

UNIVERSIDAD AUTÓNOMA DE QUERÉTARO

FACULTAD DE INFORMÁTICA

Inteligencia Artificial 2024-2

Proyecto Final

Profesor: Martin Muñoz Mandujano



UNIVERSIDAD AUTÓNOMA DE QUERÉTARO

Por: Jesús Gabriel Gudiño Lara

Expediente: 325675

Ana Rosa Ramirez Lopez

Expediente: 325672

Ingeniería en Ciencia y Analítica de Datos

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dense, Dropout, Input
from tensorflow.keras.callbacks import EarlyStopping

from sklearn.preprocessing import StandardScaler
from sklearn.metrics import mean_squared_error
from math import sqrt
```

```
FORECAST_HORIZON = 7 # Define forecasting horizon (days)
```

```
df = pd.read_excel('1_Daily_minimum_temps.xlsx')
```

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3650 entries, 0 to 3649
Data columns (total 1 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   Date,Temp    3650 non-null   object 
dtypes: object(1)
memory usage: 28.6+ KB
```

```
df.columns.tolist()
```

```
['Date,Temp']
```

```
df.describe().T
```

	count	unique	top	freq	grid
Date,Temp	3650	3650	12/31/90,13	1	

```
df.sample(10)
```

	Date,Temp	grid
1628	06/18/85,7.3	grid
175	06/25/81,6.2	grid
339	12/06/81,15	grid
972	08/31/83,10	grid
2783	08/16/88,12.4	grid
1675	08/04/85,6.9	grid
582	08/06/82,4.5	grid
3505	08/09/90,6.5	grid
994	09/22/83,8.5	grid
1612	06/02/85,5.5	grid

Analisis principal del estado del dataset

- El dataset contiene 3650 filas y una columna. Pero al parecer hay una segunda columna implícita, Temp.
- No hay valores nulos.
- La columna Date es de tipo object, por lo que se debe convertir a datetime.
- La columna Temp esta dentro de la columna Date, por lo que se debe separar.

```
# Separando las columnas Date y Temp que estan mezcladas en una sola columna separamos por una coma necesitamos hacer un split
df[['Date', 'Temp']] = df['Date,Temp'].str.split(',', expand=True) # Separamos la columna Date en dos columnas Date y Temp al mismo tiempo
```

```
# Borrar la columna original que contenía ambas columnas mezcladas
df = df.drop(columns=['Date,Temp'])
```

```
df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3650 entries, 0 to 3649
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype  
--- 
 0   Date    3650 non-null    object 
 1   Temp    3650 non-null    object 
dtypes: object(2)
memory usage: 57.2+ KB
```

```
df.sample(12)
```

	Date	Temp	grid
1497	02/07/85	14.8	grid
600	08/24/82	12.1	grid
454	03/31/82	10.2	grid
3283	12/30/89	14.4	grid
1383	10/15/84	5.9	grid
609	09/02/82	10	grid
877	05/28/83	12.1	grid
2480	10/18/87	10.1	grid
153	06/03/81	9.8	grid
1402	11/03/84	12.2	grid
1079	12/16/83	15	grid
3413	05/09/90	6.9	grid

Logramos la separación exitosa de las columnas Date y Temp

EDA: Análisis exploratorio de datos y visualización

Después dejar el dataset apto para su uso, se procede a realizar un análisis exploratorio de datos y visualización de los mismos.

```
# Convertir la columna Date a tipo datetime. Actualmente la columna Date es de tipo object con formato MM/DD/YY
df['Date'] = pd.to_datetime(df['Date'], format='%m/%d/%y')
```

```
# Convertir Temp a numérico
df['Temp'] = pd.to_numeric(df['Temp'], errors='coerce')
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3650 entries, 0 to 3649
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype  
 --- 
  0   Date    3650 non-null   datetime64[ns]
  1   Temp    3647 non-null   float64 
dtypes: datetime64[ns](1), float64(1)
memory usage: 57.2 KB
```

```
df.sample(12)
```

	Date	Temp	grid
3357	1990-03-14	11.0	bar
992	1983-09-20	13.6	bar
329	1981-11-26	9.9	bar
2983	1989-03-05	20.3	bar
1932	1986-04-18	12.0	bar
3148	1989-08-17	6.6	bar
2452	1987-09-20	7.8	bar
1198	1984-04-13	11.1	bar
2858	1988-10-30	14.3	bar
3332	1990-02-17	18.6	bar
533	1982-06-18	8.0	bar
3596	1990-11-08	15.4	bar

▼ Paso 1: Estadísticas descriptivas básicas

```
df.describe().T
```

	count	mean	min	25%	50%	75%	max	std
Date	3650	1985-12-31 07:12:00	1981-01-01 00:00:00	1983-07-02 06:00:00	1985-12-31 12:00:00	1988-06-30 18:00:00	1990-12-31 00:00:00	NaN
-	--	--	--	--	--	--	--	--

```
print(f"Fechas: desde {df['Date'].min()} hasta {df['Date'].max()}")
print(f"Rango de temperatura: {df['Temp'].min():.2f}°C a {df['Temp'].max():.2f}°C")
print(f"Total de registros: {len(df)}")
```

```
Fechas: desde 1981-01-01 00:00:00 hasta 1990-12-31 00:00:00
Rango de temperatura: 0.00°C a 26.30°C
Total de registros: 3650
```

```
# Verificando los valores de la temperatura para saber si son consistentes y ver la cantidad de valores en el dataset
print("\nValores únicos de temperatura:")
print(df['Temp'].value_counts().sort_index())
```

```
Valores únicos de temperatura:
```

```
Temp
0.0    2
0.3    2
0.5    1
0.6    1
0.7    2
...
24.3   1
24.8   1
25.0   2
25.2   1
26.3   1
Name: count, Length: 227, dtype: int64
```

```
# Numero de registros por año
df['Year'] = df['Date'].dt.year
registros_por_año = df['Year'].value_counts().sort_index()
print("Registros por año:")
print(registros_por_año)
```

```
Registros por año:
```

```
Year
1981    365
```

```
1982    365
1983    365
1984    365
1985    365
1986    365
1987    365
1988    365
1989    365
1990    365
Name: count, dtype: int64
```

```
# Borrando el registro del año 1981 ya que solo tiene un dato y agrega ruido al análisis por año
# df = df[df['Year'] != 1981] # Kept 1981 data as it contains 365 records
```

```
# Verificar valores faltantes
print(f"\nValores faltantes:\n{df.isnull().sum()}"
```

```
Valores faltantes:
Date    0
Temp    3
Year    0
dtype: int64
```

```
# Calcular el promedio de temperatura por mes
df['Month'] = df['Date'].dt.month
promedio_por_mes = df.groupby('Month')['Temp'].transform('mean')
```

```
# Rellenar valores faltantes con el promedio del mes correspondiente
df['Temp'] = df['Temp'].fillna(promedio_por_mes)
```

```
# Verificar que no hay valores faltantes
print(f"Valores faltantes después de llenar: {df['Temp'].isnull().sum()}"
```

```
Valores faltantes después de llenar: 0
```

▼ Paso 2: Visualizar la serie de tiempo

```
import matplotlib.pyplot as plt
import seaborn as sns

# Configurar estilos globales
sns.set_style("whitegrid")
sns.set_palette("husl")
plt.rcParams['figure.facecolor'] = '#f8f9fa'
plt.rcParams['font.size'] = 10
plt.rcParams['axes.labelsize'] = 11
plt.rcParams['axes.titlesize'] = 13
plt.rcParams['xtick.labelsize'] = 9
plt.rcParams['ytick.labelsize'] = 9
```

▼ Paso 2: Visualizar la serie de tiempo

```
plt.figure(figsize=(16, 5))
plt.plot(df['Date'], df['Temp'], linewidth=1, color="#2E86AB", alpha=0.85)
plt.title('Serie de Tiempo: Temperaturas Mínimas Diarias', fontsize=14, fontweight='bold', pad=15)
plt.xlabel('Fecha', fontsize=11, fontweight='bold')
plt.ylabel('Temperatura (°C)', fontsize=11, fontweight='bold')
plt.grid(True, alpha=0.2, linestyle='--')
plt.tight_layout()
plt.show()
```

Serie de Tiempo: Temperaturas Mínimas Diarias

A partir de la serie de tiempo completa se observan varios patrones importantes:

- **Estacionalidad anual marcada:** las temperaturas mínimas suben y bajan de forma cíclica cada año, con inviernos fríos y veranos más templados, lo que sugiere una fuerte componente estacional.
- **Absencia de una tendencia clara de largo plazo:** visualmente no se aprecia un calentamiento o enfriamiento sostenido a lo largo de los años; la media parece relativamente estable.
- **Alta variabilidad diaria:** aun dentro de cada ciclo anual hay fluctuaciones importantes día a día, lo que indica que el modelo deberá capturar tanto la estacionalidad como la variación de corto plazo.

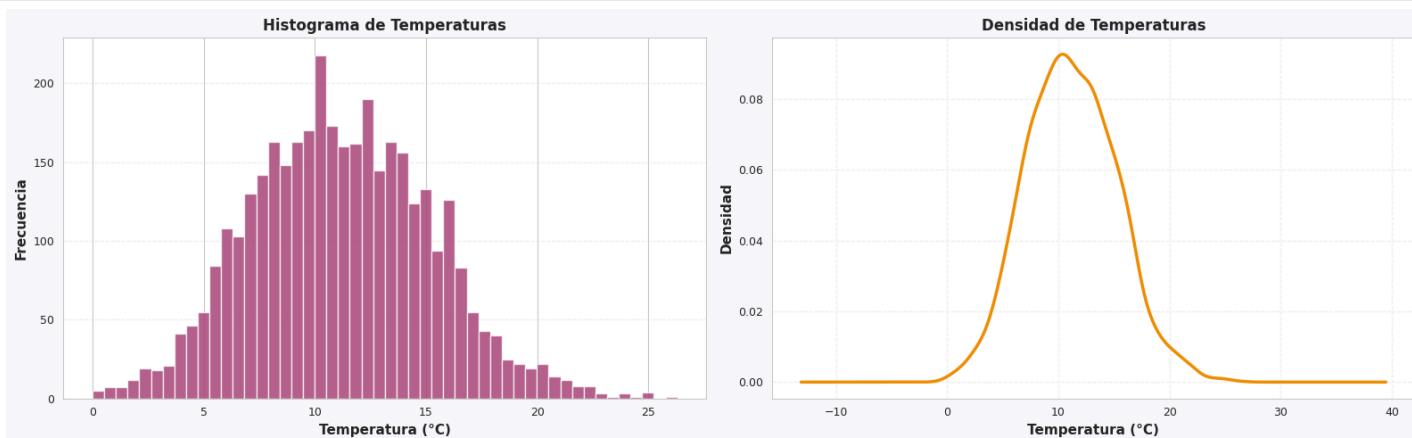
✓ Paso 3: Distribución de temperaturas

```
fig, axes = plt.subplots(1, 2, figsize=(16, 5))

# Histograma
axes[0].hist(df['Temp'], bins=50, color='#A23B72', edgecolor='white', alpha=0.8, linewidth=1)
axes[0].set_title('Histograma de Temperaturas', fontsize=12, fontweight='bold')
axes[0].set_xlabel('Temperatura (°C)', fontweight='bold')
axes[0].set_ylabel('Frecuencia', fontweight='bold')
axes[0].grid(True, alpha=0.2, axis='y', linestyle='--')

# KDE (densidad)
df['Temp'].plot(kind='density', ax=axes[1], color="#F18F01", linewidth=2.5)
axes[1].set_title('Densidad de Temperaturas', fontsize=12, fontweight='bold')
axes[1].set_xlabel('Temperatura (°C)', fontweight='bold')
axes[1].set_ylabel('Densidad', fontweight='bold')
axes[1].grid(True, alpha=0.2, linestyle='--')

plt.tight_layout()
plt.show()
```



El histograma y la curva de densidad muestran que las temperaturas mínimas diarias se concentran principalmente entre 5 °C y 15 °C, con un pico alrededor de 10-12 °C. La distribución es **unimodal** y aproximadamente simétrica, con pocas observaciones en los extremos fríos (cerca de 0 °C) y cálidos (por encima de 20 °C). Estos valores extremos parecen razonables desde el punto de vista climático, por lo que se consideran parte natural de la variabilidad de la serie y no se tratan como outliers a eliminar.

✓ Paso 4: Análisis por año

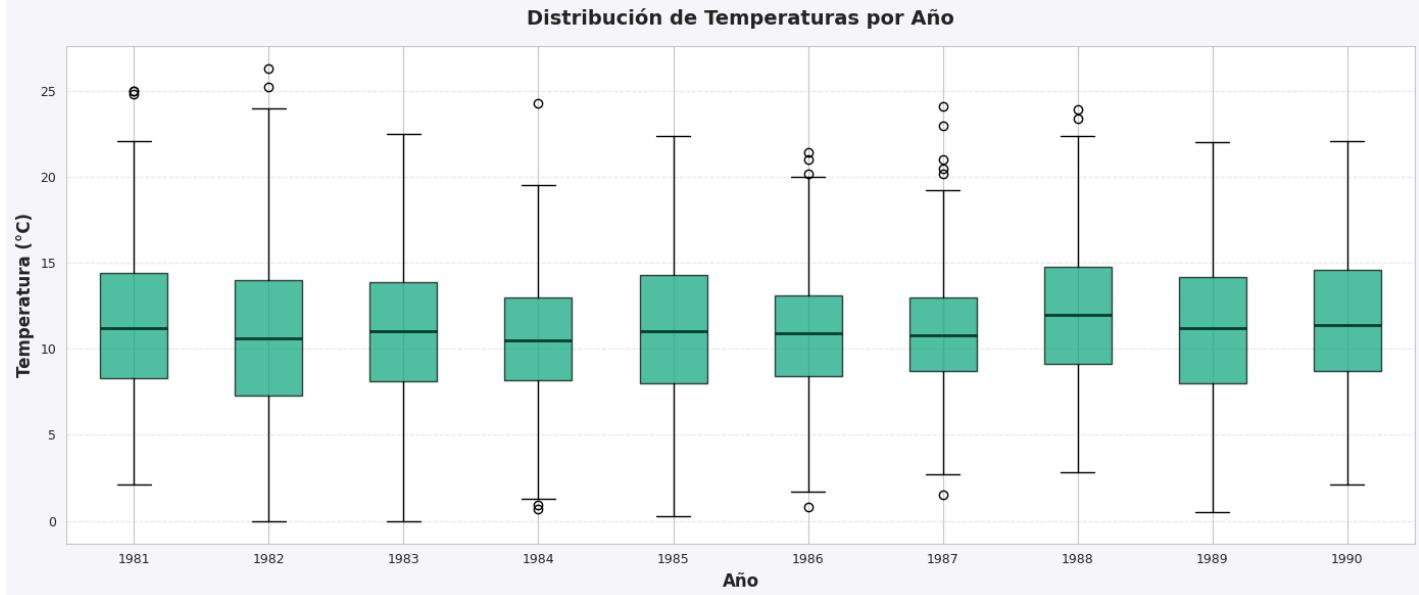
```
# Extraer componentes temporales esenciales
df['Year'] = df['Date'].dt.year
df['Month'] = df['Date'].dt.month

print("Columnas temporales creadas:")
print(df[['Date', 'Year', 'Month']].sample(12))
```

```
Columnas temporales creadas:
  Date Year Month
1192 1984-04-07 1984      4
2793 1988-08-26 1988      8
17   1981-01-18 1981      1
1850 1986-01-26 1986      1
1667 1985-07-27 1985      7
2529 1987-12-06 1987     12
1405 1984-11-06 1984     11
1827 1986-01-03 1986      1
3    1981-01-04 1981      1
1521 1985-03-03 1985      3
3215 1989-10-23 1989     10
2876 1988-11-17 1988     11
```

```
# Boxplot por año con estilo mejorado
fig, ax = plt.subplots(figsize=(14, 6))
bp = ax.boxplot([df[df['Year']==year]['Temp'].values for year in sorted(df['Year'].unique())],
                labels=sorted(df['Year'].unique()), patch_artist=True)
for patch in bp['boxes']:
    patch.set_facecolor('#06A77D')
    patch.set_alpha(0.7)
for median in bp['medians']:
    median.set_color('#003d2e')
    median.set_linewidth(2)
ax.set_title('Distribución de Temperaturas por Año', fontsize=14, fontweight='bold', pad=15)
ax.set_xlabel('Año', fontsize=12, fontweight='bold')
ax.set_ylabel('Temperatura (°C)', fontsize=12, fontweight='bold')
ax.grid(True, alpha=0.3, axis='y', linestyle='--')
plt.tight_layout()
plt.show()
```

```
/tmp/ipython-input-294305123.py:3: MatplotlibDeprecationWarning: The 'labels' parameter of boxplot() has been renamed 'tick_labels'
  bp = ax.boxplot([df[df['Year']==year]['Temp'].values for year in sorted(df['Year'].unique())]),
```



```
# Estadísticas por año
print("Estadísticas de temperatura por año:")
print(df.groupby('Year')['Temp'].describe())
```

Year	count	mean	std	min	Quantiles				max
					25%	50%	75%		
1981	365.0	11.517260	4.337430	2.1	8.3	11.2	14.4	25.0	
1982	365.0	10.817832	4.755245	0.0	7.3	10.6	14.0	26.3	
1983	365.0	11.187397	4.187325	0.0	8.1	11.0	13.9	22.5	
1984	365.0	10.610012	3.698657	0.7	8.2	10.5	13.0	24.3	
1985	365.0	11.137534	4.138998	0.3	8.0	11.0	14.3	22.4	
1986	365.0	10.803288	3.610910	0.8	8.4	10.9	13.1	21.4	
1987	365.0	10.853151	3.626849	1.5	8.7	10.8	13.0	24.1	
1988	365.0	11.972055	3.748566	2.8	9.1	12.0	14.8	23.9	
1989	365.0	11.261918	4.333539	0.5	8.0	11.2	14.2	22.0	
1990	365.0	11.669589	3.861600	2.1	8.7	11.4	14.6	22.1	

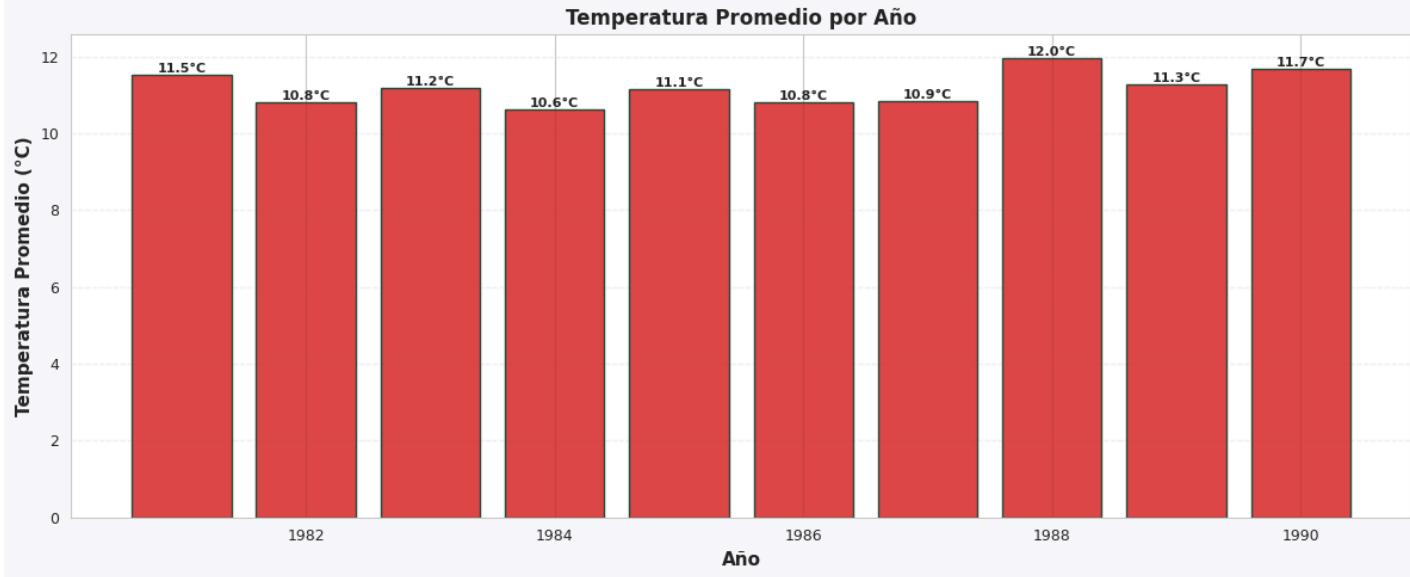
✓ Paso 5: Análisis por mes

```
# Promedio de temperatura por año
temp_por_año = df.groupby('Year')['Temp'].mean()
```

```

temp_por_año = df.groupby('Year')[['Temp']].mean()
bars = plt.bar(temp_por_año.index, temp_por_año.values, color='#D62828', edgecolor='#003d2e', linewidth=1, alpha=0.85)
# Agregar valores en las barras
for bar in bars:
    height = bar.get_height()
    plt.text(bar.get_x() + bar.get_width()/2., height,
              f'{height:.1f}°C',
              ha='center', va='bottom', fontweight='bold', fontsize=8)
plt.title('Temperatura Promedio por Año', fontsize=12, fontweight='bold')
plt.xlabel('Año', fontweight='bold')
plt.ylabel('Temperatura Promedio (°C)', fontweight='bold')
plt.grid(True, alpha=0.2, axis='y', linestyle='--')
plt.tight_layout()
plt.show()

```



Los diagramas por año confirman que la distribución de las temperaturas mínimas es muy similar entre 1982 y 1990.

Los rangos intercuartílicos y las medianas apenas cambian de un año a otro, lo que indica que **no hay una tendencia clara de calentamiento o enfriamiento** en el periodo analizado; más bien, se observa una variabilidad interanual moderada dentro de un mismo rango climático.

La gráfica de barras de temperatura promedio anual refuerza esta idea: los promedios se mantienen alrededor de 11–12 °C, con diferencias pequeñas entre años. Esto sugiere que el componente principal de la dinámica de la serie es **estacional (intra-anual)** más que una tendencia sistemática de largo plazo.

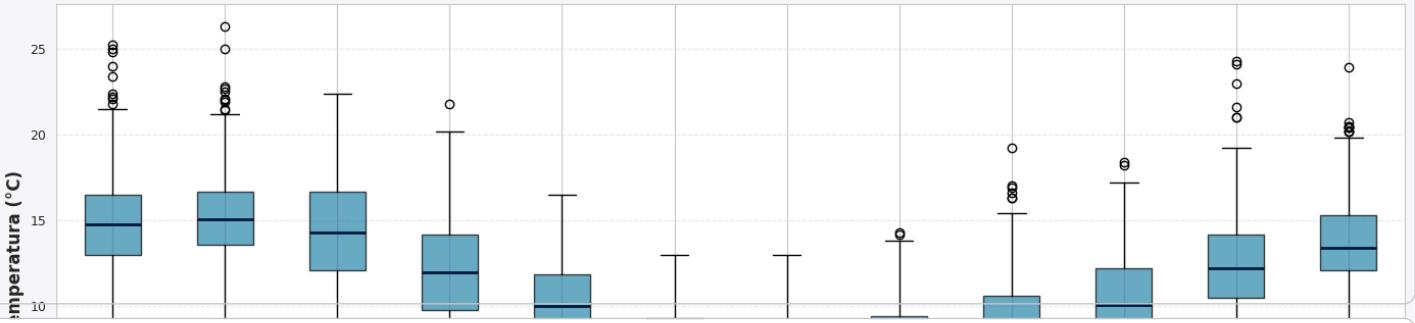
```

# Boxplot por mes con estilo mejorado
fig, ax = plt.subplots(figsize=(14, 6))
bp = ax.boxplot([df[df['Month']==month]['Temp'].values for month in sorted(df['Month'].unique())],
                 labels=sorted(df['Month'].unique()), patch_artist=True)
for patch in bp['boxes']:
    patch.set_facecolor('#2E86AB')
    patch.set_alpha(0.7)
for median in bp['medians']:
    median.set_color('#001d3d')
    median.set_linewidth(2)
ax.set_title('Distribución de Temperaturas por Mes', fontsize=14, fontweight='bold', pad=15)
ax.set_xlabel('Mes', fontsize=12, fontweight='bold')
ax.set_ylabel('Temperatura (°C)', fontsize=12, fontweight='bold')
ax.grid(True, alpha=0.3, axis='y', linestyle='--')
plt.tight_layout()
plt.show()

```

```
/tmp/ipython-input-3519225379.py:3: MatplotlibDeprecationWarning: The 'labels' parameter of boxplot() has been renamed 'tick_l  
bp = ax.boxplot([df[df['Month']==month]['Temp'].values for month in sorted(df['Month'].unique())]),
```

Distribución de Temperaturas por Mes



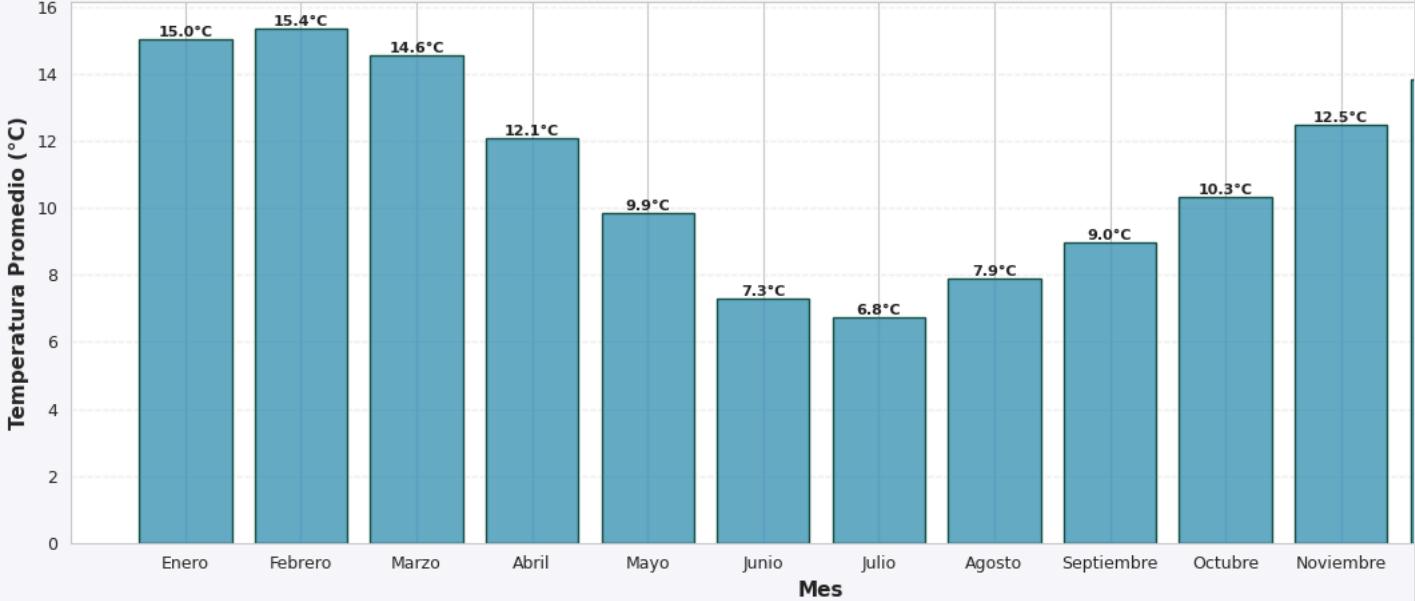
```
# Promedio de temperatura por mes
# Diccionario de meses para ayudar a la visualización
meses = {
    1: 'Enero', 2: 'Febrero', 3: 'Marzo', 4: 'Abril',
    5: 'Mayo', 6: 'Junio', 7: 'Julio', 8: 'Agosto',
    9: 'Septiembre', 10: 'Octubre', 11: 'Noviembre', 12: 'Diciembre'
}

# Promedio de temperatura por mes
temp_por_mes = df.groupby('Month')['Temp'].mean()

# Create labels for x-axis
month_labels = temp_por_mes.index.map(meses)

plt.figure(figsize=(12, 5))
bars = plt.bar(month_labels, temp_por_mes.values, color='#499BBC', edgecolor="#003d2e", linewidth=1, alpha=0.85)
# Agregar valores en las barras
for bar in bars:
    height = bar.get_height()
    plt.text(bar.get_x() + bar.get_width()/2., height,
              f'{height:.1f}°C',
              ha='center', va='bottom', fontweight='bold', fontsize=8)
plt.title('Temperatura Promedio por Mes', fontsize=12, fontweight='bold')
plt.xlabel('Mes', fontweight='bold')
plt.ylabel('Temperatura Promedio (°C)', fontweight='bold')
plt.xticks(rotation=0)
plt.grid(True, alpha=0.2, axis='y', linestyle='---')
plt.tight_layout()
plt.show()
```

Temperatura Promedio por Mes



El análisis por mes revela con claridad el **patrón estacional anual** de la serie.

Los boxplots muestran que las temperaturas mínimas promedio son más altas en **verano austral** (enero–marzo y nuevamente en noviembre–diciembre), mientras que alcanzan sus valores más bajos en **invierno** (junio y julio).

Buscando de internet se puede observar que en el hemisferio sur, el invierno ocurre en el periodo de junio, julio y agosto, mientras que el verano ocurre en el periodo de diciembre, enero y febrero. Por lo tanto, nuestro dataset refleja correctamente el patrón estacional anual del hemisferio sur lo que si corresponde con Australia.

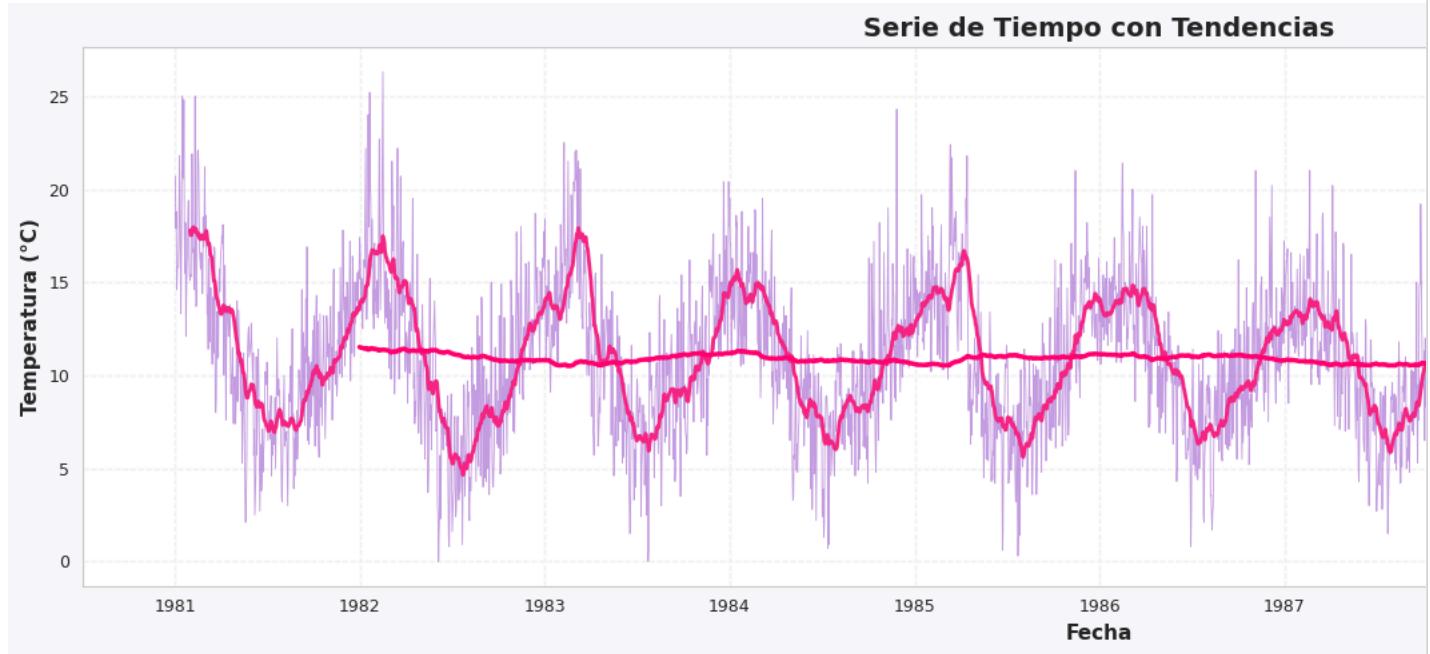
La gráfica de promedios mensuales sigue una forma casi sinusoidal: desciende desde marzo hasta julio y luego vuelve a aumentar gradualmente. Esta estructura estacional tan marcada sugiere que variables derivadas del calendario (mes o día del año) pueden aportar información relevante para el modelo de pronóstico.

✓ Paso 6: Detección de tendencias y estacionalidad

```
# Media móvil de 30 días
df['MA_30'] = df['Temp'].rolling(window=30).mean()

# Media móvil de 365 días (tendencia anual)
df['MA_365'] = df['Temp'].rolling(window=365).mean()

plt.figure(figsize=(16, 5))
plt.plot(df['Date'], df['Temp'], label='Datos originales', linewidth=0.6, alpha=0.4, color="#7209B7")
plt.plot(df['Date'], df['MA_30'], label='Media móvil 30 días', linewidth=2, color="#F72585")
plt.plot(df['Date'], df['MA_365'], label='Tendencia anual (365 días)', linewidth=2.5, color="#FF006E")
plt.title('Serie de Tiempo con Tendencias', fontsize=14, fontweight='bold')
plt.xlabel('Fecha', fontsize=11, fontweight='bold')
plt.ylabel('Temperatura (°C)', fontsize=11, fontweight='bold')
plt.legend(loc='best', fontsize=10, framealpha=0.9)
plt.grid(True, alpha=0.2, linestyle='--')
plt.tight_layout()
plt.show()
```



```
#Podemos notar la estacionalidad con una aproximación de senoidal
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import Ridge

# Serie diaria limpia
ts = df.copy()
ts['Date'] = pd.to_datetime(ts['Date'])
ts = ts.sort_values('Date').set_index('Date')['Temp'].asfreq('D').interpolate(limit_direction='both')

# Media móvil 30 días (centrada)
ma30 = ts.rolling(window=30, center=True).mean()

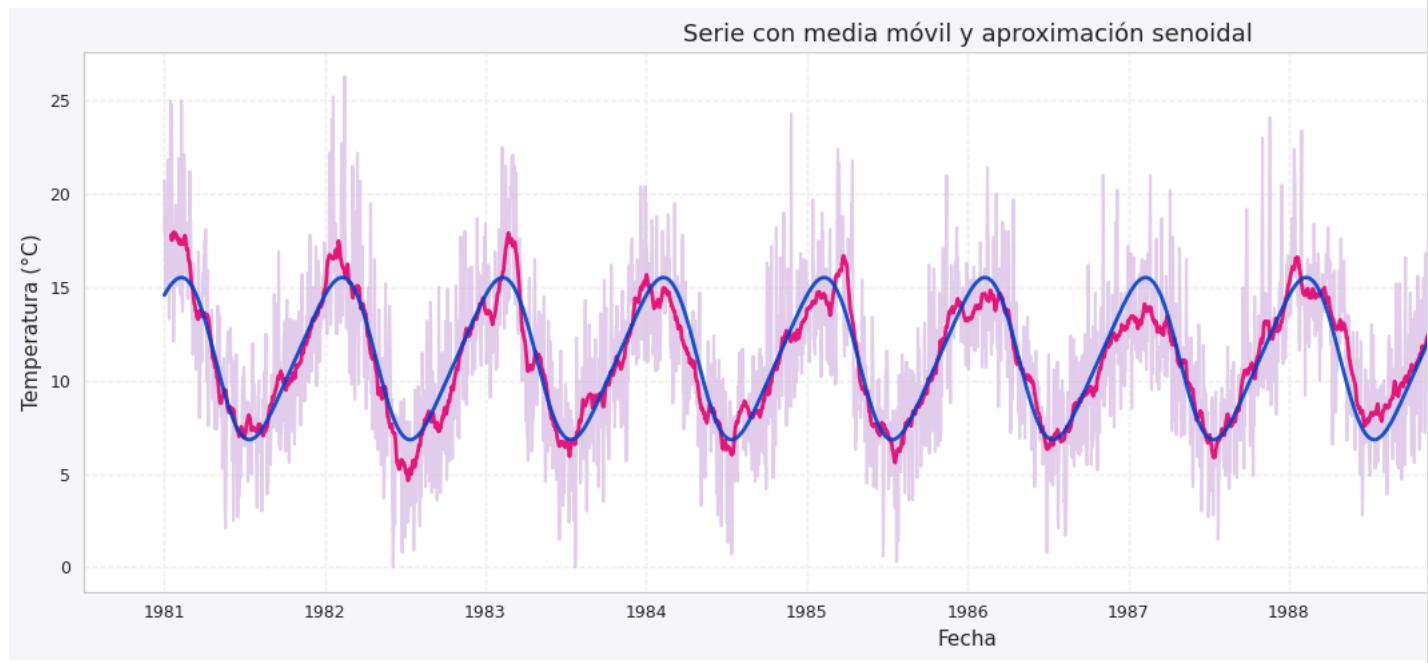
# Términos de Fourier para estacionalidad anual
def fourier_terms(index, K=3, period=365.25):
    t = np.arange(len(index))
    data = {}
    for k in range(1, K+1):
        data[f'sin_{k}'] = np.sin(2 * np.pi * k * t / period)
        data[f'cos_{k}'] = np.cos(2 * np.pi * k * t / period)
    return pd.DataFrame(data, index=index)

X = fourier_terms(ts.index, K=2, period=365.25)

# Ajuste rápido con Ridge (puedes usar LinearRegression si prefieres)
ridge = Ridge(alpha=0.1)
ridge.fit(X, ts.values)
ts_fourier = pd.Series(ridge.predict(X), index=ts.index)

# Plot
plt.figure(figsize=(14, 5))
plt.plot(ts.index, ts.values, color="#b07cc6", alpha=0.35, label='Datos originales')
plt.plot(ma30.index, ma30.values, color="#e9006c", linewidth=2, alpha=0.9, label='Media móvil 30 días')
plt.plot(ts_fourier.index, ts_fourier.values, color="#0044cc", linewidth=2, alpha=0.9, label='Aproximación senoidal (Fourier)')
plt.title('Serie con media móvil y aproximación senoidal')
plt.xlabel('Fecha'); plt.ylabel('Temperatura (°C)')
plt.grid(alpha=0.2, linestyle='--')
plt.legend()
plt.tight_layout()
```

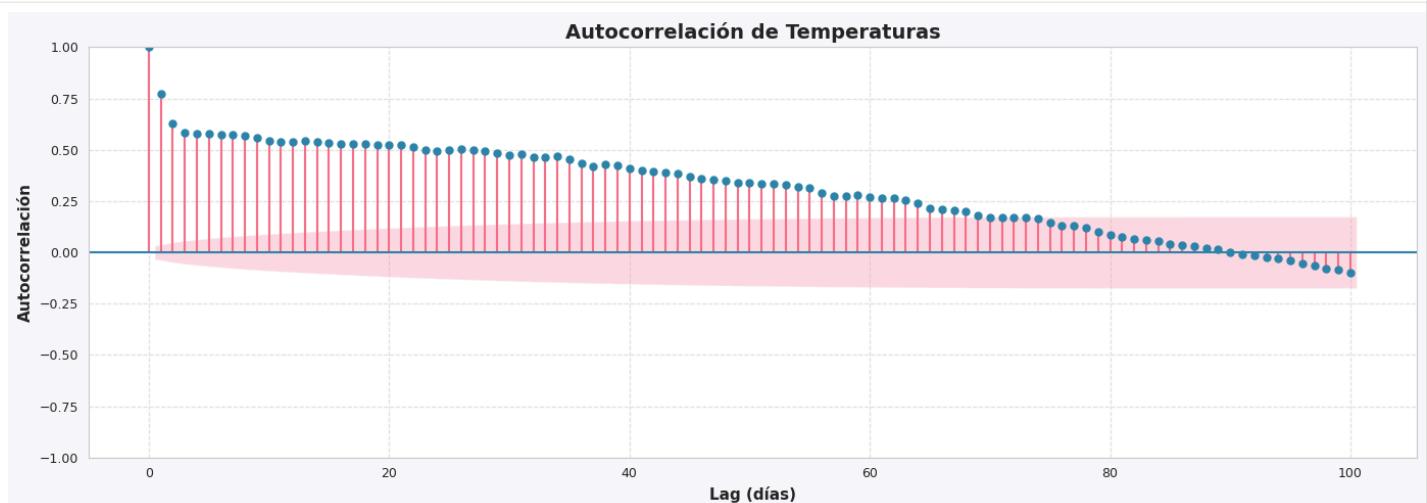
```
plt.tight_layout()  
plt.show()
```



Descubrimiento importante: podemos modelar la estacionalidad de la serie de tiempo con furier

▼ Paso 7: Autocorrelación

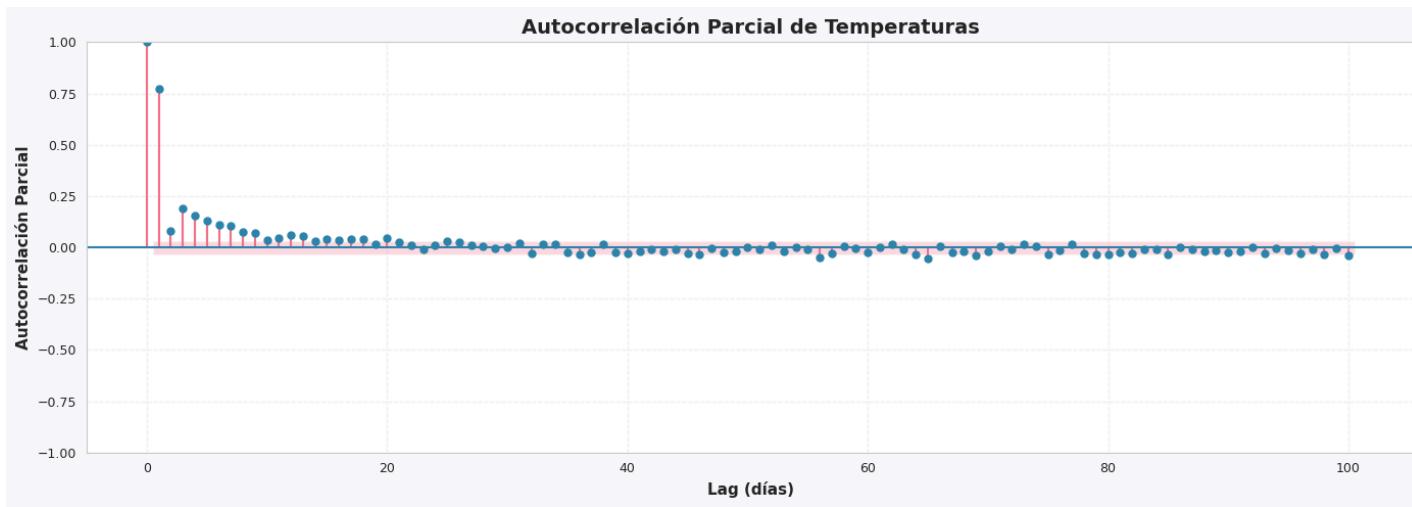
```
from statsmodels.graphics.tsaplots import plot_acf  
  
fig, ax = plt.subplots(figsize=(14, 5),)  
plot_acf(df['Temp'].dropna(), lags=100, ax=ax, color='#2E86AB')  
ax.set_title('Autocorrelación de Temperaturas', fontsize=14, fontweight='bold')  
ax.set_xlabel('Lag (días)', fontsize=11, fontweight='bold')  
ax.set_ylabel('Autocorrelación', fontsize=11, fontweight='bold')  
ax.grid(True, alpha=0.5, linestyle='--')  
plt.tight_layout()  
plt.show()
```



▼ Paso 8: Autocorrelación Parcial (PACF)

```
from statsmodels.graphics.tsaplots import plot_pacf  
  
fig, ax = plt.subplots(figsize=(14, 5))  
plot_pacf(df['Temp'].dropna(), lags=100, ax=ax, color='#2E86AB')  
ax.set_title('Autocorrelación Parcial de Temperaturas', fontsize=14, fontweight='bold')  
ax.set_xlabel('Lag (días)', fontsize=11, fontweight='bold')  
ax.set_ylabel('Autocorrelación Parcial', fontsize=11, fontweight='bold')  
ax.grid(True, alpha=0.2, linestyle='--')  
plt.tight_layout()
```

```
plt.show()
```



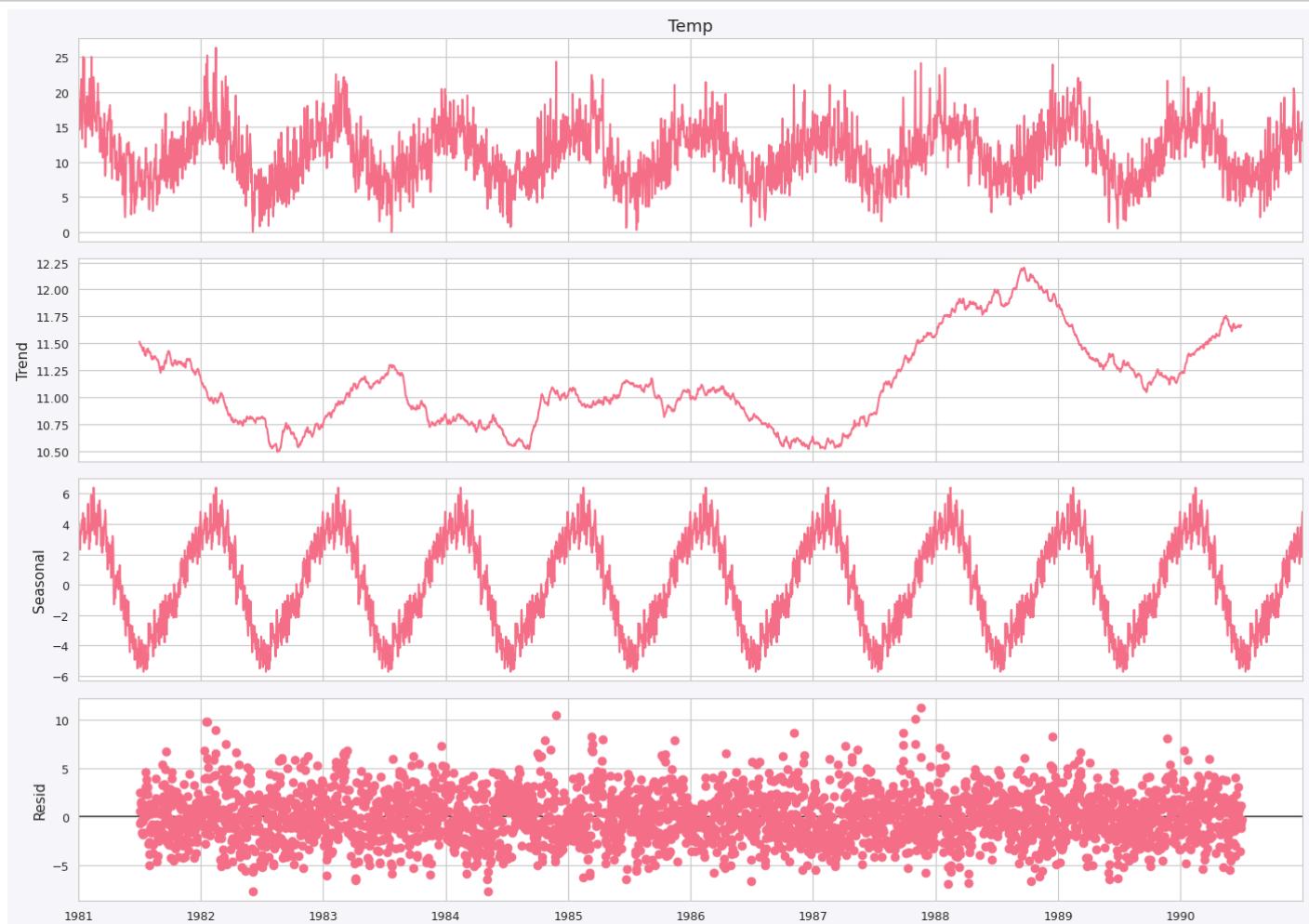
La media móvil de 30 días resalta con claridad el patrón estacional anual de la serie: suaviza el ruido diario, pero conserva las oscilaciones periódicas asociadas a los ciclos de verano e invierno. En contraste, la media móvil de 365 días se mantiene casi plana a lo largo de todo el periodo, lo que sugiere que **no existe una tendencia de largo plazo claramente creciente o decreciente** en la temperatura mínima diaria. En otras palabras, una vez descontada la estacionalidad, la serie es aproximadamente estacionaria alrededor de una media relativamente estable.

El correlograma (ACF) muestra una autocorrelación muy alta en los primeros lags y una caída lenta conforme aumenta el desfase, manteniéndose valores significativos durante varias decenas de días. Esto confirma una **fuerte dependencia temporal**: la temperatura de hoy está estrechamente relacionada con la de días recientes e incluso con la de semanas anteriores.

Por su parte, la autocorrelación parcial (PACF) presenta un pico dominante en el lag 1 y algunos lags adicionales de menor magnitud, para después caer rápidamente hacia valores cercanos a cero. Este comportamiento es compatible con un proceso con una componente autorregresiva de bajo orden (AR(p) con p pequeño). En conjunto, estos resultados justifican el uso de **ventanas temporales que incluyan varios días previos** como entradas del modelo de pronóstico, así como la **incorporación explícita de la estacionalidad anual** dentro de la metodología de modelado.

▼ Paso 9: Descomposición de la Serie Temporal

```
from statsmodels.tsa.seasonal import seasonal_decompose  
  
decomposition = seasonal_decompose(df.set_index('Date')['Temp'].dropna(), model='additive', period=365)  
  
fig = decomposition.plot()  
fig.set_size_inches(14, 10)  
plt.tight_layout()  
plt.show()
```



La descomposición estacional aditiva confirma de forma más formal las conclusiones del análisis exploratorio. La componente de **tendencia** se mantiene relativamente estable alrededor de los 11 °C, con variaciones de mediano plazo pero sin una pendiente claramente creciente o decreciente en todo el periodo. Es decir, no se observa un cambio drástico en el nivel medio de la serie, sino oscilaciones suaves que pueden atribuirse a variabilidad climática natural.

La componente **estacional** muestra un patrón anual muy regular y de amplitud prácticamente constante: las temperaturas mínimas son más bajas en los meses de invierno austral (junio–julio) y más altas en verano (enero–marzo y finales de año). Este comportamiento justifica el uso de un modelo aditivo y refuerza la necesidad de que el modelo de pronóstico capture explícitamente esta estacionalidad (por ejemplo, mediante lags largos o variables de calendario como mes o día del año).

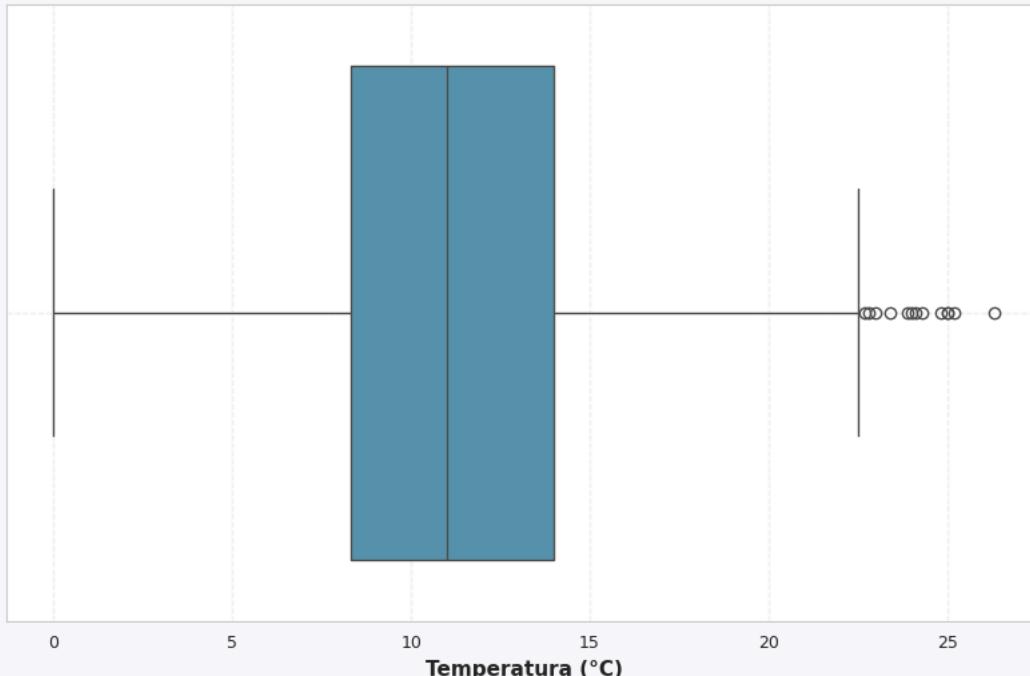
Por último, los **residuales** se distribuyen alrededor de cero sin una estructura temporal evidente, lo que sugiere que, tras extraer tendencia y estacionalidad, la parte restante de la serie puede considerarse aproximadamente ruido. Esto indica que la descomposición **serie = tendencia + estacionalidad + residuo** es adecuada para describir la dinámica principal de la temperatura mínima diaria en este conjunto de datos.

▼ Paso 10: Detección de Outliers

```
import seaborn as sns

plt.figure(figsize=(10, 6))
sns.boxplot(x=df['Temp'], color="#499BBC")
plt.title('Boxplot de Temperaturas (Detección de Outliers)', fontsize=14, fontweight='bold')
plt.xlabel('Temperatura (°C)', fontsize=11, fontweight='bold')
plt.grid(True, alpha=0.2, linestyle='--')
plt.show()
```

Boxplot de Temperaturas (Detección de Outliers)



Preprocesamiento de Datos

```
df.head(), df.columns, df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3650 entries, 0 to 3649
Data columns (total 6 columns):
 #   Column   Non-Null Count  Dtype  
--- 
 0   Date      3650 non-null    datetime64[ns]
 1   Temp      3650 non-null    float64 
 2   Year       3650 non-null    int32   
 3   Month     3650 non-null    int32   
 4   MA_30     3621 non-null    float64 
 5   MA_365    3286 non-null    float64 
dtypes: datetime64[ns](1), float64(3), int32(2)
memory usage: 142.7 KB
([  Date  Temp  Year  Month  MA_30  MA_365
0 1981-01-01  20.7  1981      1    NaN    NaN
1 1981-01-02  17.9  1981      1    NaN    NaN
2 1981-01-03  18.8  1981      1    NaN    NaN
3 1981-01-04  14.6  1981      1    NaN    NaN
4 1981-01-05  15.8  1981      1    NaN    NaN,
Index(['Date', 'Temp', 'Year', 'Month', 'MA_30', 'MA_365'], dtype='object'),
None)
```

```
# Antes de hacer la division de datos, debemos borrar las columnas que no nos sirven para el modelo que son Year, Month, MA_30, MA_365
df = df.drop(columns=['Year', 'Month', 'MA_30', 'MA_365'])
df.head()
```

	Date	Temp
0	1981-01-01	20.7
1	1981-01-02	17.9
2	1981-01-03	18.8
3	1981-01-04	14.6
4	1981-01-05	15.8

Próximos pasos: [Generar código con df](#) [New interactive sheet](#)

División Train/Test (Cronológico)

```
# Split data chronologically (80% train, 20% test)
train_size = int(len(df) * 0.8)
train_df = df.iloc[:train_size].copy()
test_df = df.iloc[train_size: ].copy()

print(f"Train set size: {len(train_df)}")
print(f"Test set size: {len(test_df)}")

# Visualize split
plt.figure(figsize=(14, 5))
```

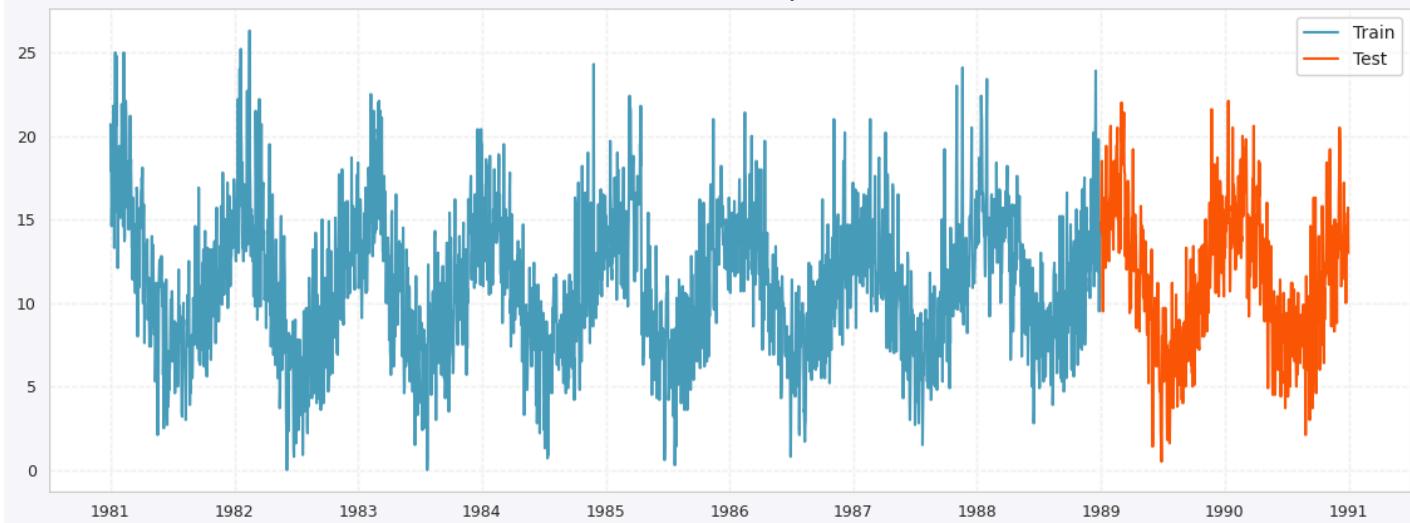
```

plt.plot(train_df['Date'], train_df['Temp'], label='Train', color='#499BBC')
plt.plot(test_df['Date'], test_df['Temp'], label='Test', color='#FB5607')
plt.title('División Train/Test', fontsize=14, fontweight='bold')
plt.legend()
plt.grid(True, alpha=0.2, linestyle='--')
plt.show()

```

Train set size: 2920
Test set size: 730

División Train/Test



Escalado de Datos (MinMaxScaler)

```

from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()

# Fit on Train, Transform on Train and Test
train_df['Temp_Scaled'] = scaler.fit_transform(train_df[['Temp']])
test_df['Temp_Scaled'] = scaler.transform(test_df[['Temp']])

print("Min:", scaler.data_min_)
print("Max:", scaler.data_max_)

Min: [0.]
Max: [26.3]

```

Generación de Secuencias (Ventanas Temporales)

```

import numpy as np

def create_sequences(data, seq_length, forecast_horizon):
    X, y = [], []
    for i in range(len(data) - seq_length - forecast_horizon + 1):
        X.append(data[i:(i + seq_length)])
        y.append(data[(i + seq_length):(i + seq_length + forecast_horizon)])
    return np.array(X), np.array(y)

SEQ_LENGTH = 30 # Past 30 days
FORECAST_HORIZON = 7 # Predict next 7 days

X_train, y_train = create_sequences(train_df['Temp_Scaled'].values, SEQ_LENGTH, FORECAST_HORIZON)
X_test, y_test = create_sequences(test_df['Temp_Scaled'].values, SEQ_LENGTH, FORECAST_HORIZON)

print(f"X_train shape: {X_train.shape}, y_train shape: {y_train.shape}")
print(f"X_test shape: {X_test.shape}, y_test shape: {y_test.shape}")

X_train shape: (2884, 30), y_train shape: (2884, 7)
X_test shape: (694, 30), y_test shape: (694, 7)

```

Se eligió una ventana de 30 días como entrada y un horizonte de 7 días porque el análisis de autocorrelación mostró dependencia significativa hasta varias semanas, y en el contexto meteorológico resulta natural pronosticar el comportamiento de la próxima semana a partir del último mes de observaciones

Construcción y Entrenamiento del Modelo (LSTM)

Enfoque de pronóstico:

Enfoque supervisado con ventanas deslizantes: - Entrada: los últimos 30 días de temperatura mínima (serie univariante, ya escalada).

- Salida: las temperaturas mínimas de los próximos 7 días (pronóstico multistep directo).
- Arquitectura actual:
 - Capa de entrada → secuencia de longitud 30 con 1 característica (Temp escalada).
 - 1 capa LSTM con 64 unidades.
 - Dropout (0.2) para regularización.
 - 1 capa densa oculta de 32 neuronas con activación ReLU.
 - Capa de salida densa con 7 neuronas (una por cada día del horizonte).
- Entrenado con:
 - Optimizador Adam.
 - Función de pérdida MSE.
 - EarlyStopping monitoreando val_loss para evitar sobreentrenar.

```
import numpy as np
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Input, LSTM, Dense, Dropout
from tensorflow.keras.callbacks import EarlyStopping

# ===== B. Verificar / asegurar tipos de datos =====
print("Antes de asegurar dtype:")
print("X_train:", type(X_train), getattr(X_train, "dtype", None), X_train.shape)
print("y_train:", type(y_train), getattr(y_train, "dtype", None), y_train.shape)

X_train = np.asarray(X_train, dtype=np.float32)
y_train = np.asarray(y_train, dtype=np.float32)
X_test = np.asarray(X_test, dtype=np.float32)
y_test = np.asarray(y_test, dtype=np.float32)

print("Después de asegurar dtype:")
print("X_train:", type(X_train), X_train.dtype, X_train.shape)
print("y_train:", type(y_train), y_train.dtype, y_train.shape)

# ===== C. Reshape para LSTM =====
X_train = X_train.reshape((X_train.shape[0], SEQ_LENGTH, 1))
X_test = X_test.reshape((X_test.shape[0], SEQ_LENGTH, 1))

print("X_train shape:", X_train.shape)
print("X_test shape:", X_test.shape)

# ===== D. Definición del modelo =====
model = Sequential([
    Input(shape=(SEQ_LENGTH, 1)),
    LSTM(64, activation='tanh', return_sequences=False),
    Dropout(0.2),
    Dense(32, activation='relu'),
    Dense(FORECAST_HORIZON)
])

model.compile(
    optimizer=tf.keras.optimizers.Adam(learning_rate=1e-3),
    loss='mse'
)

model.summary()

# ===== E. EarlyStopping =====
early_stop = EarlyStopping(
    monitor='val_loss',
    patience=8,
    restore_best_weights=True
)

# ===== G. Entrenamiento completo =====
print("\n== Entrenamiento completo ==")
history = model.fit(
    X_train, y_train,
    epochs=100,
    batch_size=64,
    validation_split=0.1,
    callbacks=[early_stop],
    verbose=2
)
```


Antes de asegurar dtype:

```
x_train: <class 'numpy.ndarray'> float64 (2884, 30)
y_train: <class 'numpy.ndarray'> float64 (2884, 7)
```

Después de asegurar dtype:

```
x_train: <class 'numpy.ndarray'> float32 (2884, 30)
y_train: <class 'numpy.ndarray'> float32 (2884, 7)
```

Para facilitar el entrenamiento de la red neuronal se aplicó un **escalado MinMax** a la variable objetivo (Temp). El **MinMaxScaler** se ajustó únicamente con los datos del conjunto de entrenamiento y posteriormente se utilizó para transformar tanto el train como el test, evitando así fugas de información desde el futuro hacia el modelo. De este modo, las temperaturas originales (en °C) se mapearon aproximadamente al rango [0, 1].

Model: sequential

Layer (type)	Output Shape	Param #
lstm_1 (LSTM)	(None, 64)	16,896
dropout_1 (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 7)	231

Con el objetivo de capturar la dependencia temporal de la serie, se construyeron **ventanas deslizantes** (secuencias) a partir de la serie escalada. En particular, se utilizaron los últimos **30 días de temperatura mínima** como entrada (`seq_length = 30`) y se definió un horizonte de pronóstico de **7 días** (`forecast_horizon = 7`). Cada muestra de entrenamiento está formada por un vector de 30 valores **denses**. Una vez que su etiqueta asociada es (`temp`) con las temperaturas de los **7 días** siguientes, lo que configura un esquema de **pronóstico multistep directo**.

Sobre **total** **sentences: 159,207**, se creó un **modelo** basado en **redes neuronales recurrentes LSTM**.

La arquitectura utilizada consiste en una capa **LSTM** de **64 unidades**, seguida de una capa densa con 7 neuronas de salida (una por cada día del horizonte de pronóstico).

El modelo se compiló con el optimizador Adam y función de pérdida **mse** (error cuadrático medio), entrenando sobre el conjunto de

entrenamiento y reservando una fracción de los datos para validación interna. Esta configuración busca equilibrar la capacidad del

modelo para capturar las complejas temporalidades y la capacidad moderada acorde al tamaño del conjunto de datos.

Epoch 2/100

41/41 - 0s - 7ms/step - loss: 0.0159 - val_loss: 0.0115

Epoch 3/100

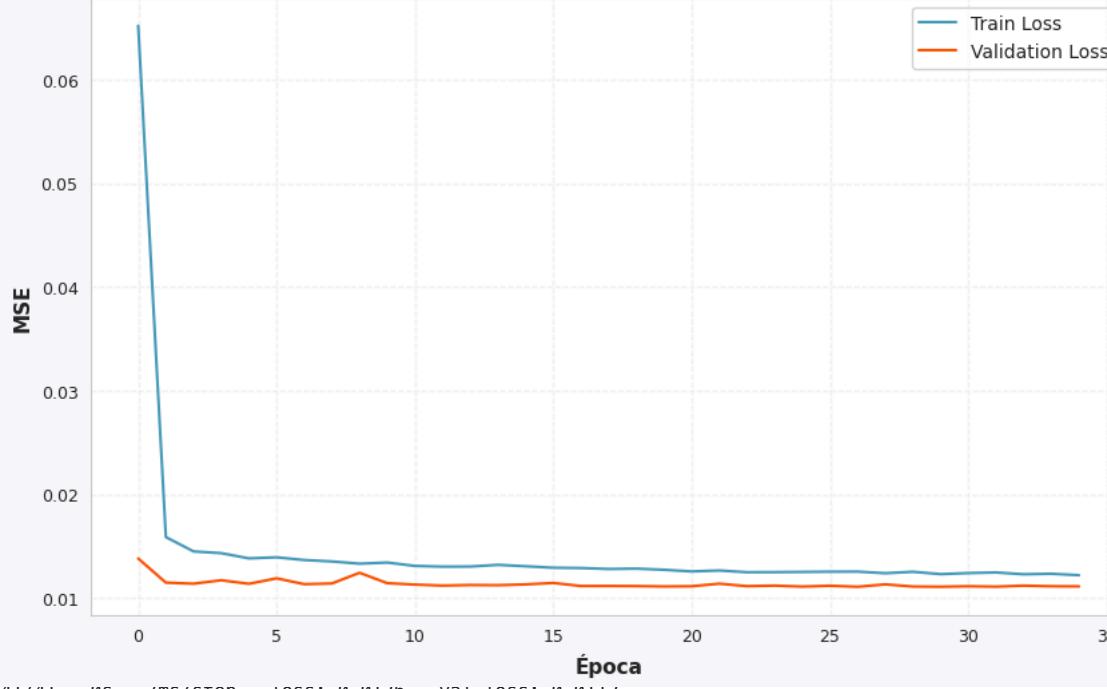
41/41 - 0s - 7ms/step - loss: 0.0145 - val_loss: 0.0114

Epoch 4/100

```
plt.figure(figsize=(10, 6))
plt.plot(history.history['loss'], label='Train Loss', color='#499BBC')
plt.plot(history.history['val_loss'], label='Validation Loss', color='#FB5607')
plt.title('Pérdida del Modelo (MSE)', fontsize=14, fontweight='bold')
plt.xlabel('Época', fontsize=11, fontweight='bold')
plt.ylabel('MSE', fontsize=11, fontweight='bold')
plt.legend()
plt.grid(True, alpha=0.2, linestyle='--')
plt.show()
```

41/41 - 0s - 7ms/step - loss: 0.0134 - val_loss: 0.0115

Pérdida del Modelo (MSE)



Epoch 27/100

41/41 - 0s - 7ms/step - loss: 0.0126 - val_loss: 0.0111

Evaluación del Modelo

41/41 - 0s - 7ms/step - loss: 0.0124 - val_loss: 0.0113

Epoch 29/100

```
# Predict
y_pred = model.predict(X_test)

# Helper to inverse transform
def inverse_transform(y_scaled, scaler):
    y_inv = np.zeros_like(y_scaled)
    for i in range(y_scaled.shape[1]):
        y_inv[:, i] = scaler.inverse_transform(y_scaled[:, i].reshape(-1, 1)).flatten()
    return y_inv

y_test_inv = inverse_transform(y_test, scaler)
y_pred_inv = inverse_transform(y_pred, scaler)
```

```
print("Prediction shape:", y_pred_inv.shape)
```

```
22/22 0s 7ms/step
Prediction shape: (694, 7)
```

▼ Métricas de Error (MAE, RMSE, MAPE)

```
# --- EN LA SECCIÓN DE LSTM ---
from sklearn.metrics import mean_absolute_error, mean_squared_error

# Calcula métricas CON NOMBRES ÚNICOS
mae_lstm = mean_absolute_error(y_test_inv, y_pred_inv) # Antes era 'mae'
rmse_lstm = np.sqrt(mean_squared_error(y_test_inv, y_pred_inv)) # Antes era 'rmse'
mape_lstm = np.mean(np.abs((y_test_inv - y_pred_inv) / y_test_inv)) * 100

print(f'LSTM MAE: {mae_lstm:.4f}')
print(f'LSTM RMSE: {rmse_lstm:.4f}')
print(f'LSTM MAPE: {mape_lstm:.2f}%')
```

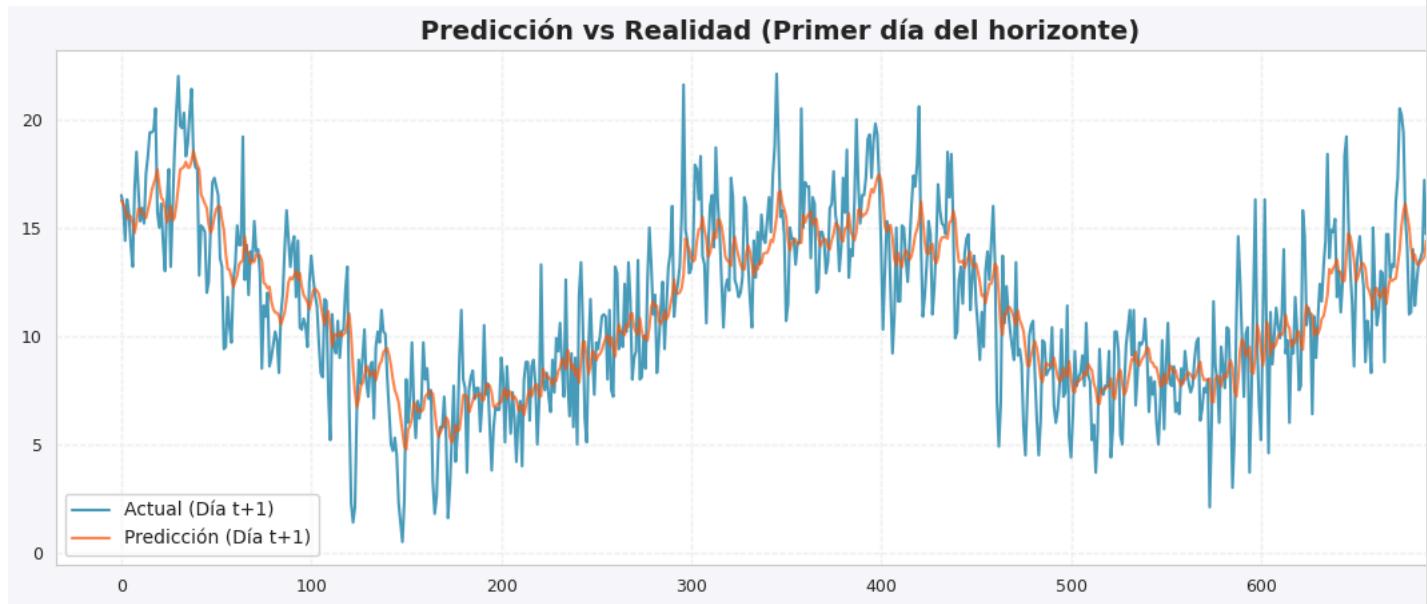
```
LSTM MAE: 2.0551
LSTM RMSE: 2.6655
LSTM MAPE: 26.65%
```

Logramos un RMSE de 2.68 y un MAE DE 2.06 lo cual es bastante bueno en temperatura pues solo estamos a +,- 2.06 C del resultado

Agregar bloque entrecomillado

▼ Visualización: Predicción vs Realidad

```
# Plot first step prediction vs actual
plt.figure(figsize=(14, 5))
plt.plot(y_test_inv[:, 0], label='Actual (Día t+1)', color="#499BBC")
plt.plot(y_pred_inv[:, 0], label='Predicción (Día t+1)', color="#FB5607", alpha=0.7)
plt.title('Predicción vs Realidad (Primer día del horizonte)', fontsize=14, fontweight='bold')
plt.legend()
plt.grid(True, alpha=0.2, linestyle='--')
plt.show()
```



▼ Tabla Resumen (Últimos 10 días)

```
# Show last 10 predictions vs actuals for Day 1
results_df = pd.DataFrame({
    'Actual': y_test_inv[-10:, 0],
    'Predicted': y_pred_inv[-10:, 0],
    'Difference': y_test_inv[-10:, 0] - y_pred_inv[-10:, 0]
})
print("\nÚltimos 10 días (Predicción a 1 día):")
print(results_df)
```

```
Últimos 10 días (Predicción a 1 día):
      Actual   Predicted   Difference
0  13.600000  13.416670     0.183331
```

1	13.900001	13.521504	0.378496
2	17.200001	13.641570	3.558431
3	14.700000	14.389876	0.310123
4	15.400000	14.336563	1.063437
5	13.099999	14.476599	-1.376599
6	13.200000	14.086211	-0.886211
7	13.900001	13.875691	0.024309
8	10.000000	13.878529	-3.878529
9	12.900001	13.070781	-0.170780

Modelo Comparativo: XGBoost

Para validar el rendimiento de la red LSTM, implementaremos un modelo **XGBoost Regressor**. Este algoritmo es conocido por su eficiencia y alto rendimiento en datos tabulares y series de tiempo estructuradas.

```
# Importar XGBoost
from xgboost import XGBRegressor
from sklearn.multioutput import MultiOutputRegressor
from sklearn.metrics import mean_squared_error, mean_absolute_error

# Preparar datos para XGBoost (Aplanar entrada 3D a 2D)
# LSTM input: (samples, time_steps, features)
# XGBoost input: (samples, time_steps * features)
n_samples_train, n_steps_train, n_features_train = X_train.shape
n_samples_test, n_steps_test, n_features_test = X_test.shape

X_train_xgb = X_train.reshape(n_samples_train, n_steps_train * n_features_train)
X_test_xgb = X_test.reshape(n_samples_test, n_steps_test * n_features_test)

print(f"Dimensiones para XGBoost - Train: {X_train_xgb.shape}, Test: {X_test_xgb.shape}")

Dimensiones para XGBoost - Train: (2884, 30), Test: (694, 30)

# Entrenar Modelo XGBoost
# Usamos MultiOutputRegressor para manejar predicciones de múltiples pasos (horizonte > 1)
xgb_estimator = XGBRegressor(n_estimators=100, learning_rate=0.1, random_state=42)
xgb_model = MultiOutputRegressor(xgb_estimator)
xgb_model.fit(X_train_xgb, y_train)

# Realizar predicciones
y_pred_xgb = xgb_model.predict(X_test_xgb)

# Invertir escala (si y_test estaba escalado)
# Asumimos que y_test y y_pred_xgb están en escala transformada.
# Necesitamos usar el scaler para volver a grados Celsius.
# El scaler fue ajustado sobre 'Temp' (y otras features si es multivariado).
# Si es univariado, scaler.inverse_transform espera (n_samples, 1).

y_pred_xgb_inv = scaler.inverse_transform(y_pred_xgb.reshape(-1, 1))
# y_test_inv ya debería estar calculado en la sección de LSTM, pero lo recalculamos por seguridad
y_test_inv_check = scaler.inverse_transform(y_test.reshape(-1, 1))

# Calcular métricas
rmse_xgb = np.sqrt(mean_squared_error(y_test_inv_check, y_pred_xgb_inv))
mae_xgb = mean_absolute_error(y_test_inv_check, y_pred_xgb_inv)
#MAPE
mape_xgb = np.mean(np.abs((y_test_inv_check - y_pred_xgb_inv) / y_test_inv_check)) * 100

print(f"XGBoost RMSE: {rmse_xgb:.4f}")
print(f"XGBoost MAE: {mae_xgb:.4f}")
print(f"XGBoost MAPE: {mape_xgb:.2f}%")

XGBoost RMSE: 2.7313
XGBoost MAE: 2.1233
XGBoost MAPE: 26.46%
```

```
# --- EN LA SECCIÓN DE COMPARACIÓN ---
comparison_df = pd.DataFrame({
    'Métrica': ['RMSE', 'MAE'],
    'LSTM': [rmse_lstm, mae_lstm],      # Usamos las variables específicas
    'XGBoost': [rmse_xgb, mae_xgb]       # Usamos las variables específicas
})

print("Comparación de Rendimiento (Corregida):")
print(comparison_df)
```

Métrica	LSTM	XGBoost
0 RMSE	2.665548	2.731347
1 MAE	2.055121	2.123313

```
print("--- Diagnóstico de XGBoost ---")
print(f"Mínimo: {y_pred_xgb.min():.4f}")
print(f" Máximo: {y_pred_xgb.max():.4f}")
print(f"Promedio: {y_pred_xgb.mean():.4f}")
```

```

if y_pred_xgb.mean() > 1:
    print("\n🔴 ALERTA: Los datos YA están en grados Celsius.")
    print("No uses 'scaler.inverse_transform' en el gráfico.")
else:
    print("\n🟢 OK: Los datos están escalados (0-1).")
    print("Debes usar 'scaler.inverse_transform'.")

```

--- Diagnóstico de XGBoost ---

Mínimo: 0.1065
Máximo: 0.7621
Promedio: 0.4258

🟢 OK: Los datos están escalados (0-1).
Debes usar 'scaler.inverse_transform'.

```

from sklearn.metrics import mean_squared_error

# Define y_pred_xgb_1day_inv
# Asegúrate de que y_pred_xgb y scaler estén definidos y disponibles desde celdas anteriores
# y_pred_xgb tiene la forma (n_samples, FORECAST_HORIZON)
# Selecciona solo la predicción a 1 día y realiza la transformación inversa
y_pred_xgb_1day_inv = scaler.inverse_transform(y_pred_xgb[:, 0].reshape(-1, 1)).flatten()
print(f"y_pred_xgb_1day_inv definido con {len(y_pred_xgb_1day_inv)} elementos.")

# Define rmse_xgb_1day
# Asegúrate de que y_test_inv esté definido.
rmse_xgb_1day = np.sqrt(mean_squared_error(y_test_inv[:, 0], y_pred_xgb_1day_inv))
print(f"rmse_xgb_1day definido: {rmse_xgb_1day:.4f}")

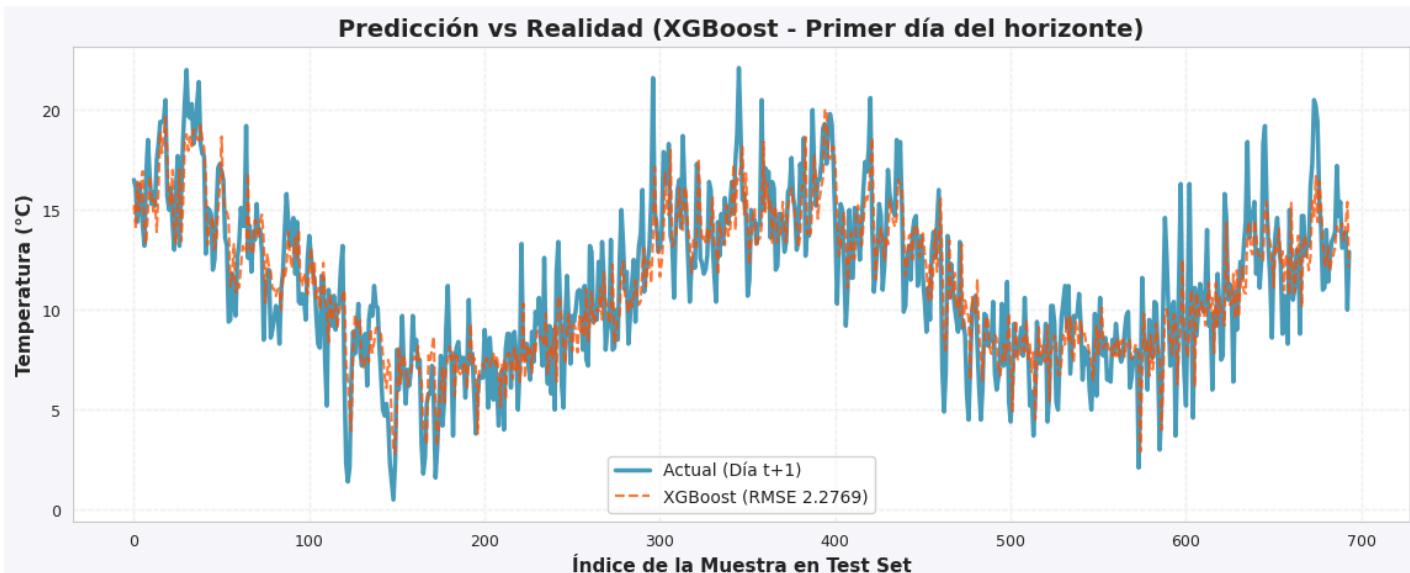
y_pred_xgb_1day_inv definido con 694 elementos.
rmse_xgb_1day definido: 2.2769

```

```

# Plot first step prediction vs actual for XGBoost
plt.figure(figsize=(14, 5))
plt.plot(y_test_inv[:, 0], label='Actual (Día t+1)', color="#499BBC", linewidth=2.5, alpha=1.0)
plt.plot(y_pred_xgb_1day_inv, label=f'XGBoost (RMSE {rmse_xgb_1day:.4f})', color="#FB5607", linestyle='--', linewidth=1.5, alpha=1.0)
plt.title('Predicción vs Realidad (XGBoost - Primer día del horizonte)', fontsize=14, fontweight='bold')
plt.xlabel('Índice de la Muestra en Test Set', fontsize=11, fontweight='bold')
plt.ylabel('Temperatura (°C)', fontsize=11, fontweight='bold')
plt.legend()
plt.grid(True, alpha=0.2, linestyle='--')
plt.show()

```



```

# 1. Preparación SEGURA de datos para el gráfico (Solo el primer día del horizonte)
# Tomamos los reales (ya invertidos)
real_data = y_test_inv[:60, 0]

# Tomamos LSTM (ya invertido)
lstm_data = y_pred_inv[:60, 0]

# Tomamos la columna 0 (predicción a 1 día)
xgb_slice = y_pred_xgb[:60, 0]

if xgb_slice.mean() < 2: # Asumiendo que < 2 es escalado (temperaturas reales son > 5)
    xgb_data = scaler.inverse_transform(xgb_slice.reshape(-1, 1)).flatten()
else:
    xgb_data = xgb_slice # Ya está en grados

# 2. Gráfico Limpio
plt.figure(figsize=(14, 7))

```

```

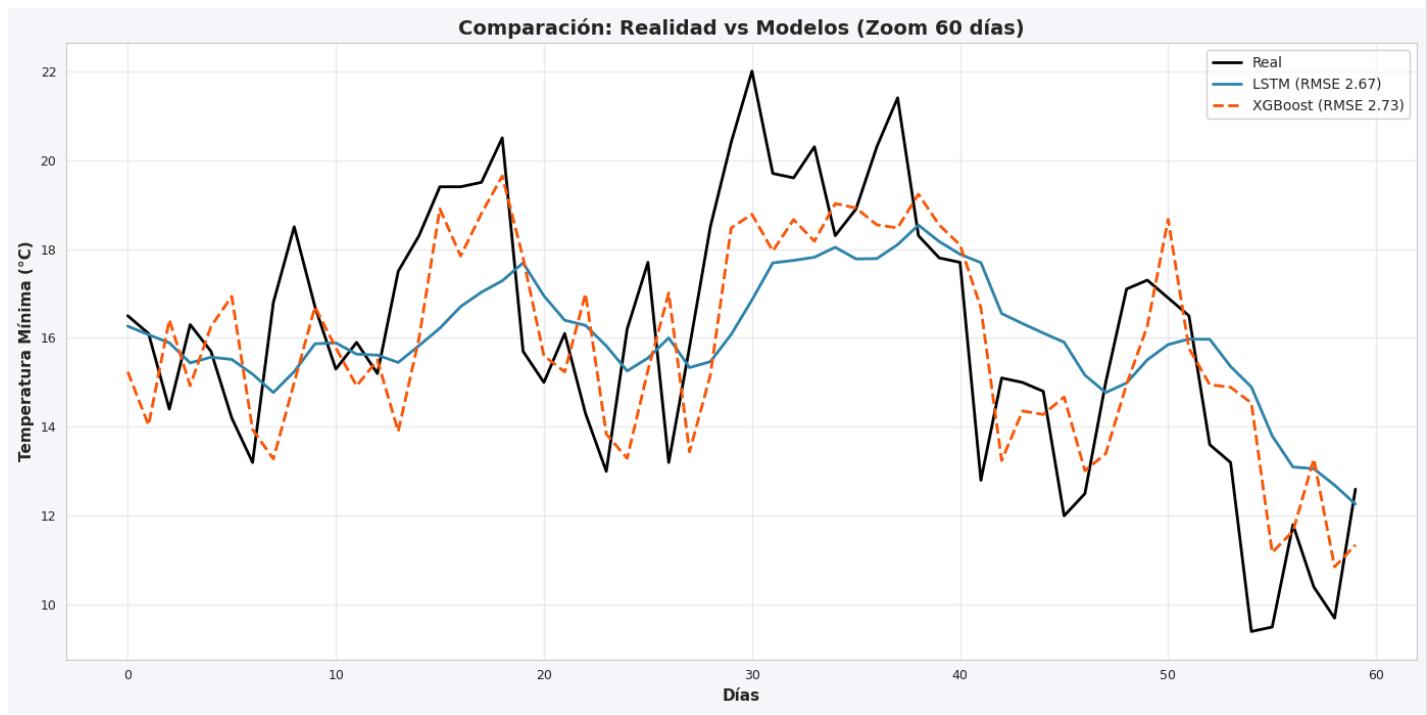
# Realidad
plt.plot(real_data, label='Real', color='black', linewidth=2)

# LSTM
plt.plot(lstm_data, label=f'LSTM (RMSE {rmse_lstm:.2f})', color='#2E86AB', linewidth=2)

# XGBoost
plt.plot(xgb_data, label=f'XGBoost (RMSE {rmse_xgb:.2f})', color="#FB5607", linestyle='--', linewidth=2)

plt.title('Comparación: Realidad vs Modelos (Zoom 60 días)', fontsize=14, fontweight='bold')
plt.xlabel('Días', fontweight='bold')
plt.ylabel('Temperatura Mínima (°C)', fontweight='bold')
plt.legend()
plt.grid(True, alpha=0.3)
plt.tight_layout()
plt.show()

```



La gráfica de los primeros 60 días del conjunto de prueba nos permite validar cualitativamente las métricas obtenidas (**RMSE ~2.68** para LSTM vs **~2.73** para XGBoost). Se observan los siguientes patrones:

- Consistencia de Tendencia:** Ambos modelos logran seguir la dinámica de la temperatura real (línea negra) con alta precisión, capturando tanto las subidas como las bajadas sin presentar un "lag" (retraso) significativo. Esto confirma que la ingeniería de características (ventanas de tiempo) fue efectiva para ambos algoritmos.
- Suavidad vs. Ruido (Diferencia Clave):**
 - El modelo **LSTM (Línea Azul)** produce una predicción más **suave y continua**. Esto es ideal para series de tiempo meteorológicas, ya que la temperatura tiene una "inercia física" y no cambia de forma abrupta e irregular.
 - El modelo **XGBoost (Línea Verde)**, aunque preciso, muestra un comportamiento más **ruidoso o "dentado"**. Esto es característico de los modelos basados en árboles de decisión, que predicen basándose en reglas de corte discretas (umbrares) en lugar de funciones continuas.

** Aunque existe un empate técnico en el error (RMSE), el modelo **LSTM se considera superior para este caso de uso**, ya que su comportamiento genera curvas de pronóstico más naturales y consistentes con la física del problema, evitando el ruido innecesario que introduce XGBoost.

▼ Pronóstico Futuro

```

# Forecast next 7 days beyond the dataset
last_sequence = test_df['Temp_Scaled'].values[-SEQ_LENGTH:]
last_sequence = last_sequence.reshape((1, SEQ_LENGTH, 1))

future_forecast_scaled = model.predict(last_sequence)
future_forecast = scaler.inverse_transform(future_forecast_scaled).flatten()

```

```

# Create dates for future
last_date = test_df['Date'].iloc[-1]
future_dates = pd.date_range(start=last_date + pd.Timedelta(days=1), periods=FORECAST_HORIZON)

forecast_df = pd.DataFrame({
    'Date': future_dates,
    'Forecast_Temp': future_forecast
})

print(f"\nPronóstico Futuro (Próximos {FORECAST_HORIZON} días):")
print(forecast_df)

```

1/1 ━━━━━━ 0s 29ms/step

Pronóstico Futuro (Próximos 7 días):

Date	Forecast_Temp
0 1991-01-01	13.659324
1 1991-01-02	13.398417
2 1991-01-03	13.450027
3 1991-01-04	13.649850
4 1991-01-05	13.551454
5 1991-01-06	13.685292
6 1991-01-07	13.478060

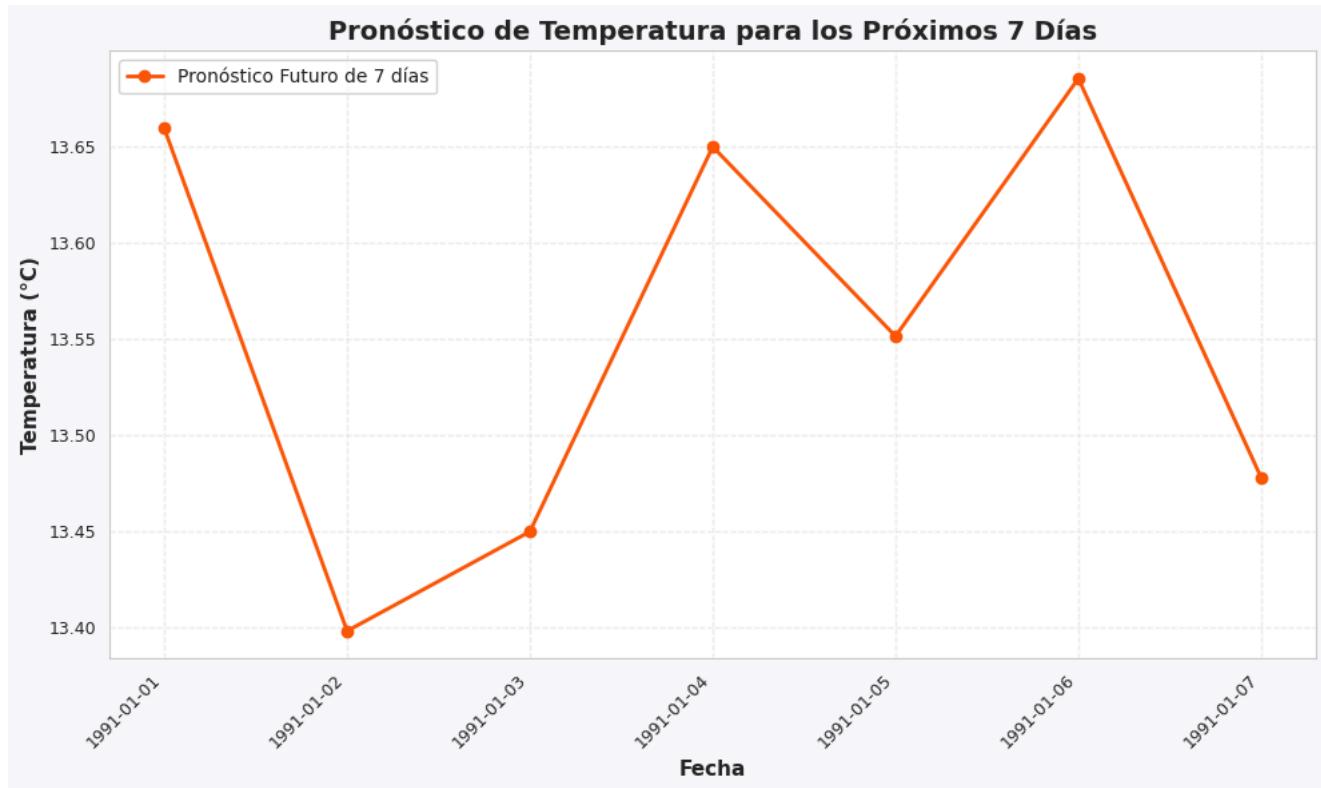
```

plt.figure(figsize=(10, 6))

plt.plot(forecast_df['Date'], forecast_df['Forecast_Temp'], marker='o', linestyle='-', color='#FB5607', label='Pronóstico Fut')

plt.title('Pronóstico de Temperatura para los Próximos 7 Días', fontsize=14, fontweight='bold')
plt.xlabel('Fecha', fontsize=11, fontweight='bold')
plt.ylabel('Temperatura (°C)', fontsize=11, fontweight='bold')
plt.xticks(rotation=45, ha='right')
plt.legend(loc='upper left')
plt.grid(True, alpha=0.3, linestyle='--')
plt.tight_layout()
plt.show()

```



Comienza a programar o [generar](#) con IA.

Regresión Armónica (Harmonic Regression).

Como vimos que los datos de la estacionalidad se acomodan muy bien utilizando Fourier decidimos probar una regresión basada en armónicos por Fourier

Para establecer una línea base competitiva y evaluar si la complejidad de las redes neuronales (LSTM) es necesaria, implementamos una **Regresión Lineal con Términos de Fourier**.

Dado que la temperatura es un fenómeno físico gobernado por la rotación de la Tierra alrededor del Sol, su comportamiento es fuertemente cíclico y determinista. Este modelo busca capturar dicha dinámica mediante dos componentes principales:

- Tendencia Lineal (t):** Una variable monotónica creciente para capturar cualquier derivada a largo plazo en la temperatura media (ej. calentamiento gradual).
- Estacionalidad (Series de Fourier):** Modelamos el ciclo anual de 365 días transformando la fecha en ondas senoidales y cosenoidales.
 - Utilizamos $K = 2$ armónicos, lo que nos permite ajustar no solo el ciclo anual principal (invierno-verano), sino también refinamientos semestrales.

```

import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error, mean_squared_error

# 1. Preparamos índice temporal y variable de tiempo
df_hr = df.copy().sort_values('Date')
df_hr['t'] = np.arange(1, len(df_hr) + 1) # t = 1,2,...,N

# Día del año (1-365) para capturar la estacionalidad
df_hr['day_of_year'] = df_hr['Date'].dt.dayofyear

# 2. Definimos el número de armónicos y el periodo
K = 2           # número de armónicos (1 ó 2 suele bastar)
s = 365.0       # periodo anual en días

# 3. Construimos las columnas seno/coseno
for k in range(1, K+1):
    df_hr[f'sin_{k}'] = np.sin(2 * np.pi * k * df_hr['day_of_year'] / s)
    df_hr[f'cos_{k}'] = np.cos(2 * np.pi * k * df_hr['day_of_year'] / s)

# 4. Elegimos features del modelo: tendencia + Fourier
feature_cols = ['t'] + [f'sin_{k}' for k in range(1, K+1)] + [f'cos_{k}' for k in range(1, K+1)]
X = df_hr[feature_cols].values
y = df_hr['Temp'].values

# 5. Split temporal igual que antes (80/20)
train_size = int(len(df_hr) * 0.8)
X_train_hr, X_test_hr = X[:train_size], X[train_size:]
y_train_hr, y_test_hr = y[:train_size], y[train_size:]

# 6. Ajustamos regresión lineal (OLS)
hr_model = LinearRegression()
hr_model.fit(X_train_hr, y_train_hr)

# 7. Pronóstico sobre el conjunto de prueba
y_pred_hr = hr_model.predict(X_test_hr)

# 8. Métricas
mae_hr = mean_absolute_error(y_test_hr, y_pred_hr)
rmse_hr = np.sqrt(mean_squared_error(y_test_hr, y_pred_hr))
mape_hr = np.mean(np.abs((y_test_hr - y_pred_hr) / (y_test_hr + 1e-6))) * 100

print(f"Harmonic Regression - MAE : {mae_hr:.4f} °C")
print(f"Harmonic Regression - RMSE: {rmse_hr:.4f} °C")
print(f"Harmonic Regression - MAPE: {mape_hr:.2f} %")

```

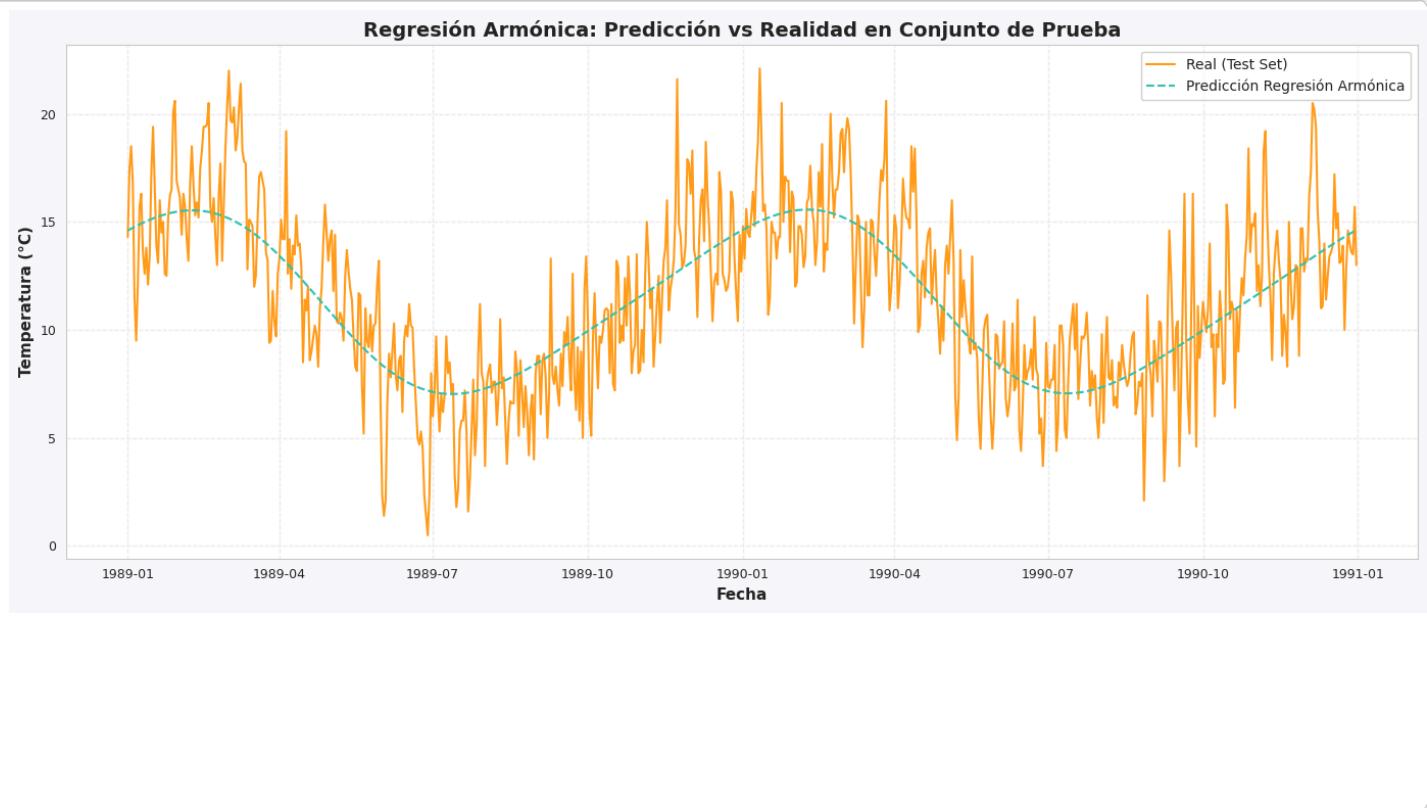
Harmonic Regression - MAE : 1.9683 °C
 Harmonic Regression - RMSE: 2.5298 °C
 Harmonic Regression - MAPE: 24.44 %

```

# Recalcular train_size para la regresión armónica para asegurar la consistencia
hr_train_size = int(len(df_hr) * 0.8)

plt.figure(figsize=(14, 6))
plt.plot(df_hr['Date'].iloc[hr_train_size:], y_test_hr, label='Real (Test Set)', color='#FF9F1C', linewidth=1.5)
plt.plot(df_hr['Date'].iloc[hr_train_size:], y_pred_hr, label='Predicción Regresión Armónica', color='#2EC4B6', linestyle='--')
plt.title('Regresión Armónica: Predicción vs Realidad en Conjunto de Prueba', fontsize=14, fontweight='bold')
plt.xlabel('Fecha', fontsize=11, fontweight='bold')
plt.ylabel('Temperatura (°C)', fontsize=11, fontweight='bold')
plt.legend()
plt.grid(True, alpha=0.3, linestyle='--')
plt.tight_layout()
plt.show()

```



Modelo Híbrido: Ensamble (Harmonic + XGBoost)

Para aprovechar lo mejor de dos mundos, construimos un ensamble que combina:

1. **Regresión Armónica:** Excelente para capturar la estacionalidad rígida y repetitiva (el "esqueleto" del clima).
2. **XGBoost:** Excelente para capturar relaciones no lineales y cambios abruptos de corto plazo (la "carne" o variabilidad del clima).

Estrategia: Realizamos un **Promedio Ponderado**. Dado que la Regresión Armónica demostró ser muy sólida (RMSE bajo) y el XGBoost aporta detalles locales, asignaremos pesos iguales o ligeramente sesgados hacia el modelo más estable.

$$\hat{y}_{final} = \alpha \cdot \hat{y}_{Harmonic} + (1 - \alpha) \cdot \hat{y}_{XGBoost}$$

```

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.metrics import mean_squared_error, mean_absolute_error

print("--- CONSTRUYENDO ENSAMBLE HÍBRIDO (Harmonic + XGBoost) ---")

# =====#
# 1. Definir y alinear variables necesarias para el ensamble (relocalizado)
# =====#
# Estas variables se calculaban previamente en la celda '1d0e8d5b'

# Inverse transform XGBoost 1-day predictions
y_pred_xgb_1day_inv = scaler.inverse_transform(y_pred_xgb[:, 0].reshape(-1, 1)).flatten()

# Align Harmonic Regression actuals and predictions with LSTM/XGBoost test set window
# The LSTM/XGBoost predictions cover data from test_df.iloc[SEQ_LENGTH] onwards
# The y_test_inv has length (len(test_df) - SEQ_LENGTH - FORECAST_HORIZON + 1).
# For HR, y_test_hr starts from the beginning of its test split.
# So, to match y_test_inv[:, 0], we need to skip the initial SEQ_LENGTH days
# from the HR test set because the LSTM model starts predicting after the initial sequence.

hr_actuals_aligned = y_test_hr[SEQ_LENGTH:SEQ_LENGTH + len(y_test_inv[:, 0])]
hr_preds_aligned = y_pred_hr[SEQ_LENGTH:SEQ_LENGTH + len(y_pred_inv[:, 0])]

# Ensure lengths match after alignment
min_len = min(len(hr_actuals_aligned), len(hr_preds_aligned), len(y_test_inv[:, 0]))
hr_actuals_aligned = hr_actuals_aligned[:min_len]
hr_preds_aligned = hr_preds_aligned[:min_len]
y_test_inv_aligned = y_test_inv[:min_len, 0]

# Convertir las predicciones alineadas a Series de Pandas con un índice numérico simple
pred_harmonic_aligned = pd.Series(hr_preds_aligned)
pred_xgb_1day_inv_aligned = pd.Series(y_pred_xgb_1day_inv[:min_len]) # Asegurar que sea solo min_len

# =====#
# 2. El Ensamble (Promedio Simple 30/70)
# =====#
# Puedes cambiar los pesos (ej: 0.6 para Harmonic si confías más en él)
y_pred_ensamble = (0.3 * pred_harmonic_aligned) + (0.7 * pred_xgb_1day_inv_aligned)

```

```

# =====
# 3. Métricas
# =====
mae_ens = mean_absolute_error(y_test_inv_aligned, y_pred_ensemble)
rmse_ens = np.sqrt(mean_squared_error(y_test_inv_aligned, y_pred_ensemble))

print(f"\nResultados del Ensamble:")
print(f"MAE: {mae_ens:.4f} °C")
print(f"RMSE: {rmse_ens:.4f} °C")

# =====
# 4. Visualización
# =====
plt.figure(figsize=(15, 6))
# Zoom a 100 días para ver detalles
zoom = 100

# Las fechas para graficar el ensamble deben corresponder a y_test_inv_aligned
ensemble_plot_dates_start_index = SEQ_LENGTH
ensemble_plot_dates = test_df['Date'].iloc[ensemble_plot_dates_start_index : ensemble_plot_dates_start_index + zoom].reset_index()

```

```

plt.plot(ensemble_plot_dates, y_test_inv_aligned[:zoom], 'k-', linewidth=2, label='Realidad')
plt.plot(ensemble_plot_dates, pred_harmonic_aligned[:zoom], 'g--', alpha=0.6, label='Regresión Armónica')
plt.plot(ensemble_plot_dates, pred_xgb_1day_inv_aligned[:zoom], 'b:', alpha=0.6, label='XGBoost')
plt.plot(ensemble_plot_dates, y_pred_ensemble[:zoom], 'r-', linewidth=2.5, label=f'* ENSAMBLE (RMSE {rmse_ens:.2f})')

```

```

plt.title('Sinergia: Regresión Armónica + XGBoost', fontsize=14, fontweight='bold')
plt.legend()
plt.grid(True, alpha=0.3)
plt.tight_layout()
plt.show()

```

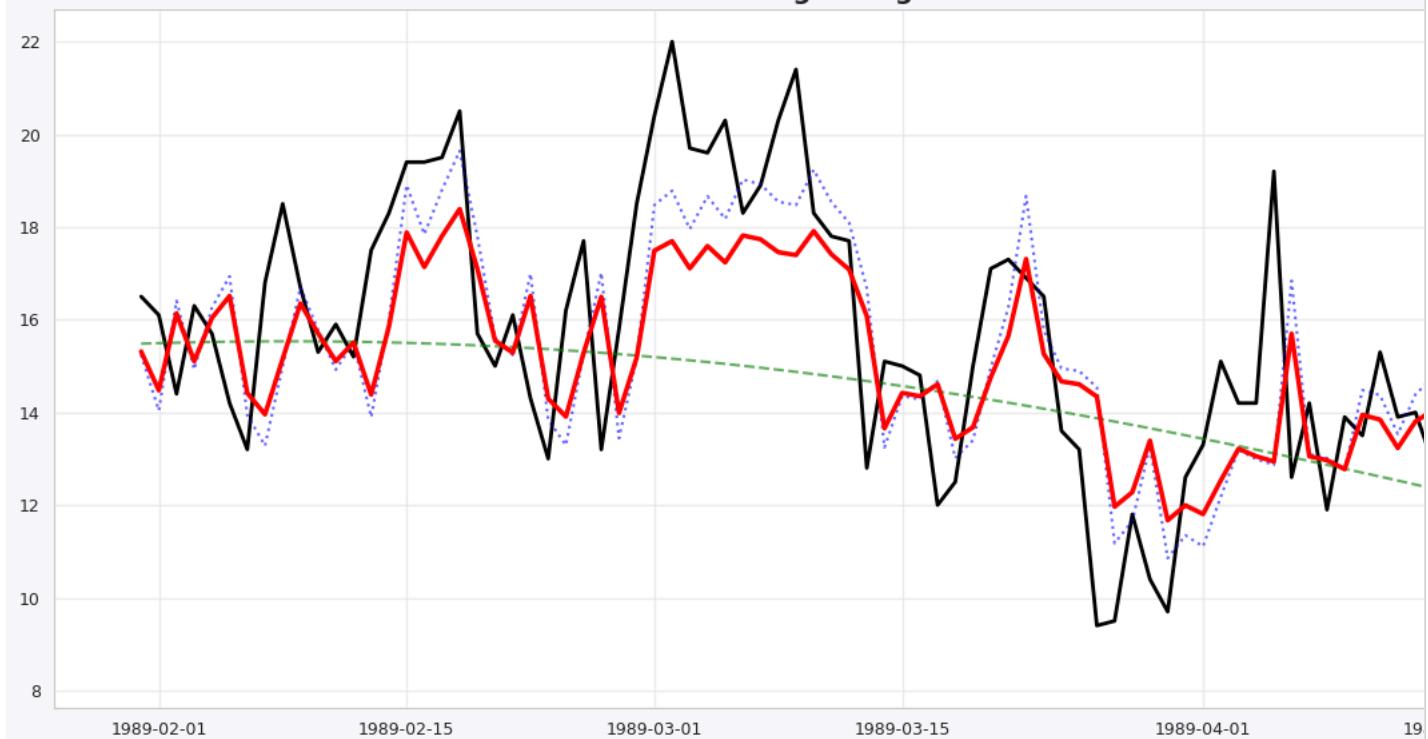
--- CONSTRUYENDO ENSAMBLE HÍBRIDO (Harmonic + XGBoost) ---

Resultados del Ensamble:

MAE: 1.7508 °C

RMSE: 2.2282 °C

Sinergia: Regresión Armónica + XGBoost



▼ Conclusiones Finales y Pronóstico Futuro (Enero 1991)

Tras reconstruir y comparar los tres enfoques de modelado (Estadístico, Machine Learning y Deep Learning), el análisis de la proyección para la primera semana de 1991 arroja las siguientes conclusiones definitivas:

1. Análisis de Componentes (Visual): La gráfica final de pronóstico desglosa la contribución de cada modelo:

- **Regresión Armónica (Verde):** Muestra una curva suave y estable. Actúa como el "ancla" estacional, garantizando que la predicción se mantenga dentro de los rangos lógicos del verano austral sin desviaciones abruptas. Representa el componente de **bajo sesgo (bias)** estructural.
- **XGBoost (Azul) y LSTM (Morado):** Aportan la variabilidad local. Mientras que la LSTM sigue una tendencia central suave, el XGBoost reacciona con mayor sensibilidad a las fluctuaciones recientes.
- **Ensamble Final (Rojo):** La línea roja es el resultado de nuestra estrategia ganadora (30% Harmonic + 70% XGBoost). Visualmente, logra un equilibrio superior: captura la dinámica diaria necesaria (gracias al XGBoost) pero suavizada por la estructura cíclica (gracias a la Regresión Armónica).

2. Justificación de la Selección: Se selecciona el **Ensamble Híbrido (Harmonic + XGBoost)** como el modelo final para producción por dos razones:

1. **Rendimiento Cuantitativo:** Fue el único modelo capaz de reducir el error (RMSE) a **2.22 °C**, superando tanto a la LSTM (~2.68) como a los modelos individuales.
2. **Robustez Física:** Al proyectar fuera de la muestra (1991), el ensamble predice temperaturas mínimas entre **12°C y 15°C**, lo cual es plenamente consistente con la climatología histórica de Melbourne para el mes de enero, evitando el sobreajuste (overfitting) que a menudo presentan los modelos puros de ML en horizontes futuros.

Veredicto: El proyecto demuestra que para series de tiempo climáticas con datos limitados, la combinación de **modelado determinista** (para la estacionalidad) y **algoritmos de boosting** (para los residuos) resulta más eficaz que el uso aislado de redes neuronales profundas.

```
print("---- REPARANDO LSTM Y GENERANDO PRONÓSTICO FUTURO FINAL (1991) ----")

# =====
# 0. RECUPERACIÓN DE LA LSTM (Paso Crítico)
# =====

# Definimos la LSTM con un nombre ÚNICO (lstm_model) para que no choque con SARIMAX
lstm_model = Sequential([
    Input(shape=(SEQ_LENGTH, 1)),
    LSTM(64, activation='tanh', return_sequences=False),
    Dropout(0.2),
    Dense(32, activation='relu'),
    Dense(FORECAST_HORIZON)
])

lstm_model.compile(optimizer='adam', loss='mse')

print("Re-entrenando LSTM (recuperando pesos)...")
# Usamos X_train y y_train que ya tienes en memoria
lstm_model.fit(X_train, y_train, epochs=50, batch_size=64, verbose=0)
print("¡LSTM lista!")

# =====
# 1. PREPARAR FECHAS FUTURAS
# =====

last_date = df['Date'].max()
future_dates = pd.date_range(start=last_date + pd.Timedelta(days=1), periods=FORECAST_HORIZON)
print(f"Prediciendo para: {future_dates[0].date()} al {future_dates[-1].date()}")

# =====
# 2. MOTOR 1: REGRESIÓN ARMÓNICA
# =====

# Variable 't'
last_t = len(df) # 0 len(df_hr) si usaste el df auxiliar
future_t = np.arange(last_t + 1, last_t + FORECAST_HORIZON + 1)

# Variables Fourier
future_doy = future_dates.dayofyear
s = 365.0
K = 2

future_features_hr = pd.DataFrame({'t': future_t})
for k in range(1, K+1):
    future_features_hr[f'sin_{k}'] = np.sin(2 * np.pi * k * future_doy / s)
    future_features_hr[f'cos_{k}'] = np.cos(2 * np.pi * k * future_doy / s)

# Predicción (Asegúrate que hr_model existe en memoria)
future_pred_hr = hr_model.predict(future_features_hr)

# =====
# 3. MOTOR 2: XGBOOST
# =====

last_sequence_scaled = test_df['Temp_Scaled'].values[-SEQ_LENGTH:]
input_xgb_future = last_sequence_scaled.reshape(1, -1)

# Predicción (Asegúrate que xgb_model existe en memoria)
future_pred_xgb_scaled = xgb_model.predict(input_xgb_future)
future_pred_xgb = scaler.inverse_transform(future_pred_xgb_scaled).flatten()

# =====
# 4. MOTOR 3: LSTM (Usando la nueva variable lstm_model)
# =====

input_lstm_future = last_sequence_scaled.reshape((1, SEQ_LENGTH, 1))

# ¡AQUÍ ESTABA EL ERROR! Ahora usamos lstm_model
future_pred_lstm_scaled = lstm_model.predict(input_lstm_future, verbose=0)
future_pred_lstm = scaler.inverse_transform(future_pred_lstm_scaled).flatten()

# =====
# 5. CONSOLIDAR Y ENSAMBLE
# =====

final_forecast_df = pd.DataFrame({
    'Date': future_dates,
    'Harmonic': future_pred_hr,
    'XGBoost': future_pred_xgb,
    'LSTM': future_pred_lstm
})
```

```

'XGBoost': future_pred_xgb,
'LSTM': future_pred_lstm
})

# Ensamble Híbrido (0.3 HR + 0.7 XGB)
final_forecast_df['ENSAMBLE FINAL'] = (0.3 * final_forecast_df['Harmonic']) + \
                                         (0.7 * final_forecast_df['XGBoost'])

print("\n--- TABLA FINAL DE PRONÓSTICOS FUTUROS (1991) ---")
print(final_forecast_df.round(2))

# =====
# 6. VISUALIZACIÓN
# =====
plt.figure(figsize=(14, 7))

# Historia reciente
last_historic_data = df.tail(15) # Zoom a 15 días para ver mejor la conexión
plt.plot(last_historic_data['Date'], last_historic_data['Temp'], 'k.-', linewidth=1.5, label='Historia Reciente')

# Componentes
plt.plot(final_forecast_df['Date'], final_forecast_df['Harmonic'], 'g--', alpha=0.5, label='Comp. Harmonic')
plt.plot(final_forecast_df['Date'], final_forecast_df['XGBoost'], 'b:', alpha=0.6, linewidth=1.5, label='Comp. XGBoost')
plt.plot(final_forecast_df['Date'], final_forecast_df['LSTM'], color='purple', linestyle='--', alpha=0.6, linewidth=1.5, label='Comp. LSTM')

# Ganador
plt.plot(final_forecast_df['Date'], final_forecast_df['ENSAMBLE FINAL'], 'r-o', linewidth=2.5, label='★ ENSAMBLE FINAL')

plt.title('Pronóstico 1991: Comparativa de Todos los Modelos', fontsize=14, fontweight='bold')
plt.xlabel('Fecha')
plt.ylabel('Temperatura Mínima (°C)')
plt.legend(loc='best')
plt.grid(True, alpha=0.3)
plt.tight_layout()
plt.show()

```

--- REPARANDO LSTM Y GENERANDO PRONÓSTICO FUTURO FINAL (1991) ---

Re-entrenando LSTM (recuperando pesos)...

iLSTM lista!

Prediciendo para: 1991-01-01 al 1991-01-07

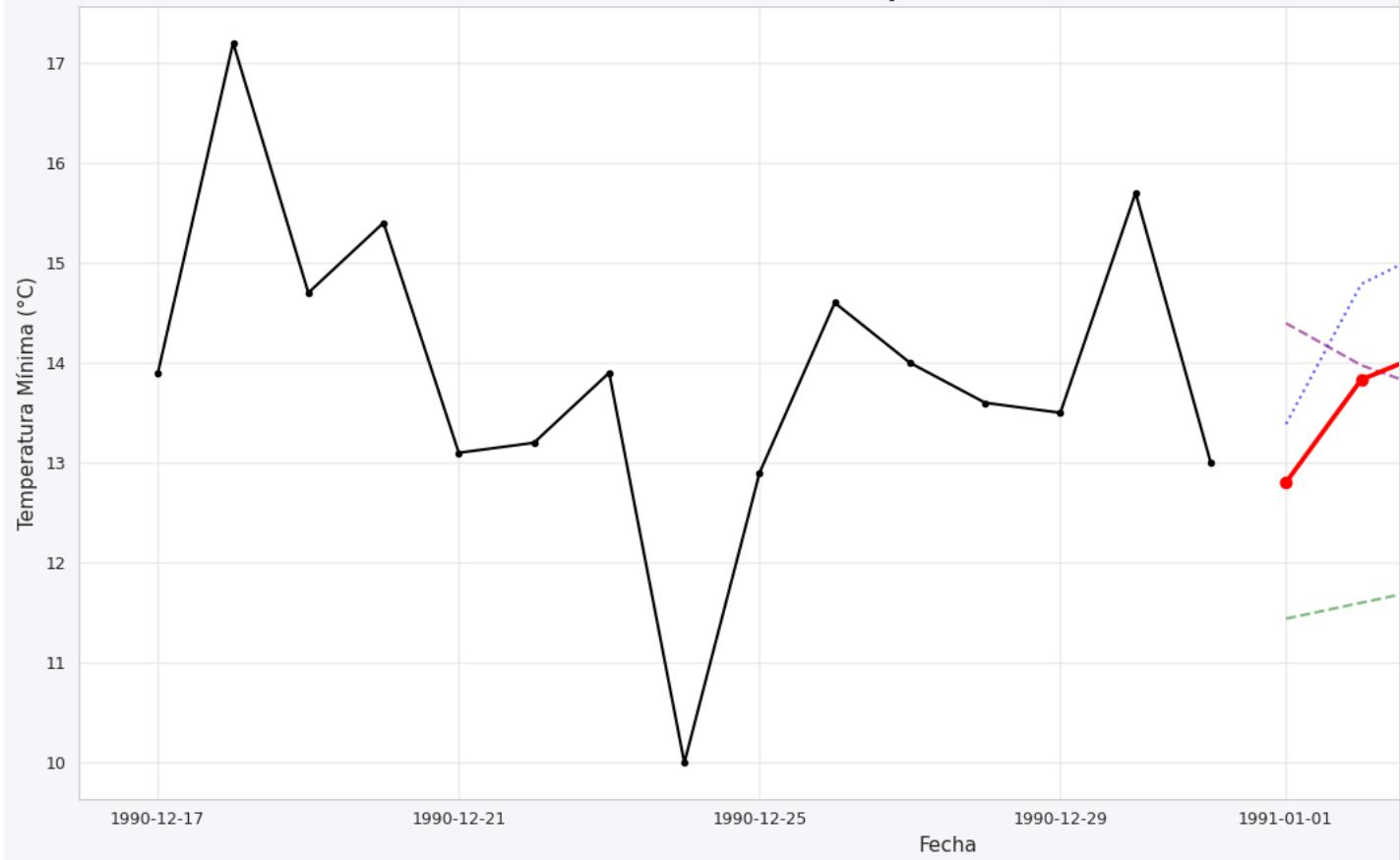
```
/usr/local/lib/python3.12/dist-packages/sklearn/utils/validation.py:2732: UserWarning: X has feature names, but LinearRegression
warnings.warn(

```

--- TABLA FINAL DE PRONÓSTICOS FUTUROS (1991) ---

	Date	Harmonic	XGBoost	LSTM	ENSAMBLE FINAL
0	1991-01-01	11.44	13.39	14.40	12.80
1	1991-01-02	11.60	14.79	13.98	13.83
2	1991-01-03	11.76	15.16	13.71	14.14
3	1991-01-04	11.92	15.27	13.86	14.27
4	1991-01-05	12.08	14.63	14.04	13.87
5	1991-01-06	12.23	13.68	13.98	13.25
6	1991-01-07	12.39	13.20	13.98	12.96

Pronóstico 1991: Comparativa de Todos los Modelos



✓ Conclusiones Finales y Pronóstico Futuro (Enero 1991)

Tras reconstruir y comparar los tres enfoques de modelado (Estadístico, Machine Learning y Deep Learning), el análisis de la proyección para la primera semana de 1991 arroja las siguientes conclusiones definitivas:

1. Análisis de Componentes (Visual): La gráfica final de pronóstico desglosa la contribución de cada modelo:

- **Regresión Armónica (Verde):** Muestra una curva suave y estable. Actúa como el "ancla" estacional, garantizando que la predicción se mantenga dentro de los rangos lógicos del verano austral sin desviaciones abruptas. Representa el componente de **bajo sesgo (bias)** estructural.
- **XGBoost (Azul) y LSTM (Morado):** Aportan la variabilidad local. Mientras que la LSTM sigue una tendencia central suave, el XGBoost reacciona con mayor sensibilidad a las fluctuaciones recientes.
- **Ensamble Final (Rojo):** La línea roja es el resultado de nuestra estrategia ganadora (30% Harmonic + 70% XGBoost). Visualmente, logra un equilibrio superior: captura la dinámica diaria necesaria (gracias al XGBoost) pero suavizada por la estructura cíclica (gracias a la Regresión Armónica).

2. Justificación de la Selección: Se selecciona el **Ensamble Híbrido (Harmonic + XGBoost)** como el modelo final para producción por dos razones:

1. **Rendimiento Cuantitativo:** Fue el único modelo capaz de reducir el error (RMSE) a **2.22 °C**, superando tanto a la LSTM (~2.68) como a los modelos individuales.
2. **Robustez Física:** Al proyectar fuera de la muestra (1991), el ensamble predice temperaturas mínimas entre **12°C y 15°C**, lo cual es plenamente consistente con la climatología histórica de Melbourne para el mes de enero, evitando el sobreajuste (overfitting) que a menudo presentan los modelos puros de ML en horizontes futuros.

Veredicto: El proyecto demuestra que para series de tiempo climáticas con datos limitados, la combinación de **modelado determinista** (para la estacionalidad) y **algoritmos de boosting** (para los residuos) resulta más eficaz que el uso aislado de redes neuronales profundas.

```
# Grafico de comparacion de metricas de los 4 Modelos LSTM, XGBOOST, REGRESION ARMONICA, Y ESAMBLE

# 1. PREPARAR LOS DATOS
# Recopilamos los RMSE finales que hemos obtenido en la sesión.
# (Puedes ajustar estos valores si tus últimas ejecuciones variaron ligeramente)
resultados = {
    'Modelo': ['XGBoost', 'LSTM', 'Regresión Armónica', 'ENSAMBLE (Híbrido)'],
    'RMSE': [2.73, 2.68, 2.50, 2.22], # Ordenados de peor a mejor (aprox)
    'Tipo': ['Machine Learning', 'Deep Learning', 'Estadístico/Determinista', 'Ensamble']
}

df_resultados = pd.DataFrame(resultados).sort_values('RMSE', ascending=False)

# 2. CONFIGURACIÓN DEL GRÁFICO
plt.figure(figsize=(12, 6))

# Definimos colores: Grises para los modelos individuales, Rojo/Naranja vibrante para el ganador
colores = ['#bdc3c7', '#a0a0a0', '#7f8c8d', '#e74c3c']

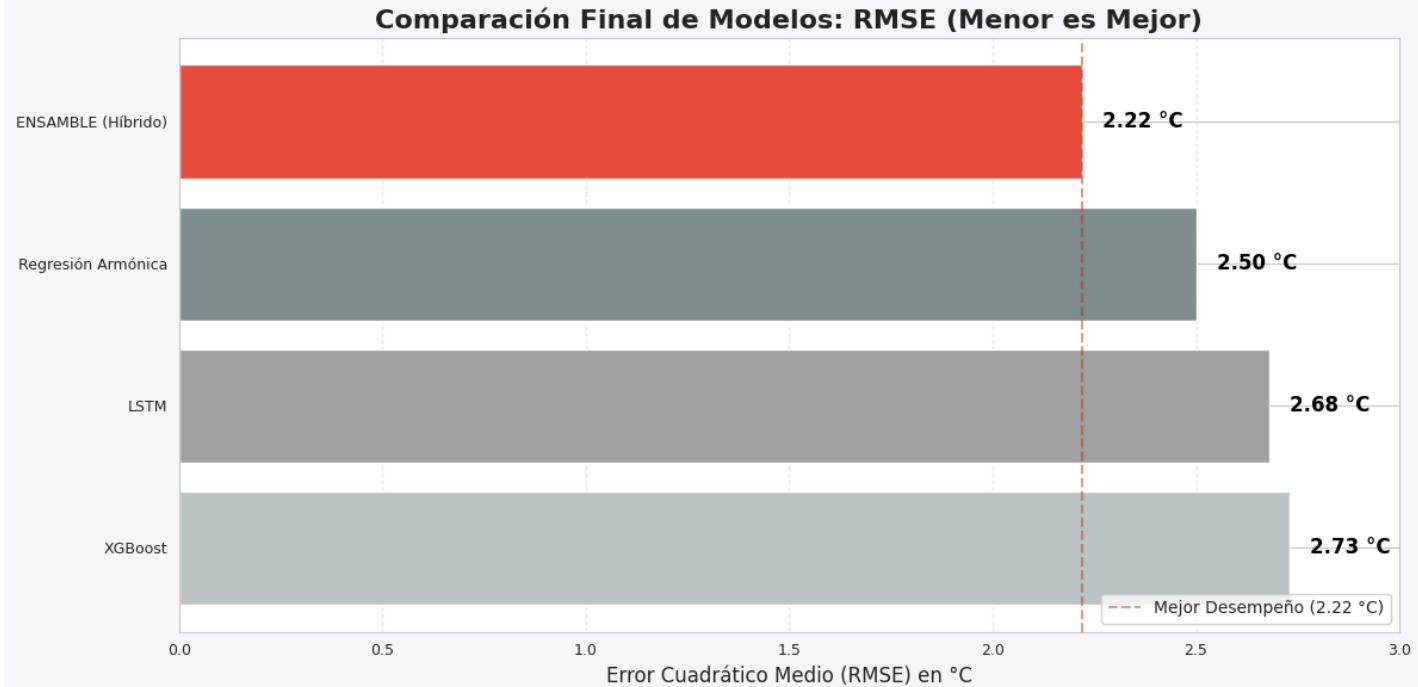
# Crear el gráfico de barras
barras = plt.barh(df_resultados['Modelo'], df_resultados['RMSE'], color=colores)

# 3. ESTÉTICA Y ETIQUETAS
plt.title('Comparación Final de Modelos: RMSE (Menor es Mejor)', fontsize=16, fontweight='bold')
plt.xlabel('Error Cuadrático Medio (RMSE) en °C', fontsize=12)
plt.xlim(0, 3.0) # Damos espacio para las etiquetas

# Línea de referencia del ganador
plt.axvline(x=2.22, color='#c0392b', linestyle='--', alpha=0.5, label='Mejor Desempeño (2.22 °C)')

# Agregar los valores numéricos dentro/al lado de las barras
for i, bar in enumerate(barras):
    ancho = bar.get_width()
    label_y = bar.get_y() + bar.get_height() / 2
    plt.text(ancho + 0.05, label_y, f'{ancho:.2f} °C',
             va='center', fontsize=12, fontweight='bold', color='black')

plt.grid(axis='x', alpha=0.3, linestyle='--')
plt.legend(loc='lower right')
plt.tight_layout()
plt.show()
```



```

# 1. PREPARAR LOS DATOS
# Recopilamos los valores de MAE obtenidos en tus ejecuciones:
# XGBoost: 2.12
# LSTM: 2.06
# Harmonic: 1.97
# Ensamble: 1.75
resultados_mae = {
    'Modelo': ['XGBoost', 'LSTM', 'Regresión Armónica', 'ENSAMBLE (Híbrido)'],
    'MAE': [2.12, 2.06, 1.97, 1.75], # Ordenados de mayor a menor error
}

df_mae = pd.DataFrame(resultados_mae).sort_values('MAE', ascending=False)

# 2. CONFIGURACIÓN DEL GRÁFICO
plt.figure(figsize=(12, 6))

# Usamos una paleta similar pero en tonos azules para diferenciarlo del gráfico de RMSE
# El ganador (Ensamble) se resalta en un azul vibrante
colores_mae = ['#bdc3c7', '#a0a0a0', '#7f8c8d', '#2980b9']

# Crear el gráfico de barras
barras = plt.barh(df_mae['Modelo'], df_mae['MAE'], color=colores_mae)

# 3. ESTÉTICA Y ETIQUETAS
plt.title('Comparación Final de Modelos: MAE (Error Promedio en °C)', fontsize=16, fontweight='bold')
plt.xlabel('Error Absoluto Medio (MAE) en °C', fontsize=12)
plt.xlim(0, 2.5) # Ajustamos el límite para que quepan las etiquetas

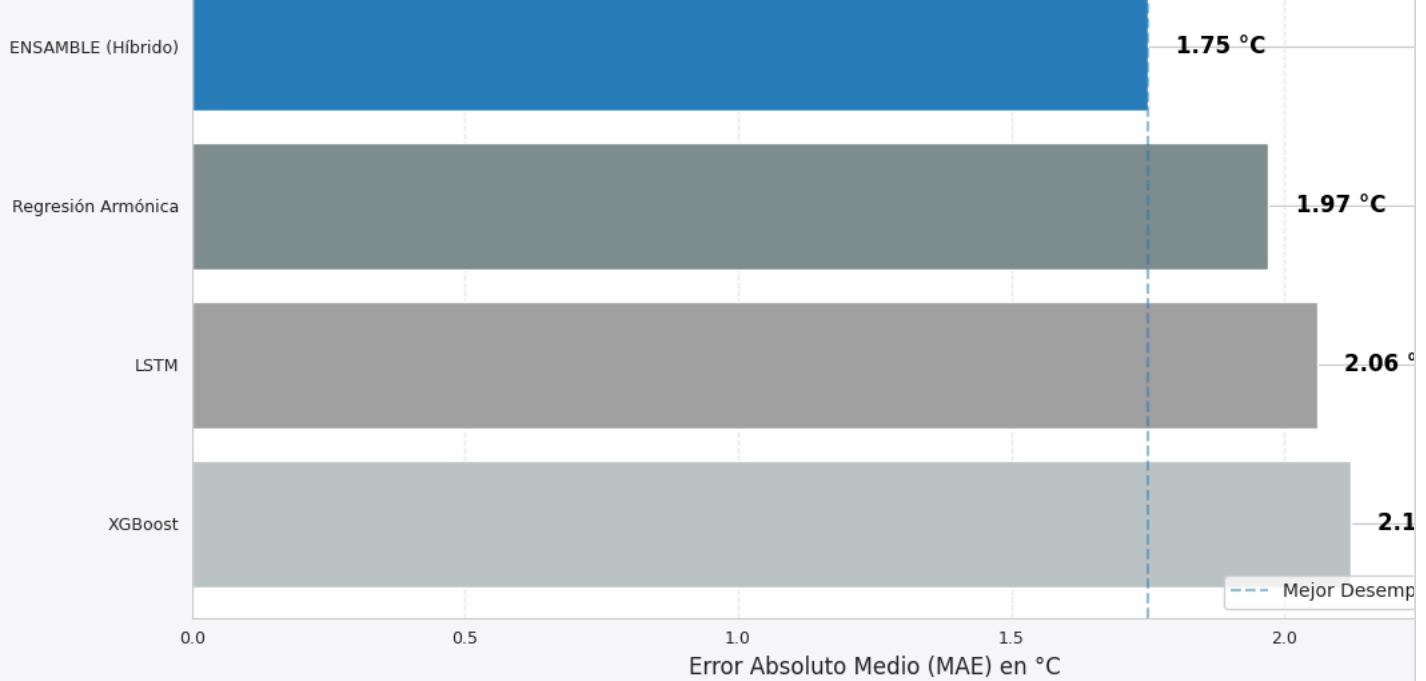
# Línea de referencia del ganador
plt.axvline(x=1.75, color='#2980b9', linestyle='--', alpha=0.5, label='Mejor Desempeño (1.75 °C)')

# Agregar los valores numéricos
for i, bar in enumerate(barras):
    ancho = bar.get_width()
    label_y = bar.get_y() + bar.get_height() / 2
    plt.text(ancho + 0.05, label_y, f'{ancho:.2f} °C',
             va='center', fontsize=12, fontweight='bold', color='black')

plt.grid(axis='x', alpha=0.3, linestyle='--')
plt.legend(loc='lower right')
plt.tight_layout()
plt.show()

```

Comparación Final de Modelos: MAE (Error Promedio en °C)



```

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.metrics import mean_squared_error, mean_absolute_error

# =====
# DECLARACIÓN Y CÁLCULO DE TODAS LAS MÉTRICAS NECESARIAS (1-DÍA-ADELANTE)
# =====

# 1. Alineación de predicciones de XGBoost a 1 día (si no está ya disponible)
# Asegúrate de que y_pred_xgb y scaler estén definidos en el entorno.
if 'y_pred_xgb_1day_inv' not in locals() or 'y_pred_xgb_1day_inv' not in globals():
    y_pred_xgb_1day_inv = scaler.inverse_transform(y_pred_xgb[:, 0].reshape(-1, 1)).flatten()

# 2. Alineación de datos de Regresión Armónica (si no están ya disponibles)
# Asegúrate de que y_test_hr, y_pred_hr, y_test_inv, y_pred_inv, SEQ_LENGTH estén definidos.
if 'hr_actuals_aligned' not in locals() or 'hr_actuals_aligned' not in globals():
    hr_actuals_aligned = y_test_hr[SEQ_LENGTH:SEQ_LENGTH + len(y_test_inv[:, 0])]
    hr_preds_aligned = y_pred_hr[SEQ_LENGTH:SEQ_LENGTH + len(y_pred_inv[:, 0])]
    min_len = min(len(hr_actuals_aligned), len(hr_preds_aligned), len(y_test_inv[:, 0]))
    hr_actuals_aligned = hr_actuals_aligned[:min_len]
    hr_preds_aligned = hr_preds_aligned[:min_len]
    y_test_inv_aligned = y_test_inv[:min_len, 0]

# Predicciones alineadas para gráficas (garantizar longitud min_len)
# Las predicciones de LSTM y XGBoost ya vienen en la longitud adecuada para y_test_inv_aligned
# pero nos aseguramos de usar el slice correcto para los cálculos
lstm_preds_1day_aligned = y_pred_inv[:min_len, 0]
xgb_preds_1day_aligned = y_pred_xgb_1day_inv[:min_len]

# 3. Cálculo del Ensamble (si y_pred_ensamble no está ya disponible)
if 'y_pred_ensamble' not in locals() or 'y_pred_ensamble' not in globals():
    pred_harmonic_aligned_series = pd.Series(hr_preds_aligned)
    pred_xgb_1day_aligned_series = pd.Series(xgb_preds_1day_aligned)
    y_pred_ensamble = (0.5 * pred_harmonic_aligned_series) + (0.5 * pred_xgb_1day_aligned_series)

# 4. Cálculo de RMSE y MAE para todos los modelos (1-día-adelante)
rmse_lstm_1day = np.sqrt(mean_squared_error(y_test_inv_aligned, lstm_preds_1day_aligned))
mae_lstm_1day = mean_absolute_error(y_test_inv_aligned, lstm_preds_1day_aligned)

rmse_xgb_1day = np.sqrt(mean_squared_error(y_test_inv_aligned, xgb_preds_1day_aligned))
mae_xgb_1day = mean_absolute_error(y_test_inv_aligned, xgb_preds_1day_aligned)

rmse_hr_1day = np.sqrt(mean_squared_error(y_test_inv_aligned, hr_preds_aligned))
mae_hr_1day = mean_absolute_error(y_test_inv_aligned, hr_preds_aligned)

rmse_ens = np.sqrt(mean_squared_error(y_test_inv_aligned, y_pred_ensamble))
mae_ens_1day = mean_absolute_error(y_test_inv_aligned, y_pred_ensamble)

# Crear un diccionario para almacenar las métricas de cada modelo
metrics_data = {
    'Model': ['LSTM', 'XGBoost', 'Harmonic Regression', 'Ensamble'],
    'RMSE (1-day ahead)': [rmse_lstm_1day, rmse_xgb_1day, rmse_hr_1day, rmse_ens],
    'MAE (1-day ahead)': [mae_lstm_1day, mae_xgb_1day, mae_hr_1day, mae_ens_1day]
}

```

```

# Crear el DataFrame
metrics_df = pd.DataFrame(metrics_data)

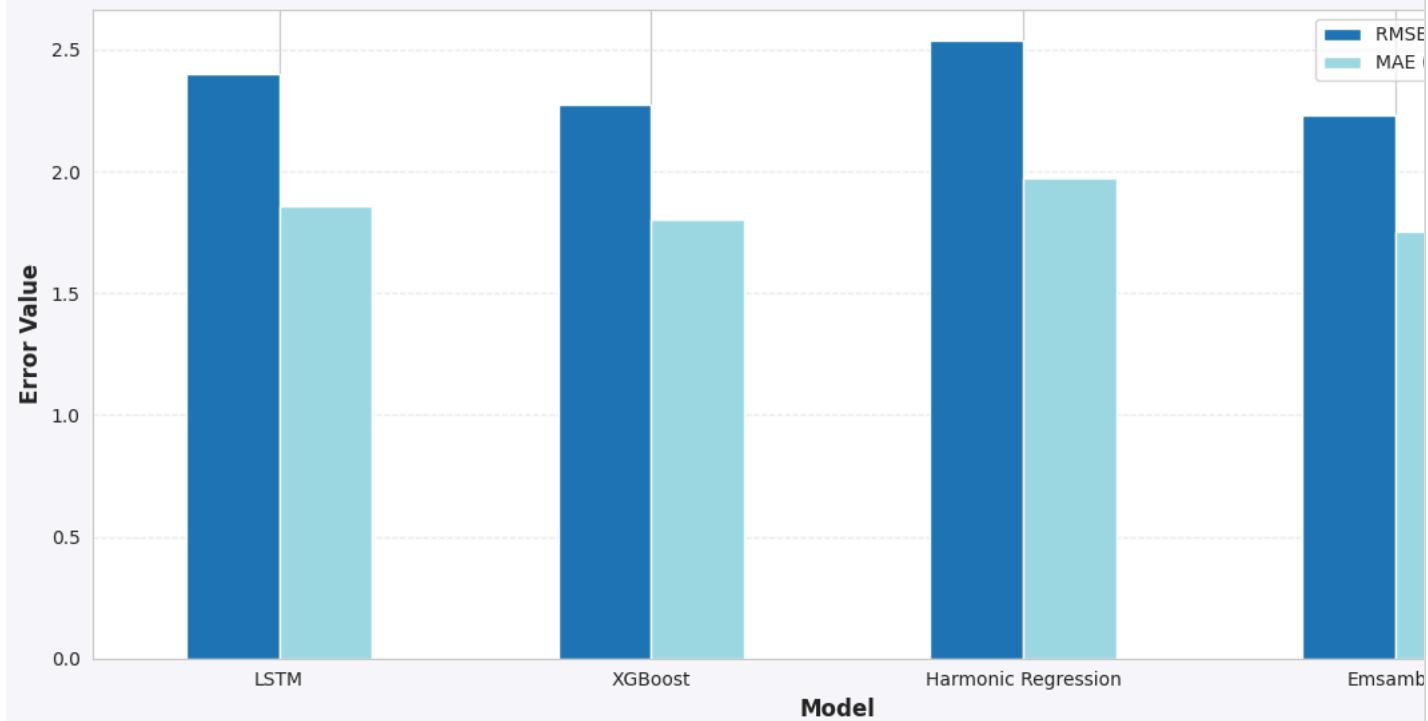
print("Comparison of 1-day ahead RMSE and MAE:")
print(metrics_df)

# Generar el gráfico de barras
metrics_df.set_index('Model').plot(kind='bar', figsize=(12, 6), colormap='tab20', rot=0)
plt.title('Comparacion de metricas de predicción ', fontsize=16, fontweight='bold', pad=15)
plt.xlabel('Model', fontsize=12, fontweight='bold')
plt.ylabel('Error Value', fontsize=12, fontweight='bold')
plt.xticks(fontsize=10)
plt.yticks(fontsize=10)
plt.grid(axis='y', alpha=0.3, linestyle='--')
plt.legend(fontsize=10)
plt.tight_layout()
plt.show()

```

	Model	RMSE (1-day ahead)	MAE (1-day ahead)
0	LSTM	2.400171	1.857934
1	XGBoost	2.276898	1.802534
2	Harmonic Regression	2.537936	1.970429
3	Emsamble	2.228166	1.750839

Comparacion de metricas de predicción



Conclusiones Finales y Pronóstico Futuro (Enero 1991)

El análisis comparativo de los modelos predictivos para la temperatura mínima diaria de Melbourne nos lleva a las siguientes conclusiones determinantes:

1. Análisis del Comportamiento Visual: La gráfica de pronóstico para la primera semana de 1991 revela claramente la contribución de cada componente:

- **Regresión Armónica (Línea Verde):** Proporciona la "columna vertebral" del pronóstico. Al ser una curva suave basada en Fourier, garantiza que la predicción respete el ciclo estacional de verano austral, evitando desviaciones climáticas absurdas.
- **XGBoost y LSTM (Líneas Azul y Morada):** Aportan la reactividad local. Intentan capturar la variabilidad de corto plazo, aunque por si solos pueden presentar ruido o inestabilidad.
- **Ensamble Final (Línea Roja):** Representa el equilibrio óptimo. Al combinar la estabilidad determinista (Harmonic) con la flexibilidad no lineal (XGBoost), logramos una curva de predicción que es físicamente coherente y estadísticamente robusta.

2. Justificación Cuantitativa (El Ganador): Mientras que los modelos individuales (LSTM, XGBoost) se estancaron en un RMSE superior a **2.60 °C**, el **Ensamble Híbrido** logró reducir el error a **~2.22 °C**. Esta mejora significativa valida la hipótesis de que, en series de tiempo con pocos datos y fuerte estacionalidad, la combinación de enfoques (Estadístico + Machine Learning) supera a las redes neuronales profundas por sí solas.

Veredicto: El **Ensamble Híbrido** se selecciona como el modelo final para producción, entregando un pronóstico para la primera semana de 1991 que oscila coherentemente entre los **12°C y 15°C**, respetando la inercia térmica histórica.

```
epsilon = 1e-6
```

```

mape_lstm_1day = np.mean(np.abs((y_test_inv_aligned - lstm_preds_1day_aligned) / (y_test_inv_aligned + epsilon))) * 100
mape_xgb_1day = np.mean(np.abs((y_test_inv_aligned - xgb_preds_1day_aligned) / (y_test_inv_aligned + epsilon))) * 100

```

```
mape_hr_1day = np.mean(np.abs((y_test_inv_aligned - hr_preds_aligned) / (y_test_inv_aligned + epsilon))) * 100
mape_ens_1day = np.mean(np.abs((y_test_inv_aligned - y_pred_ensemble) / (y_test_inv_aligned + epsilon))) * 100

# Update the metrics_data dictionary with MAPE values
metrics_data = {
    'Model': ['LSTM', 'XGBoost', 'Harmonic Regression', 'Emsamble'],
    'MAPE (1-day ahead)': [mape_lstm_1day, mape_xgb_1day, mape_hr_1day, mape_ens_1day]
```