB) real con un ajuste estacional, siendo un variable numérica continua cuya periodicidad es trimestral, mientras las demás variables utilizadas en este análisis abarcan dos categorías principales las cuales son macroeconómicas y financieras. El objetivo principal es comprender en profundidad las tendencias, patrones y características del PIB a lo largo del tiempo. A continuación, se presentan las diferentes variables utilizadas en el análisis junto con una breve descripción.

1. **PIB Real:** la medida de actividad económica por excelencia de un país. La palabra "Real" hace referencia al cálculo ya que se utilizan precios de un año base para que estos sean constantes y así lograr que esta medida no se vea afectada por la inflación. Esta es una variable numérica de tipo continua y es la variable objetivo, hace parte de la categoría de variables macroeconómicas.

2. **Desempleo:** el número de personas que se encuentran en edad para trabajar y hacen parte de la fuerza de trabajo, pero no reciben un salario ni por parte de una empresa ni por trabajo independiente. Su categoría es macroeconómica y es una variable numérica continua.

3. **IPC:** índice de precios al consumidor, este es un indicador que mide la variación de los precios de los bienes y servicios más representativos de los hogares. Variable numérica de tipo continua cuya categoría es macroeconómica.

4. **Variación del IPC:** muestra el cambio de precios de canasta representativa de los hogares en un periodo de tiempo. Variable numérica de tipo continua cuya categoría es macroeconómica.

5. **Capacidad de Utilización de Manufacturas:** la máxima producción posible de una empresa manufacturera, medida en unidades de producción por período. Su categoría es macroeconómica y es una variable numérica continua.

6. **Tasa de Intervención del Banco de la República:** es el instrumento de política monetaria que usa el Banco de la República cuyo objetivo es tener un efecto que permita al ente estatal controlar la tasa de inflación. Esta es una variable numérica y su categoría es macroeconómica.

7. **TES 1 año:** la tasa de interés que pagan los bonos del estado en el mercado de renta fija con plazo de un año, se le considera una tasa de interés de corto plazo. Esta es una variable numérica y su categoría es financiera.

8. **TES 5 años:** la tasa de interés que pagan los bonos del estado en el mercado de renta fija con plazo de 5 años, se le considera una tasa de interés de mediano plazo. Esta es una variable numérica y su categoría es financiera.

9. **TES 10 años:** la tasa de interés que pagan los bonos del estado en el mercado de renta fija con plazo de 10 años, se le considera una tasa de interés de largo plazo. Esta es una variable numérica y su categoría es financiera.

10. **Nivel Yield:** medido como el promedio de las tasas de los bonos del estado de corto, mediano y largo plazo. categoría financiera.

11. **Pendiente Yield:** diferencia entre los rendimientos de corto y largo plazo. Esta es una variable numérica continua y su categoría es financiera.

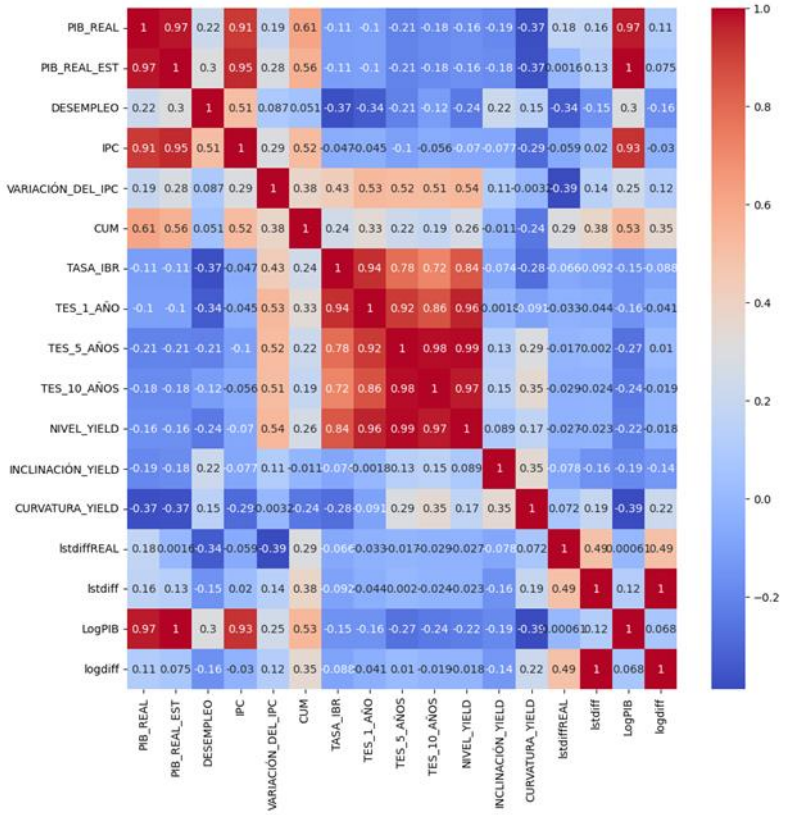
12. **Curvatura Yield:** doble producto del rendimiento de mediano plazo menos el de más corto y más largo plazo. Variable numérica continua y con categoría financiera.

Preparación de los datos.

Los datos fueron preparados mediante un exhaustivo proceso de Extracción, Transformación y Carga (ETL). En primer lugar, se llevó a cabo la extracción manual de los datos de dos fuentes principales: el Departamento Administrativo Nacional de Estadística (DANE) y el Banco de la República. Estos datos fueron recopilados y posteriormente unificados en un conjunto de datos común, que se almacenó en un archivo de Excel alojado en el servicio de nube proporcionado por Google.

Posteriormente, se procedió a realizar un proceso de transformación de los datos. En primer lugar, se aplicaron diversas técnicas de transformación a la variable objetivo con el objetivo de cumplir con la hipótesis de estacionariedad para una serie de tiempo. Esta etapa implicó la aplicación de métodos como la diferenciación de logaritmos u otras transformaciones específicas según las características de la variable en cuestión.

Además, se llevó a cabo una transformación sobre las variables macroeconómicas y financieras utilizadas en el análisis. Estas variables fueron escaladas con el propósito de llevarlas a una magnitud similar, lo que facilita la comparación y la correcta interpretación de su influencia en el análisis de la serie de tiempo. Las variables fueron estandarizadas por medio del paquete *standar scaler* de la librería de Sklearn de python, la cual re-escala las variables por medio de la resta de cada dato con la media de la variable en cuestión sobe su desviación estándar. Algo importante a mencionar es que se encontró que no todas las variables tenían una alta correlación lineal y dado a este escalamiento de los datos decidimos solo escoger la métrica Euclídea y no de Mahalanobis para la implementación del KNN en nuestra variable objetivo. Esto dado a que no fue necesario por la no fuerte correlación en la mayoría de las variables (ver mapa de correlaciones).



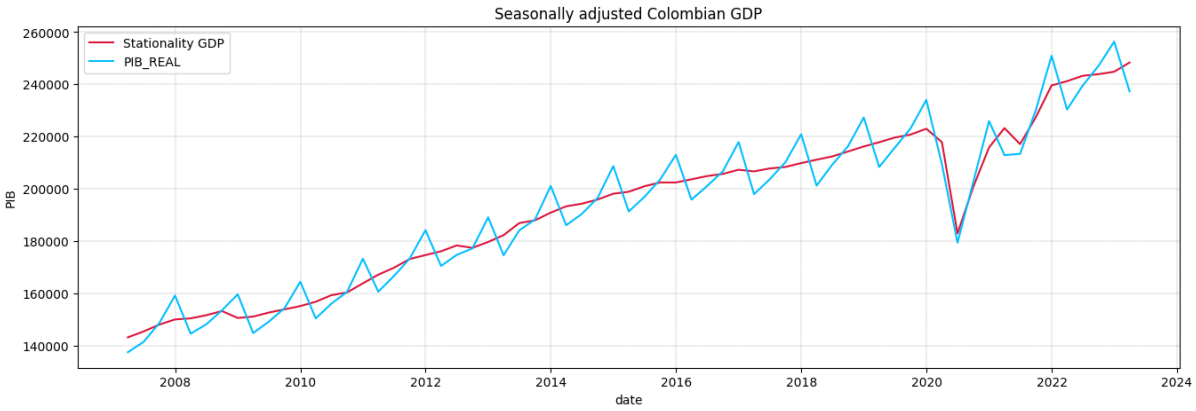
Gráfica 1. Matriz de correlaciones lineales entre variables.

Finalmente, el proceso de carga de los datos se realizó en diferentes momentos del análisis, lo cual implicó una conexión directa entre el entorno utilizado y el servicio de nube ofrecido por Google. Esta conexión aseguró la integridad y disponibilidad de los datos en todo momento, permitiendo un flujo eficiente y seguro para su posterior análisis.

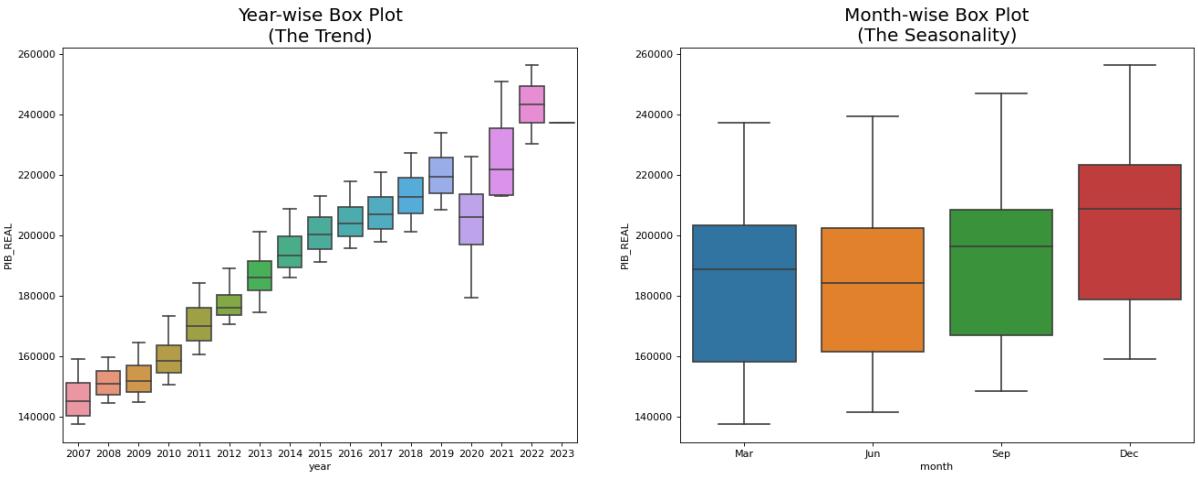
En resumen, la preparación de los datos involucró un minucioso proceso de ETL, desde la extracción manual de las fuentes primarias hasta la transformación de las variables objetivo y macroeconómicas/financieras, y finalmente, su carga en el entorno de análisis a través de una conexión directa con el servicio de nube de Google. Estos pasos garantizaron la calidad, coherencia y confiabilidad de los datos utilizados en el estudio de series de tiempo.

Análisis descriptivo de los datos y hallazgos importantes.

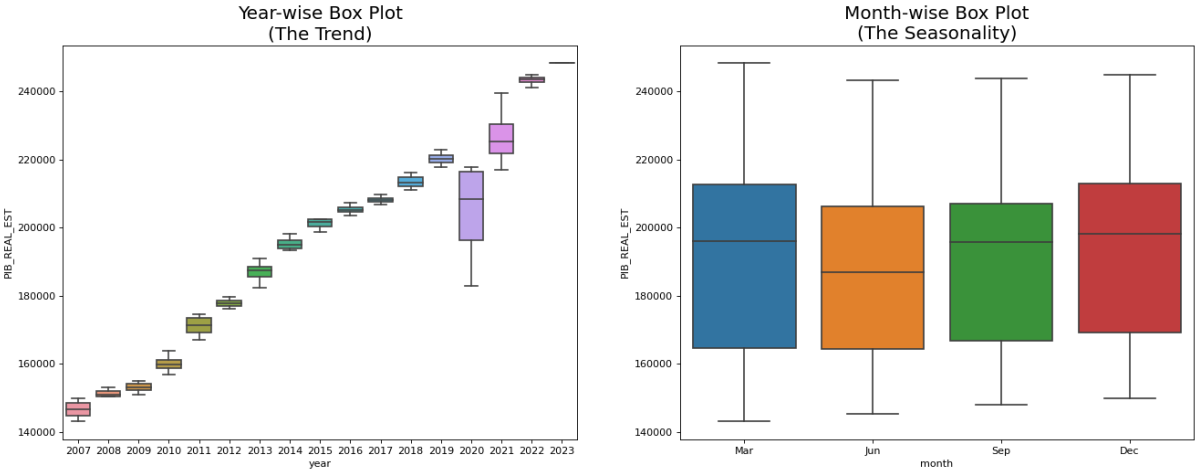
Uno de los primeros pasos en el momento de trabajar con una serie de tiempo es identificar sus diferentes componentes, en este caso se decidió hacer un análisis a las variables del PIB real y el PIB real con ajuste estacional. La gráfica 1 muestra la comparación entre los datos entregados por el Banco de la República desde el primer trimestre de 2007. En esta se observa a primera instancia como la variable del PIB real presenta tanto un componente tendencial alcista durante todo el periodo, excepto el primer trimestre de 2020, año relacionado a la pandemia del Covid-19, como un componente estacional donde se identifica cómo existe una caída del PIB real en el primer trimestre de cada año, mientras que por su parte la variable del PIB real con ajuste estacional sólo presenta un componente tendencial alcista similar.



Gráfica 2. Comparación entre el PIB real y el PIB con corregimiento estacional.



Gráfica 3. Comparación año a año del PIB real.

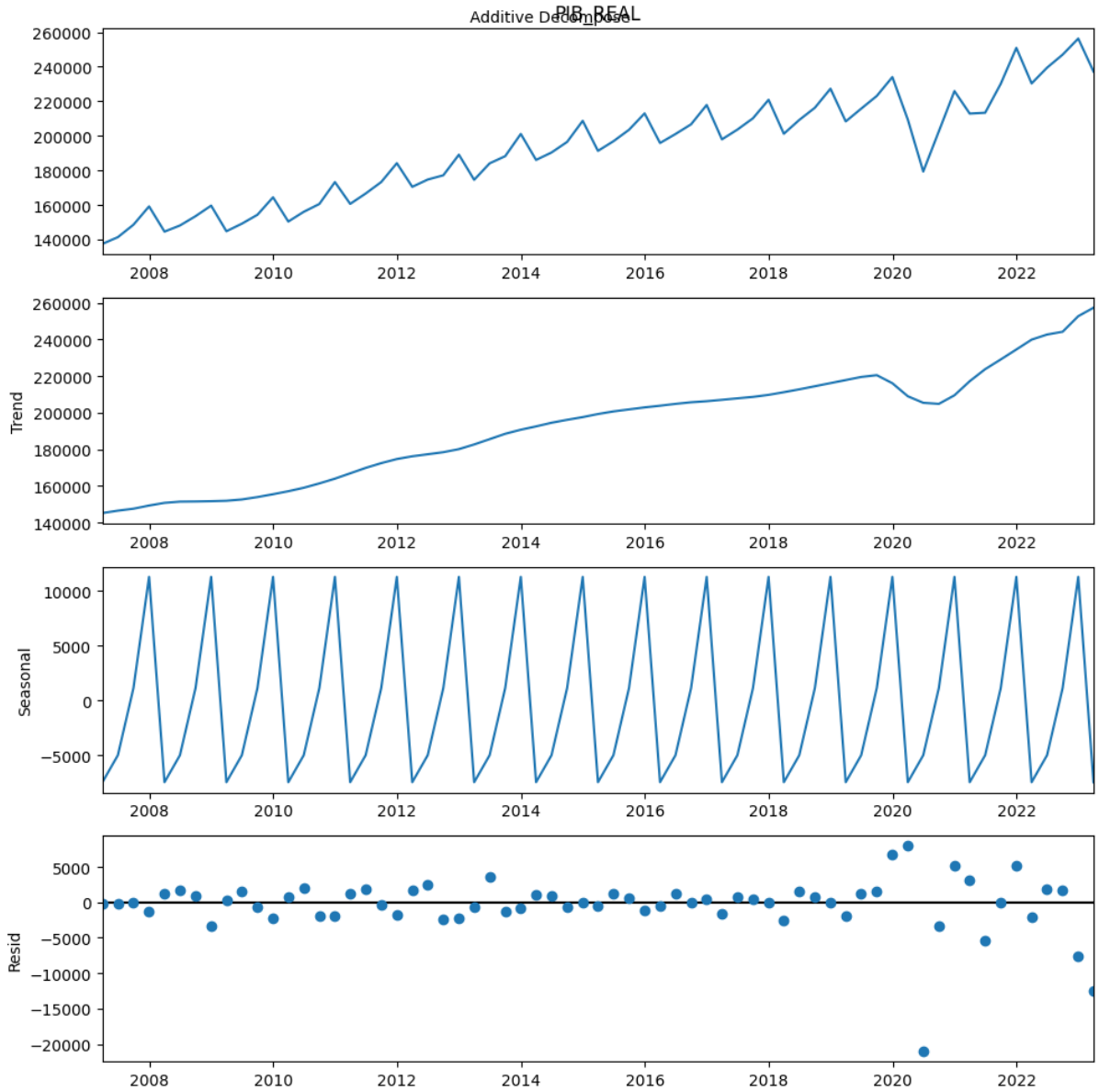


Gráfica 4. Comparación año a año del PIB real con ajuste estacional.

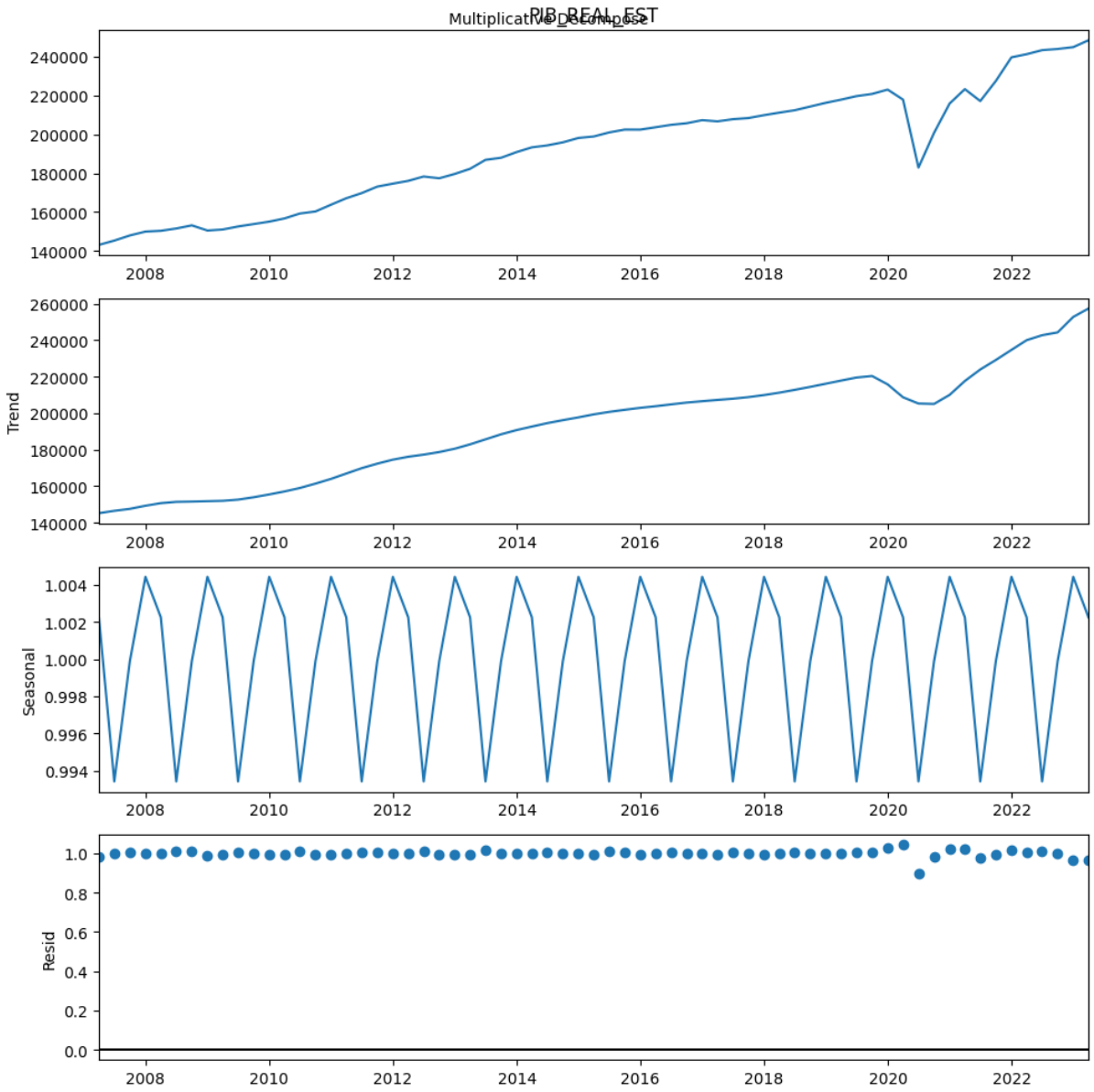
En las gráficas 2 y 3 se puede observar cómo en ambos casos ambas variables objetivo poseen un componente de tendencia alcista y lo que parecía en una primera instancia como datos atípicos en el primer y segundo trimestre de 2020 se confirma que no lo son al no estar por fuera de los bigotes de las cajas en ambos casos. Para asegurar el rechazo de posibles *outliers* se empleó también el método heurístico descrito por Yan (2012) y que fue adoptado por Martínez et al (2019) el cual considera que un valor es atípico si su valor

absoluto es cuatro veces mayor que el máximo del valor absoluto de la mediana de los tres puntos consecutivos que se encuentran antes y después de nuestra observación. De igual forma se observa de nuevo como el componente estacional fue eliminado en la serie de tiempo del PIB real con ajuste ya que no es tan notoria la diferencia entre las gráficas de caja.

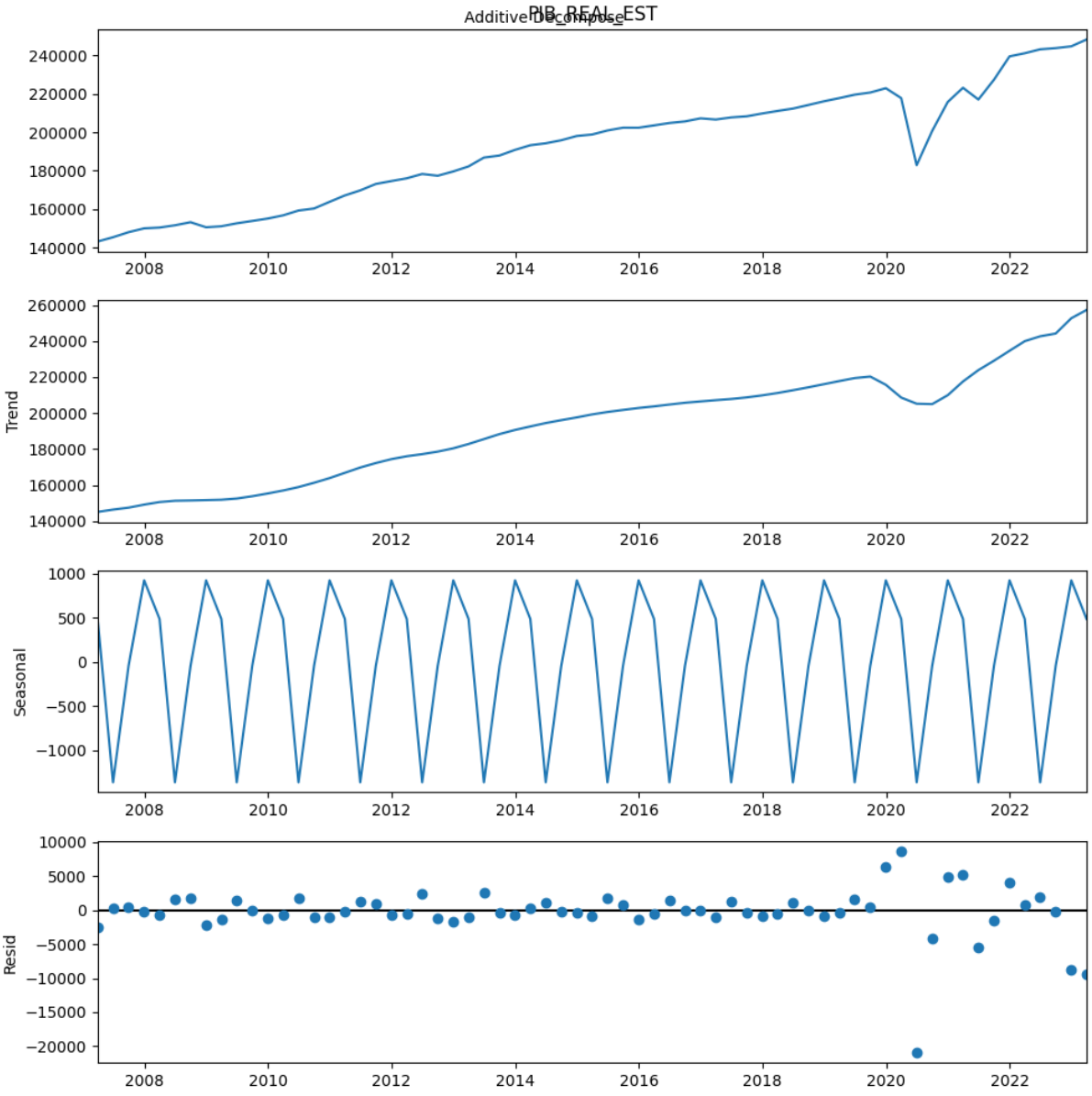
Por su lado al momento de obtener la descomposición de ambas series de tiempo se observa como ambas series parecen seguir una descomposición aditiva (ver gráficas 5 y 6), a primera instancia se observa como el componente tendencial y estacional no varía en ninguna de las descomposiciones, donde se encuentra la mayor diferencia es en el componente residual, ambas parecen seguir un patrón pero se observa como en la descomposición aditiva hacia el final de la serie esta deja de seguir un patrón y empieza a ser más aleatoria. Usando estas descomposiciones de la serie de tiempo aditiva, procedimos a utilizar la parte desestacionalizada de la serie PIB real y se procedió a sacar el logaritmo de estos datos desestacionalizados con el objetivo de aplicar la prueba de Dickey-Fuller al tener una variable con corrección de escala. Dicha variable aprobó dicha prueba y cuando se procedió a graficar, obtuvimos una gráfica muy similar a la obtenida con los datos del PIB que entrega el banco de la república con ajuste estacional (ver gráficas 2 y 5). Esto nos dio la seguridad de que estos datos con corrección estacional son confiables.



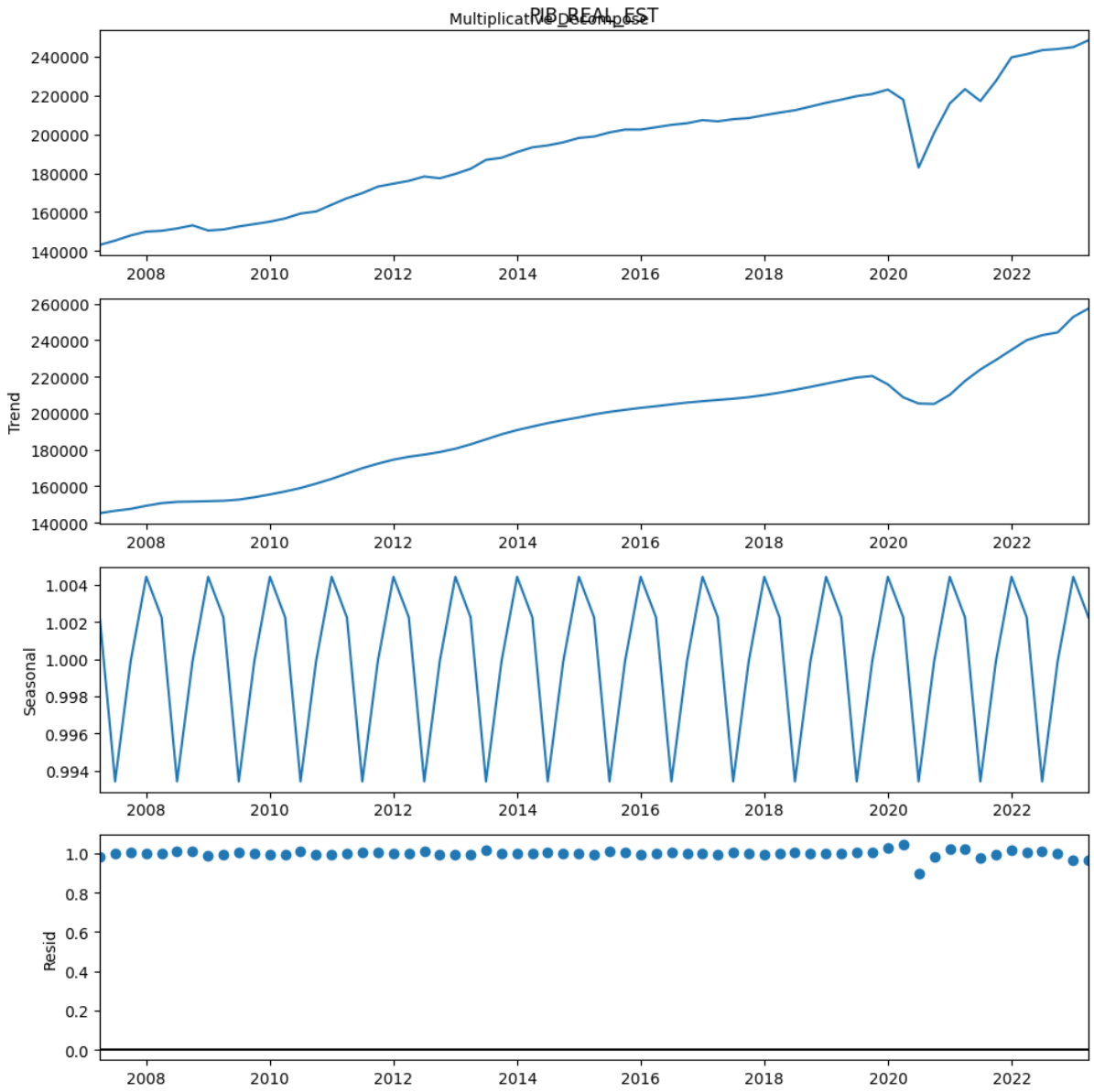
Gráfica 5. Descomposición Aditiva y Multiplicativa del PIB real.



Gráfica 5. Descomposición Aditiva y Multiplicativa del PIB real.



Gráfica 5. Descomposición Aditiva y Multiplicativa del PIB real ajustado.



Gráfica 5. Descomposición Aditiva y Multiplicativa del PIB real ajustado.

Con el fin de evaluar la estacionariedad de la serie de tiempo, se aplicaron las pruebas de Dickey-Fuller y KPPS. Sin embargo, los resultados iniciales sobre ambas series de tiempo no proporcionaron pruebas suficientes de estacionariedad. Por esta razón se decidió aplicar diferentes transformaciones sobre ambas series de tiempo buscando volver las series estacionarias, en primer lugar, se aplicó una diferencia sobre los datos, luego un logaritmo y por último una diferencia sobre el logaritmo de la serie de tiempo, es importante señalar

que esta transformación implicó la pérdida del primer dato tanto para el PIB real como para el logaritmo. Para abordar esta pérdida de datos y mantener la continuidad en el análisis, se procedió a reemplazar el primer valor faltante en cada variable. Este reemplazo se realizó utilizando la media de los tres datos siguientes correspondientes a cada predictor. Por último, se volvió a comprobar la estacionariedad de la serie con las diferentes transformaciones realizadas y se encontró al imputar el primer dato nulo de las diferencias del PIB real y de las diferencias logarítmicas se cumplía el rechazo de la hipótesis nula (ver gráfica 7).

```python
from statsmodels.tsa.stattools import adfuller, kpss

# ADF Test (Test de Dickey  Fuller)
result = adfuller(df1.logdiff.values, autolag='AIC')
print(f'ADF Statistic: {result[0]}')
print(f'p-value: {result[1]}')
for key, value in result[4].items():
    print('Critial Values:')
    print(f'   {key}, {value}')

# KPSS Test
result = kpss(df1.logdiff.values, regression='c')
print('\nKPSS Statistic: %f' % result[0])
print('p-value: %f' % result[1])
for key, value in result[3].items():
    print('Critial Values:')
```

```
ADF Statistic: -3.2856552800556766
p-value: 0.015531987650177026
Critial Values:
   1%, -3.542412746661615
Critial Values:
   5%, -2.910236235808284
Critial Values:
   10%, -2.5927445767266866

KPSS Statistic: 0.380768
p-value: 0.085445
Critial Values:
```

Gráfico 7. Resultados obtenidos de las diferentes pruebas de estacionariedad.

Sin embargo, al usar el PIB real ajustado pudimos observar mejores valores respecto p-value. Por esta razón decidimos trabajar con el PIB real con ajuste estacional con el propósito de no tener dudas respecto al rechazo de la hipótesis nula ya que los valores del PIB real ajustado con una diferencia o la diferencia del logaritmo cumplía con ser estacionaria con valores del *p-value* muy bajos, del orden de 10^{-11} tanto para las diferencias

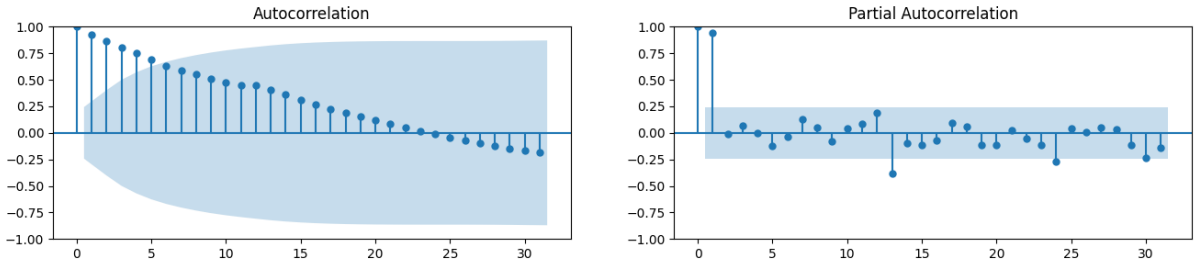
de esta variable objetivo como las diferencias logarítmicas (ver gráfico 8). Como valor agregado se descompuso la serie del PIB real eliminando el componente estacional y se volvió a sacar el resultado de las diferentes pruebas, dando como resultado que la serie es estacionaria.

```
Análisis para el PIB Real con Ajuste Estacional

Análisis para una diferencia
ADF Statistic: -2.6201240724210884
p-value: 0.08890894941729538

KPSS Statistic: 0.180900
p-value: 0.100000
---------------------------------------------------------------------------
Análisis para la diferencia del logaritmo
ADF Statistic: -2.5226308029754874
p-value: 0.11006465035935897

KPSS Statistic: 0.234055
p-value: 0.100000
Análisis para el PIB Real con Ajuste Estacional
---------------------------------------------------------------------------
Análisis para una diferencia
ADF Statistic: -7.556424233183984
p-value: 3.0908669371372585e-11

KPSS Statistic: 0.092937
p-value: 0.100000
---------------------------------------------------------------------------
Análisis para la diferencia del logaritmo
ADF Statistic: -7.530676147765089
p-value: 3.58555834534179e-11

KPSS Statistic: 0.108138
p-value: 0.100000
```

Gráfico 8. Resultados obtenidos de las diferentes pruebas de estacionariedad.

Por último, se realizó un análisis de la significancia de los rezagos sobre la serie de tiempo con ajuste estacional, observando como en la autocorrelación los primeros 4 rezagos son significativos mientras en que la autocorrelación parcial solo los dos primeros son

significativos, con la sorpresa de hallar un rezago algo con un nivel de significancia no muy elevado en el rezago 14.



Gráfica 9. Análisis de la autocorrelación y autocorrelación parcial del PIB real ajustado.

Selección de Modelos. (supervisados y no supervisados), Ingeniería de Características, Entrenamiento, Evaluación.

Modelos.

En el contexto teórico, se han explorado distintos modelos para el análisis y la predicción de series de tiempo, incluyendo el enfoque univariado y multivariado del algoritmo K vecinos más cercanos (KNN), el modelo de media móvil integrado autorregresivo (ARIMA) y sus extensiones ARIMAX, SARIMA y SARIMAX.

El KNN multivariado se emplea tanto para clasificación como para regresión en el ámbito de las series de tiempo. Su aplicación se ha evidenciado en diversas investigaciones, como el análisis de la producción lechera en EE. UU. y el Reino Unido, donde ha demostrado su utilidad y eficacia al considerar múltiples variables relevantes.

Por otra parte, el KNN multivariante también se ha utilizado en la predicción de series de tiempo, como en el caso del índice S&P 500. En este enfoque, se busca a los vecinos más cercanos utilizando una única variable para llevar a cabo la predicción. Utilizamos distancia Euclidiana como se mencionó anteriormente. Los hiper-parámetros a tener en cuenta en los

modelos de KNN fue el vecino (k) más cercano y la distribución de los pesos por distancia inversa en donde los vecinos más cercanos tienen un mayor peso.

El modelo ARIMA, por su parte, constituye un método estadístico ampliamente reconocido y aplicado en el análisis y la predicción de series de tiempo. Ha demostrado ser efectivo en diversos estudios, adaptándose a diferentes combinaciones de términos autorregresivos, diferencias y medias móviles.

Asimismo, se han considerado las extensiones del modelo ARIMA, como el ARIMAX, que permite la inclusión de variables exógenas en el modelo, y el SARIMA y SARIMAX, que incorporan componentes de estacionalidad en los datos.

Por último, se ha hecho mención del modelo ARX, una extensión del ARIMA que se enfoca exclusivamente en términos autorregresivos y variables exógenas.

Estos modelos han sido seleccionados debido a su relevancia y eficacia en la predicción de series de tiempo en diversas áreas de estudio, lo que justifica su elección en el presente proyecto.

KNN multivariado

En el caso del KNN multivariado, se llevaron a cabo múltiples validaciones cruzadas con el fin de determinar el valor óptimo de K. Los resultados revelaron que diferentes escenarios presentaban preferencias distintas en cuanto a la elección de K. Para algunos escenarios, se obtuvieron mejores resultados con K=2, mientras que en otros casos se encontró que K=7 o K=11 eran las opciones más adecuadas. Dichos cálculos de este hiper-parámetro se ejecutaron cambiando el número de divisiones del dataset. Se obtuvieron

dichos valores de K para los diferentes escenarios al dividir el dataset (parámetro *Cross Validation*) en 5, 7 y 10 obteniendo el mismo valor K.

A pesar de estas variaciones, en términos generales, el escenario 1 del KNN multivariado con K=2 demostró tener el menor error cuadrático medio (MSE=0.003958), seguido muy de cerca por el escenario 6, que consideraba todas las variables y también utilizaba K=2, obteniendo un MSE de 0.00403. Estos hallazgos son altamente alentadores, ya que tanto el KNN univariado como el multivariado exhibieron las mejores métricas de desempeño con K=2.

Es igualmente significativo resaltar que el KNN multivariado logró superar por una mínima diferencia a la mejor regresión lineal obtenida. En el escenario 4, mediante la aplicación de la técnica de regularización Ridge, se alcanzó un MSE de 0.003962. Es importante mencionar que para cada escenario de las regresiones lineales se estimó como mejor modelo (lineal, Ridge y Lasso) a la mejor regresión en términos de las métricas MAPE, MSE (de los datos de prueba) y R^2. Esto cumpliendo el marco teórico del equilibrio entre estas métricas de sesgo y variabilidad para determinar el mejor modelo en una regresión lineal.

Estos resultados enfatizan la efectividad y el potencial del enfoque del KNN multivariado en el ámbito de la predicción de series de tiempo, superando a otros modelos como la regresión lineal. La elección del valor K=2 como la opción más favorable resalta la importancia de considerar la proximidad de los vecinos en el proceso de predicción, revelando un patrón consistente en cuanto a la relevancia de esta configuración en el contexto de los modelos KNN. El KNN se comporta como un regresor inteligente, en el qua a partir de distancias Euclidianas más cercanas entre los predictores, corrige y establece

mínimos óptimos para la variable objetivo. Cabe recalcar que las métricas utilizadas para cada escenario en KNN multivariado fueron MSE y MAE.

Evaluación.

Durante la evaluación se decidió calcular varios modelos los cuales se dividen en dos grupos, los modelos univariados y los multivariados, los primeros son los modelos cuya variable independiente es la misma variable dependiente rezagada mientras que los multivariados son modelos que incluyen el rezago de la variable dependiente y un conjunto de variables X's diferentes a la dependiente, es importante especificar que en para el grupo de modelos multivariados se decidió seguir la misma metodología de escenarios del trabajo de Maccarrone et al. (2021) donde para cada modelo multivariado se evaluó un conjunto de 6 escenarios los cuales son: curva Yield, variables macroeconómicas y curva Yield, variables macroeconómicas, variables proxies, variables macroeconómicas y proxies, una combinación de todas las anteriores. Dicha metodología obedece y es coherente con las metodologías *backward* y *forward-step wise selection* de la ingeniería de características, paso muy importante a la hora de utilizar las diferentes técnicas y algoritmos de *machine learning*, ya que como podemos observar el escenario 6 contiene todas las variables y los demás escenarios contienen menos variables en una mezcla de estos que son coherentes con los conceptos financieros ya expuestos. Además, es de suma importancia mencionar que para **todos** los modelos el *dataset* fue divido en un 80% en datos de entrenamiento y un 20% en datos de prueba, *i.e*, datos que el modelo entrenado no mira para luego comparar dichos datos con las predicciones con el modelo. Para poder determinar el modelo cuya predicción sea la más cercana a la real se decidió usar la métrica del Error Cuadrático Medio (MSE), al igual que se usa en el trabajo de Maccarrone et al. (2021). El MSE es

ampliamente utilizado y su principal objetivo es medir el promedio de los errores al cuadrado, donde los errores al cuadrado se calculan como la diferencia entre el valor estimado y el valor verdadero, todo elevado al cuadrado.

Análisis y Conclusiones.

Después de haber obtenido todas las predicciones de todos los modelos y los diferentes escenarios y se recopilaron los datos de los errores cuadráticos medios en dos tablas. La tabla 1 recoge todos los MSE's para los modelos multivariados en cada uno de los escenarios mientras que en la tabla 2 se encuentran todos los MSE's para los modelos univariados.

ESCENARIO	REGRESIÓN	ARX	ARIMAX	AJUSTE POLINOMIAL	SVM	KNN
Escenario 1	Lineal: $3.947x10^{-3}$	$4.04x10^{-3}$	$4.04x10^{-3}$	$5.85x10^{-2}$	$4.2x10^{-3}$	$3.958x10^{-3}$
Escenario 2	Lasso: $4.189x10^{-3}$	$4.30x10^{-3}$	$4.27x10^{-3}$	$22x10^{-2}$	$4.294x10^{-3}$	$4.290x10^{-3}$
Escenario 3	Lasso: $4.189x10^{-3}$	$4.23x10^{-3}$	$4.26x10^{-3}$	$4.143x10^{-2}$	$4.286x10^{-3}$	$4.311x10^{-3}$
Escenario 4	Ridge: $3.90x10^{-3}$	$3.96x10^{-3}$	$3.99x10^{-3}$	$4.131x10^{-2}$	$4.286x10^{-3}$	$4.118x10^{-3}$
Escenario 5	Ridge: $3.84x10^{-3}$	$3.83x10^{-3}$	$3.81x10^{-3}$	$5.012x10^{-2}$	$4.286x10^{-3}$	$4.053x10^{-3}$
Escenario 6	Ridge: $3.908x10^{-3}$	$3.84x10^{-3}$	$3.94x10^{-3}$	$3.63x10^{-2}$	$4.286x10^{-3}$	$4.049x10^{-3}$

Tabla 1. Errores cuadráticos medios para cada modelo multivariado en cada escenario.

MODELO	MSE
AR (2)	$4.22x10^{-3}$
AR (4)	$4.23x10^{-3}$
ARIMA	$4.03x10^{-3}$
KNN	$2.7x10^{-3}$

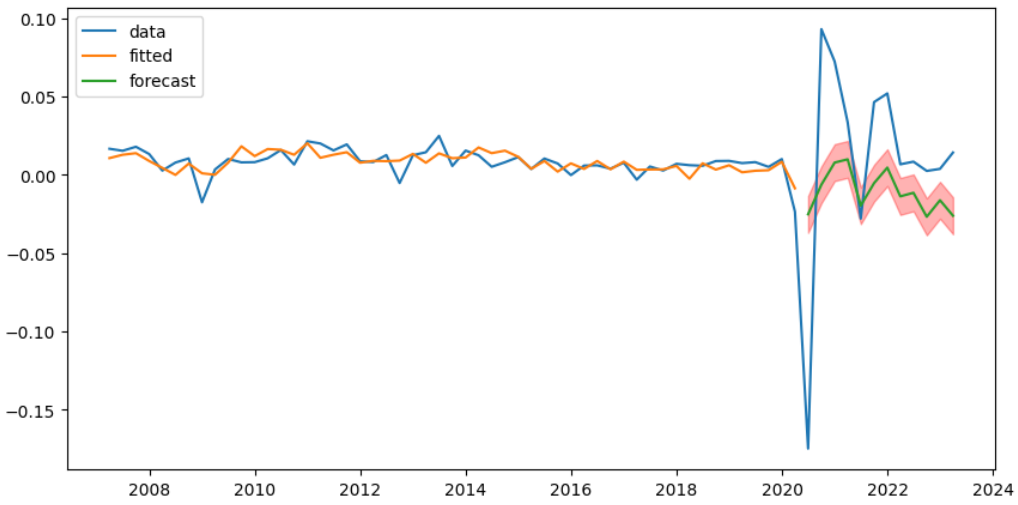
Tabla 2. Errores cuadráticos medios para cada modelo univariado.

De la tabla 1 se observa primero como al usar modelos multivariados el modelo con menor error cuadrático medio es el ARIMAX cuando se usa el escenario 5 con un MSE de 0.00381, después de este se encuentran los modelos de ARX con escenario 5 y 6, la

regresión Ridge en los últimos tres escenarios y el KNN con escenario 1 junto con el ARIMAX con escenario 6, también cabe destacar que todos estos modelos en estos escenarios tienen un rango de valores muy cercanos. Es importante destacar como en los modelos ARX, Ridge y ARIMAX destacan sobre todo los escenarios 5 y 6 mientras en el KNN el 1. También se observa como los peores resultados los obtuvo el ajuste polinomial en todos los escenarios.

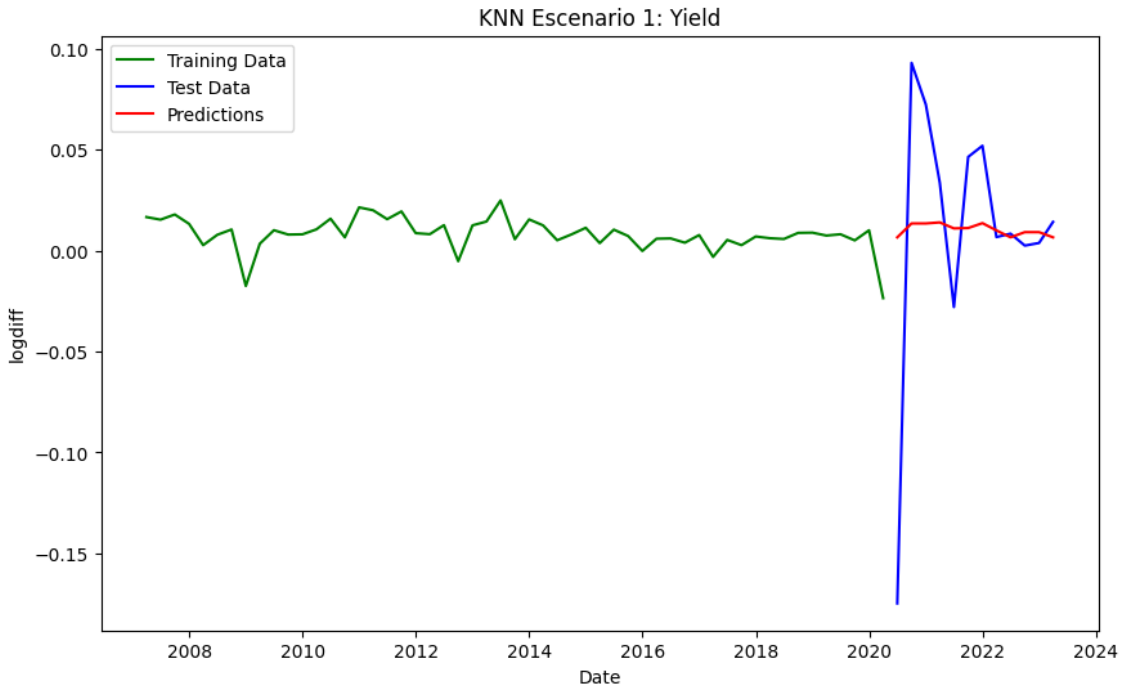
Por su lado para el caso de los modelos univariados se observa como el mejor modelo es el KNN el cual obtiene un error cuadrático medio de 0.0027, mientras los demás modelos siguen valores muy parecidos a los de los modelos multivariados.

En la gráfica 8 se observa como el modelo ARIMAX en el escenario 5 predijo los últimos 12 trimestres de la serie de tiempo al igual que el rango en el que se movían las predicciones.



Gráfica 10. Predicción del modelo ARIMAX en el escenario 5.

En la gráfica 11 se puede observar los resultados para el KNN multivariado con k=2 con los datos de entrenamiento en color verde, los de prueba en color azul y las predicciones en color rojo.



Gráfica 11. Predicción del modelo KNN multivariado en el escenario 1 con K=2.

Conclusiones generales del Proyecto.

En conclusión, el modelo KNN en su enfoque univariado se destaca como el más efectivo entre todos los modelos estudiados. Su rendimiento sobresaliente en la predicción de series de tiempo supera a otras técnicas analizadas en términos de precisión y capacidad predictiva. Esto resalta la importancia de considerar este tipo de modelos en el proceso de predicción, lo que conduce a resultados más precisos y confiables.

Es notable observar la consistencia de los resultados obtenidos en este estudio con los resultados encontrados en un documento de referencia con la diferencia en que el presente trabajo se basa en datos del Producto Interno Bruto (PIB) de Colombia, mientras que el documento de referencia se enfoca en el PIB de los Estados Unidos, los resultados muestran una coherencia que indica la aplicabilidad de las técnicas y enfoques utilizados en diferentes contextos como el de Colombia y permiten a los *"Policy Makers"* implementar diferentes técnicas con el objetivo de obtener siempre la mejor predicción del PIB real.

En cuanto a la métrica de evaluación, se utilizó el Error Cuadrático Medio (MSE), observándose como en el caso de los modelos univariados el menos MSE se obtiene con el KNN mientras que en el caso multivariado utilizando un modelo ARIMAX donde las variables independientes eran macroeconómicas y financiera. Esta métrica permite evaluar la precisión del modelo en la predicción de series de tiempo. El valor obtenido indica un nivel bajo de error, lo cual es altamente deseable en términos de calidad y confiabilidad de las predicciones realizadas.

En conjunto, estos hallazgos y conclusiones respaldan la relevancia y efectividad del enfoque KNN univariado en la predicción de series de tiempo, así como la consistencia de los resultados obtenidos en diferentes contextos. Los resultados del MSE refuerzan la confiabilidad y precisión de las predicciones realizadas. Estos hallazgos contribuyen al avance del conocimiento en el campo del análisis de series de tiempo y ofrecen perspectivas valiosas para futuras investigaciones y aplicaciones prácticas en diversos ámbitos.

Bibliografía.

Arunraj, N.S., Ahrens, D., Fernandes, M. (2016) Application of SARIMAX Model to Forecast Daily Sales in Food Retail Industry. Int. J. Operat. Res. Inf. Syst. 7, 2 (April 2016), 1–21. https://doi.org/10.4018/IJORIS.2016040101

Ban, T., Zhang, R., Pang, S., Sarrafzadeh, A., Inoue, D. (2013). Referential KNN Regression for Financial Time Series Forecasting. In: Lee, M., Hirose, A., Hou, ZG., Kil, R.M. (eds) Neural Information Processing. ICONIP 2013. Lecture Notes in Computer Science, vol 8226. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-42054-2_75

Chávez Quisbert, Nicolás. (1997). MODELOS ARIMA. Revista Ciencia y Cultura, (1), 23-30. Recuperado en 29 de mayo de 2023, de http://www.scielo.org.bo/scielo.php?script=sci_arttext&pid=S2077-33231997000100005&lng=es&tlng=es.

Fattah, J., Ezzine, L., Aman, Z., Haj El Moussami, and Abdeslam Lachhab. “Forecasting of demand using ARIMA model,” International Journal of Engineering Business Management, October 2018

Harvey, A.C. (1990). ARIMA Models. In: Eatwell, J., Milgate, M., Newman, P. (eds) Time Series and Statistics. The New Palgrave. Palgrave Macmillan, London. https://doi.org/10.1007/978-1-349-20865-4_2

Jain. D (2020) Cross Validation using KNN. Recuperado de: https://towardsdatascience.com/cross-validation-using-knn-6babb6e619c8

Kwon, K. Cho, W. and Na, J. (2016) ARIMAX and ARX Models with Social Media Information to Predict Unemployment Rate. Journal of Advanced Management Science, Vol. 4, No. 5, pp. 401-404 doi: 10.12720/joams.4.5.401-404

Pereira, J.M., Basto, M. & Ferreira-da-Silva, A. (2016). The Logistic Lasso and Ridge Regression in Predicting Corporate Failure. Procedia Economics and Finance. 39. 634-641. 10.1016/S2212-5671(16)30310-0.

Sharma, P. (2023). Ridge and Lasso Regression Explained. Recuperado de: https://www.tutorialspoint.com/ridge-and-lasso-regression-explained

Tajmouati, S., Wahbi, B., Bedoui, A., Abarda, A., Dakkon M. (2021) Applying k-nearest neighbors to time series forecasting: two new approaches.

Wahyudi, S.T. (2017). The ARIMA Model for the Indonesia Stock Price.