

UNIVERSIDAD COMPLUTENSE DE MADRID
MÁSTER BIG DATA, DATA SCIENCE & IA

Módulo

Predicción de Series Temporales

Estudiante

Allan Villarreal Chacón

Índice

Introducción.....	3
Representación Gráfica y Descomposición.....	4
Partición en Train y Test.....	5
Modelo de Suavizado Exponencial.....	6
Autocorrelación y Autocorrelación Parcial.....	8
Predicciones e Intervalos de Confianza.....	10
Predicciones con Distintos Modelos.....	11
Conclusiones Generales.....	12

Introducción

Para la realización del presente trabajo se ha seleccionado una serie temporal extraída del Instituto Nacional de Estadística (INE) que contiene el Índice de Comercio al por Menor sin Estaciones de Servicio.

Esta serie recoge la evolución mensual de la actividad del comercio minorista en España, excluyendo la venta de combustible, con lo cual se pretende enfocar el estudio en el comportamiento del consumo de bienes de carácter no energético.

La serie contiene 311 observaciones mensuales, lo que permite analizar su comportamiento, identificar tendencias, observar estacionalidad y detectar posibles ciclos económicos.

Desde una perspectiva de negocio, el estudio de la serie temporal propuesta, permite a las empresas del sector minorista anticipar cambios en la demanda y planificar sus inventarios así como ajustar sus estrategias comerciales.

El análisis predictivo de esta serie puede servir como herramienta de apoyo a la toma de decisiones estratégicas enfocadas en reducir la incertidumbre en un entorno económico cambiante.

Entre los objetivos del presente análisis se encuentran los siguientes:

- Identificar la presencia de tendencia y estacionalidad.
- Construir un modelo que permita generar predicciones útiles.
- Interpretar los resultados desde un enfoque orientado a la toma de decisiones empresariales.

El código Python completo, utilizado en el presente trabajo, se incluye en el cuaderno de Jupyter que se adjunta; aún así, dentro del presente documento se incluyen algunos fragmentos del código que se consideraron relevantes para mejorar la comprensión de los pasos llevados a cabo durante el análisis.

Representación Gráfica y Descomposición

Para proceder con la representación gráfica y descomposición de la serie temporal elegida, se procedió primeramente con la conversión de los datos contenidos en la columna Período a Datetime, así como con la conversión de los valores contenidos en la columna Total a tipo Float, tal como se muestra en el siguiente código:

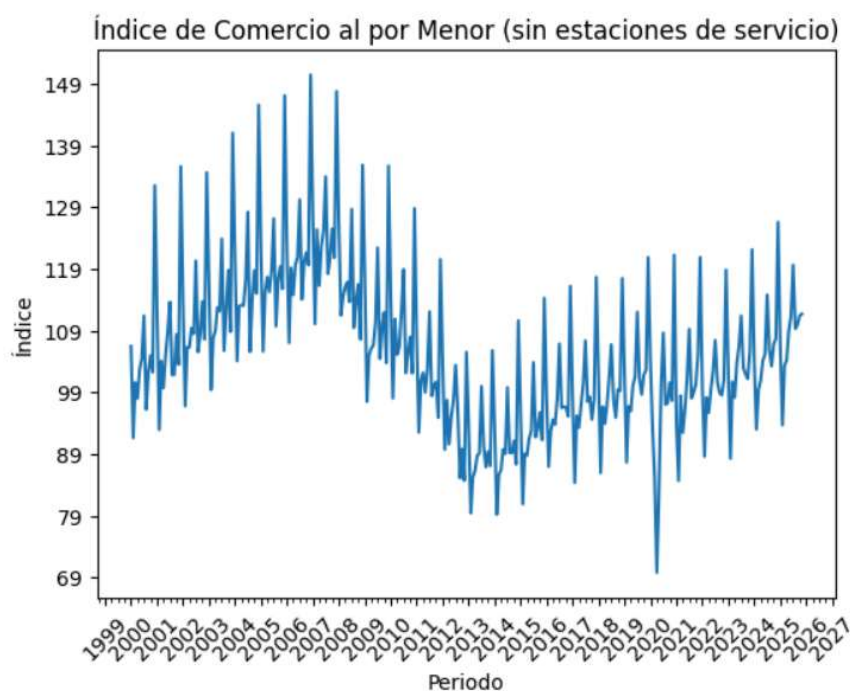
```
#Convertimos la columna Período a Datetime
df["Período"] = pd.to_datetime(df["Período"], format='%Ym%m')

#Ordenamos los valores por fecha
df = df.sort_values('Período')

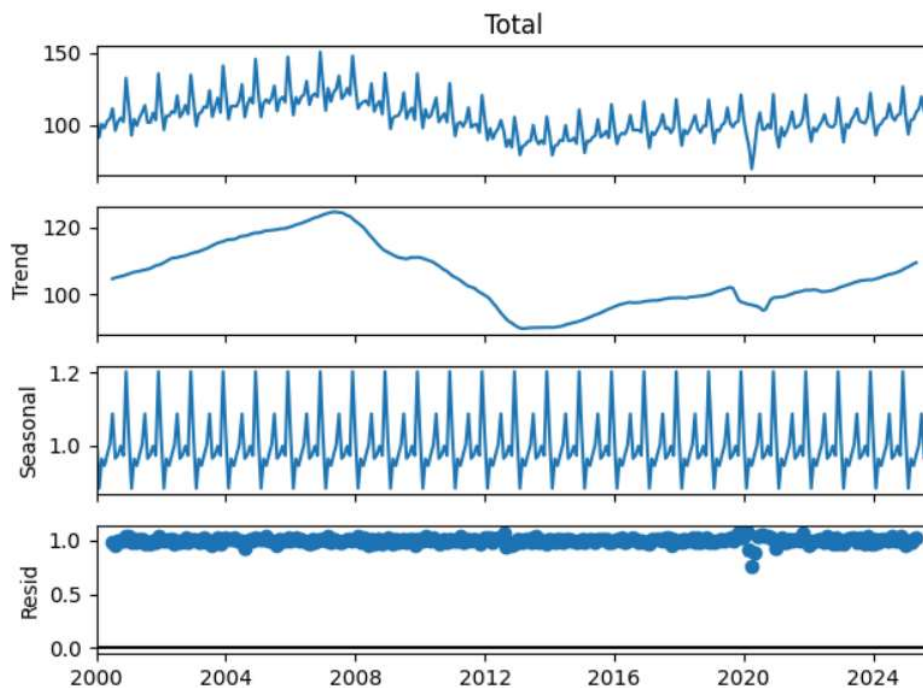
#Establecemos la fecha como índice
df.set_index('Período', inplace=True)

#Convertimos los valores del índice a número
df["Total"] = df["Total"].str.replace(",", ".").astype(float)
```

Tras realizar las conversiones necesarias, se obtiene la siguiente representación gráfica de la serie temporal:



Posteriormente se procedió con la descomposición de la serie temporal, cuyo resultado se muestra a continuación:



Tanto la representación gráfica como la descomposición de la serie temporal muestran el comportamiento estacional de los datos, así como las caídas ocasionadas por eventos como la crisis económica que se presentó durante la primera década de este siglo, o la pandemia de COVID-19 en 2020.

Partición en Train y Test

En concordancia con el enunciado se reservan los últimos 10 datos observados para realizar predicciones posteriores y probar los distintos modelos que se tendrán en cuenta para el presente estudio.

La separación de dichos datos se llevó a cabo según se muestra en el siguiente fragmento de código:

```
n_test = 10
train = df[:-n_test]
test = df[-n_test:]
```

Modelo de Suavizado Exponencial

Según se solicita en el enunciado propuesto, se pusieron a prueba diferentes modelos de suavizado exponencial, con el fin de comparar los resultados y elegir el modelo que mejores características ofrezca en el contexto del presente proyecto.

Se pusieron a prueba los modelos SES, Holt, Holt-Winters Aditivo y Holt-Winters Multiplicativo sobre el conjunto de datos de train, según se detalla en el siguiente fragmento de código:

```
# --- Utilizando SES ---
model_ses = SimpleExpSmoothing(train['Total']).fit()

# --- Utilizando Holt ---
model_holt = ExponentialSmoothing(
    train['Total'],
    trend='add',
    seasonal=None
).fit()

# --- Utilizando Holt-Winters aditivo---
model_hw_add = ExponentialSmoothing(
    train['Total'],
    trend='add',
    seasonal='add',
    seasonal_periods=12
).fit()

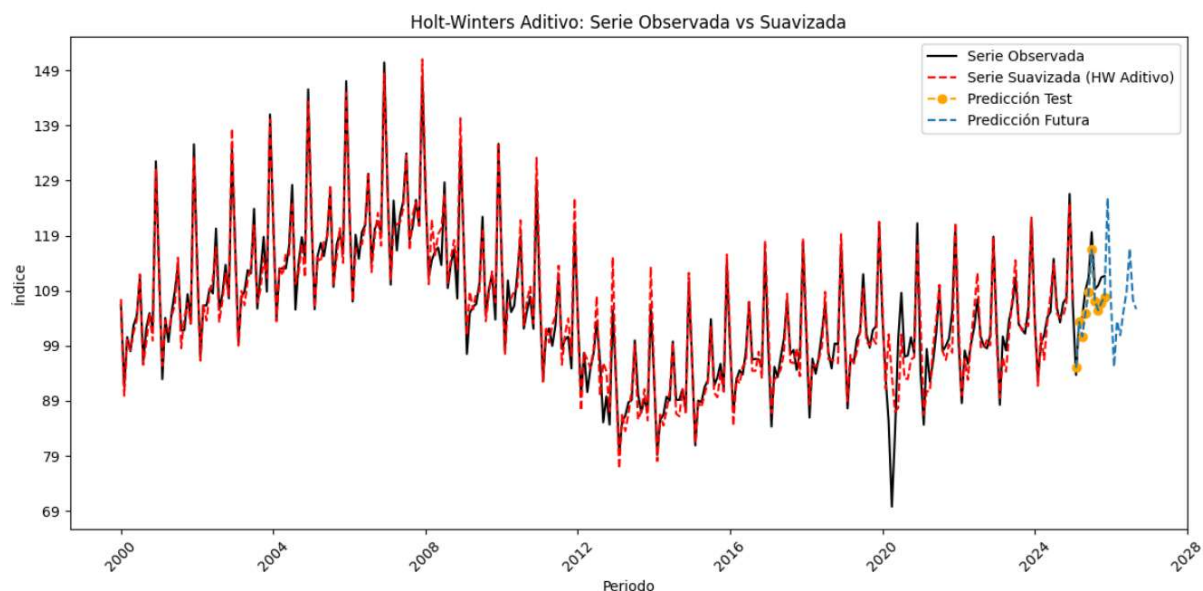
# --- Utilizando Holt-Winters multiplicativo ---
model_hw_mul = ExponentialSmoothing(
    train['Total'],
    trend='add',
    seasonal='mul',
    seasonal_periods=12
).fit()
```

Los modelos fueron posteriormente evaluados utilizando MAE y RMSE, tras lo cual se obtuvieron los siguientes resultados:

	Modelo	MAE	RMSE
0	SES	4.948856	6.476559
1	Holt	4.216349	6.236822
2	HW Aditivo	2.967027	3.273195
3	HW Multiplicativo	3.245457	3.719490

Tal como puede observarse en la tabla anterior el modelo Holt-Winters Aditivo es el que muestra los valores de MAE y RMSE más favorables por lo que se elige dicho modelo como el idóneo para el estudio de la serie temporal elegida.

A continuación se adjunta la representación gráfica de la serie observada y la serie suavizada con una predicción a 10 períodos



Tal como puede observarse, el modelo seleccionado reproduce adecuadamente tanto la tendencia como la estacionalidad de la serie temporal sobre los datos observados; pero además, las predicciones muestran un comportamiento estacional similar a los datos observados lo que refuerza la elección del modelo HW Aditivo como el apropiado para el estudio.

El modelo aditivo seleccionado presenta los siguientes parámetros:

$$\alpha = 0.38974$$

$$\beta = 0.00000$$

$$\gamma = 0.23001$$

Sustituyendo dichos valores en las ecuaciones teóricas del modelo, obtenemos las siguientes expresiones:

$$L_t = 0.38974(Y_t - S_{t-12}) + 0.61026(L_{t-1} + T_{t-1})$$

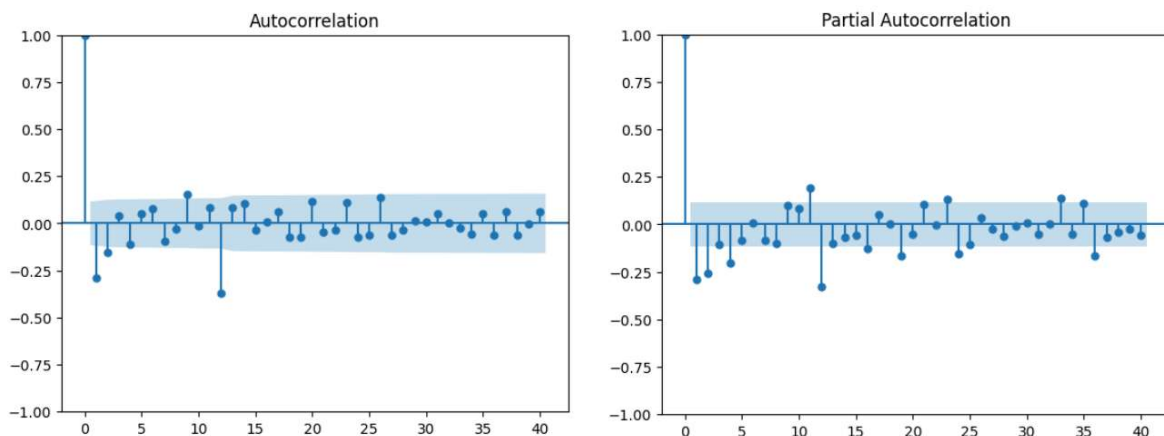
$$T_t = T_{t-1} = 0.01604$$

$$S_t = 0.23001(Y_t - L_t) + 0.76999S_{t-12}$$

Dado que $\beta = 0$, se determina que la mejor estrategia de cara a la interpretación de los datos es mantener la pendiente inicial estimada, ya que no existe razón estadística suficiente para justificar la actualización de la tendencia en cada periodo.

Autocorrelación y Autocorrelación Parcial

Se procedió con la representación gráfica de la serie y las funciones de Autocorrelación y Autocorrelación Parcial, tal como se solicita en el enunciado; dichas representaciones gráficas se presentan a continuación:



A partir de las gráficas anteriores se determina un modelo inicial definido por la expresión SARIMA(0,1,1)(0,1,1)_{12}; sin embargo antes de considerarlo definitivo se llevan a cabo algunas comprobaciones en concordancia con lo solicitado en el enunciado propuesto.

En primer lugar se obtiene el sumario de parámetros del modelo utilizando el siguiente código:

```
model_sarima = SARIMAX(
    train['Total'],
    order=(0,1,1),
    seasonal_order=(0,1,1,12),
    enforce_stationarity=False,
    enforce_invertibility=False
)
results_sarima = model_sarima.fit()
print(results_sarima.summary())
```

Cuya ejecución devuelve el siguiente resultado:

```

SARIMAX Results
=====
Dep. Variable:                Total    No. Observations:                301
Model:                SARIMAX(0, 1, 1)x(0, 1, 1, 12)    Log Likelihood                -711.925
Date:                Thu, 19 Feb 2026    AIC                1429.850
Time:                08:54:27    BIC                1440.689
Sample:                01-01-2000    HQIC                1434.200
                - 01-01-2025
Covariance Type:                opg
=====
              coef    std err          z      P>|z|      [0.025     0.975]
-----
ma.L1          -0.5461      0.025    -21.480      0.000     -0.596     -0.496
ma.S.L12       -0.6980      0.042    -16.469      0.000     -0.781     -0.615
sigma2         10.4568      0.380     27.541      0.000      9.713     11.201
=====
Ljung-Box (L1) (Q):                1.95    Jarque-Bera (JB):                1498.11
Prob(Q):                0.16    Prob(JB):                0.00
Heteroskedasticity (H):                2.81    Skew:                -1.60
Prob(H) (two-sided):                0.00    Kurtosis:                14.00
=====

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Tal como puede observarse, el MA1 exhibe un coeficiente de -0.5461 y un p_value de 0.000, mientras el MA estacional (lag12) muestra un coeficiente de -0.6980 y un p_value de 0.000, tras lo cual se confirma que ambos son significativos y además se acepta como verdadera la estacionalidad anual y relevante de la serie.

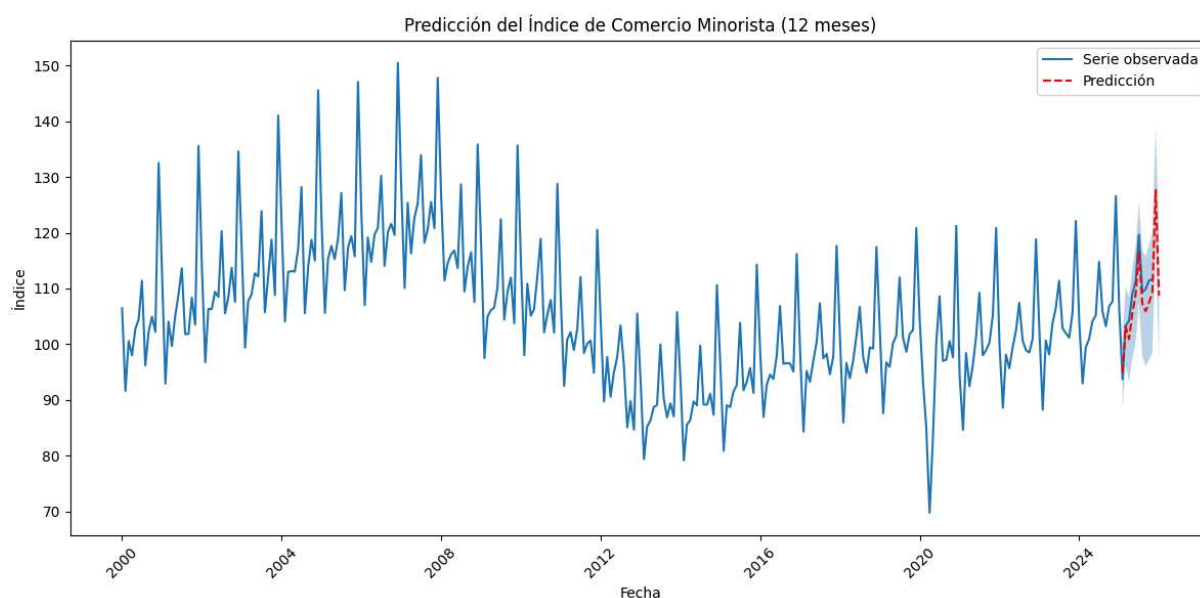
Adicionalmente se determina que el modelo es parsimonioso, lo cual, en conjunto con su significancia estadística confirma la idoneidad del modelo escogido, que puede expresarse mediante la siguiente expresión algebraica:

$$(1-B)(1-B12)y_t=(1-\theta_1B)(1-\Theta_1B12)\epsilon_t$$

Predicciones e Intervalos de Confianza

Para la elaboración de estas predicciones se escoge como intervalo mínimo, un ciclo de 12 meses; esto porque, en el contexto de la serie temporal seleccionada, 1 año, es el periodo mínimo con sentido completo para el estudio.

El código utilizado para esta sección se incluye en el cuaderno de Jupyter que se adjunta en la entrega. La gráfica obtenida tras la ejecución de dicho código se muestra a continuación:

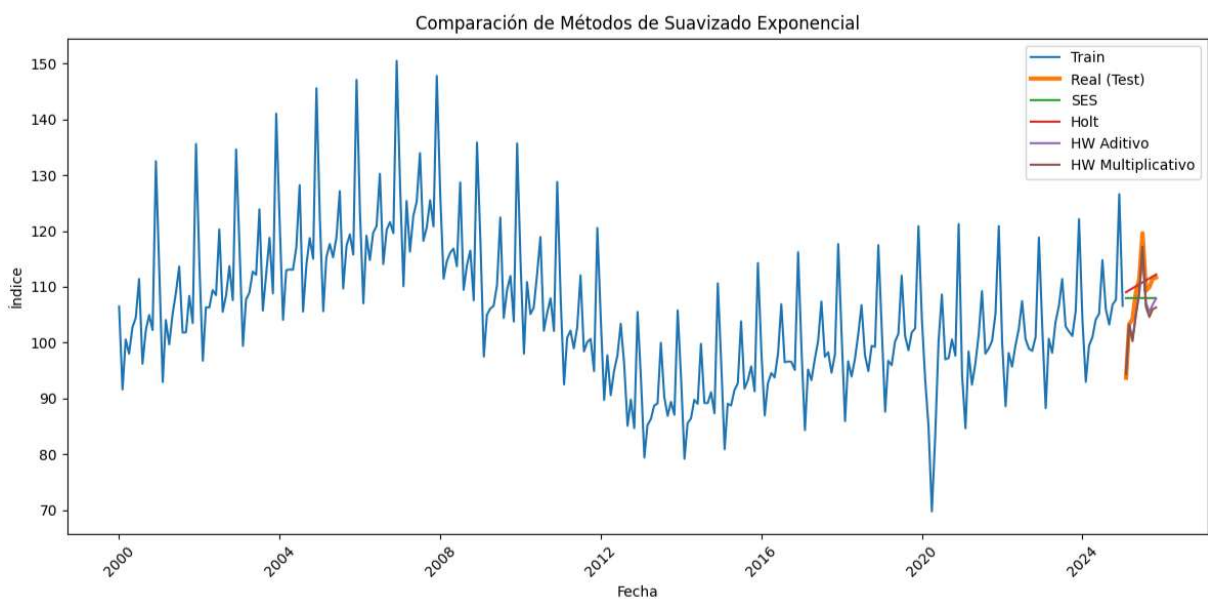


En la gráfica anterior, la línea punteada representa la predicción, mientras la banda celeste sombreada representa un intervalo de confianza del 95%. Se puede observar como el

modelo mantiene la estacionalidad anual esperada, pero también es notorio cómo la incertidumbre aumenta conforme aumenta el horizonte temporal.

Predicciones con Distintos Modelos

Se utilizaron los modelos probados anteriormente en el presente documento para realizar predicciones sobre la serie temporal, con el fin de comparar los resultados obtenidos; la representación gráfica correspondiente se muestra a continuación:



Además, en atención al enunciado propuesto, se llevó a cabo una comparación numérica entre los modelos, utilizando el MAE y el RMSE como referencia, la tabla siguiente muestra los resultados:

	Modelo	MAE	RMSE
0	SES	4.938509	6.474799
1	Holt	4.216349	6.236822
2	HW Aditivo	2.967027	3.273195
3	HW Multiplicativo	3.245457	3.719490

Tal como se discutió en un punto anterior, el modelo con los valores de error más bajos es el modelo que resulta idóneo para el análisis de la serie temporal elegida, en este caso, el modelo Holt-Winters Aditivo.

La idoneidad del modelo anteriormente mencionado se corrobora, además por medio del estudio de la gráfica adjunta, dónde se nota cómo el HW-Aditivo presenta el menor error de predicción en el conjunto de validación.

Conclusiones Generales

- El análisis exploratorio y la descomposición de la serie temporal mostraron una clara tendencia creciente a largo plazo con perturbaciones originadas por eventos extraordinarios como la crisis económica de principios de siglo y la pandemia de COVID-19.
- Se evaluaron 4 modelos de suavizado exponencial (SES, Holt, HW- Aditivo, HW-Multiplicativo) y se demostró que el modelo aditivo es el que mejor se ajusta a la estacionalidad de la serie elegida para el presente estudio.
- Se identificó el modelo SARIMA(0,1,1)(0,1,1)₁₂, que presentó parámetros estadísticamente significativos, además de una capacidad predictiva coherente con el comportamiento histórico de los datos.
- Las predicciones para los siguientes 12 meses mantuvieron el patrón estacional y presentaron un incremento progresivo de la incertidumbre según aumenta el horizonte temporal.