Universidad de Oviedo
Universidá d'Uviéu
University of Oviedo

Máster Universitario en Análisis de Datos para Inteligencia de Negocios

Trabajo de Fin de Máster

Estudio comparativo de modelos clásicos de series temporales y métodos de Machine Learning para la predicción de la temperatura diaria de Gijón

Rodrigo Álvarez Fernández

Tutoras
Agustina Bouchet
Irene Mariñas del Collado

ÍNDICE

Índice de figuras

Índice de tablas

Declaración de originalidad

De acuerdo con lo expresado en el *artículo 8.3 del Reglamento para la elaboración y defensa del Trabajo Fin de Máster de la Universidad de Oviedo*, aprobado por su Consejo de Gobierno el 17 de julio de 2020 (BOPA de 7 de agosto de 2020), quiero expresar lo siguiente:

Yo, **RODRIGO ALVAREZ FERNANDEZ**, con DNI en relación a la memoria que presento ante el Tribunal, para su valoración como ***Trabajo Final en el Máster Universitario en Análisis de Datos para la Inteligencia de Negocios (MANADINE),*** quiero ***DECLARAR*** que ***s***oy el autor de la misma, habiendo citado debidamente las fuentes utilizadas en su desarrollo.

Para que conste, firmo el presente documento.

Oviedo, 17 de julio de 2023

Fdo.-

Resumen

Las series temporales son secuencias de datos organizados en orden cronológico y están muy presentes en nuestro día a día. Este trabajo se centra en el análisis de series temporales con un enfoque específico en el estudio de la temperatura media diaria en Gijón. Se presentan los conceptos fundamentales de las series temporales y se abordan los pasos de un preprocesamiento de una base de datos, incluyendo estrategias para tratar con datos faltantes. Además, se introducen y aplican varios modelos y métodos para el análisis y predicción de este tipo de datos, como los modelos de suavizado exponencial y SARIMA y los métodos de *Seasonal Naive*, KNN y SVM. Mediante el uso de estos modelos y métodos, se busca comprender y predecir los patrones y tendencias de la temperatura en Gijón, además de comparar la capacidad predictora de estos, proporcionando información valiosa para futuras investigaciones.

Abstract

Time series are sequences of data organized in chronological order and are very present in our daily lives. This work focuses on the analysis of time series with a specific focus on the study of the mean daily temperature in Gijón. The fundamental concepts of time series are presented, and the steps of preprocessing a database are used, including strategies for dealing with missing data. Additionally, various models and methods for the analysis and prediction of this type of data are introduced and applied, such as exponential smoothing models and SARIMA, as well as Seasonal Naive, KNN, and SVM methods. By using these models and methods, the aim is to understand and predict temperature patterns and trends in Gijón, as well as to compare the predictive ability of these models, providing valuable information for future research.

Capítulo 1

Introducción

En la actualidad, con el creciente nivel de digitalización, se dispone de una amplia cantidad de datos que permiten extraer conclusiones de manera cada vez más exhaustiva. Un enfoque relacionado con el tratamiento de datos es la predicción de valores futuros, como pueden ser las ventas mensuales que va a tener una empresa, qué productos pueden ser recomendados para un usuario en una página web, o los valores de mercado que se podrían tener en las siguientes semanas.

Dentro de la predicción de datos existe un campo que es el de las series temporales, en el cual se enfoca el presente trabajo. Este tipo de estructuras de datos son conjuntos de observaciones recogidas a intervalos regulares de tiempo. El tamaño de dicho intervalo se conoce como la frecuencia, la cual puede variar y, dependiendo de esta se puede estar hablando de series de datos trimestrales, anuales, semanales, etc. Es esta estructura de las series temporales la que permite el uso de una amplia variedad de estrategias para predecir valores futuros [1].

La capacidad de predecir el comportamiento de una variable haciendo uso de su pasado es lo que hace que las series temporales sean ampliamente útiles en diversas situaciones. Estos múltiples usos incluyen, por ejemplo, la aplicación por parte de una empresa para predecir el nivel de ingresos netos en un periodo de tiempo [2]. Otro campo es la salud pública, donde las organizaciones utilizan estas aplicaciones para predecir la propagación de enfermedades y planificar respuestas efectivas en casos de brotes de enfermedades [3]. O la climatología, donde los científicos pueden predecir el clima futuro y las tendencias climáticas a largo plazo [4]. Existen numerosos campos de aplicación en esta disciplina que abarcan una amplia gama de áreas, lo que demuestra no solo su versatilidad, sino también su gran relevancia.

Una aplicación con mucha utilidad es la predicción del tiempo atmosférico diario, siendo este un problema importante en la industria meteorológica. Existen diversas técnicas para predecir variables climáticas, como modelos de series temporales, modelos numéricos de predicción o métodos de *Machine Learning*. El análisis de estas variables utilizando dichas técnicas, puede proporcionar información valiosa sobre las estaciones, las tendencias o los patrones en los datos climáticos, lo que permite predecir las

condiciones meteorológicas a futuro [5]. El estudio de las diferentes técnicas que existen resulta de gran relevancia puesto que puede ayudar a proporcionar predicciones precisas y fiables del tiempo futuro. Estas predicciones no solo son importantes tanto a nivel de usuario sino también de sociedad, ayudando con problemas de carácter de planificación urbana, industria agrícola, seguridad de las personas o incluso investigación del tiempo atmosférico. Cabe destacar que predecir estos datos es todo un desafío, debido a la complejidad de los datos climáticos, variabilidad temporal y la interacción de múltiples otras variables [6].

Este trabajo se centra en la teoría de series temporales e intenta comparar la capacidad predictora de distintos modelos y métodos existentes para predecir la temperatura media diaria de la ciudad de Gijón. Para abordar la predicción de esta variable, se han considerado diversas metodologías teniendo en cuenta los conocimientos a priori sobre las características de los datos (estacionalidad, frecuencia diaria, tendencia, entre otros). A continuación, se indican las metodologías seleccionadas junto con una motivación de por qué se escogieron:

1. *Seasonal Naive*: Este enfoque es simple y rápido de implementar, lo cual puede ser adecuado cuando se necesita una predicción rápida sin utilizar métodos más complejos. Este método asume que el valor futuro será igual al valor observado en la misma temporada del año anterior. Si la temperatura diaria tiene una fuerte estacionalidad y sigue patrones similares año tras año, este enfoque puede proporcionar resultados satisfactorios. Un ejemplo ilustrativo se encuentra en [7], donde se predicen las lluvias en India utilizando *Seasonal Naive* junto con otras metodologías.

2. Suavizado Exponencial: El suavizado exponencial es una técnica ampliamente utilizada para pronosticar series temporales. El suavizado exponencial de Holt-Winters, es especialmente adecuado cuando hay tendencia y estacionalidad. Si la temperatura diaria muestra patrones a largo plazo o cambios graduales, este puede capturar y proyectar adecuadamente esas características. En [8] se ejemplifica el uso de suavizado exponencial para el modelado y predicción de parámetros meteorológicos en Pakistán.

3. Modelo estacional autorregresivo integrado de media móvil (SARIMA, de sus siglas en inglés, *Seasonal Autoregressive Integrated Moving Average)*: El modelo SARIMA es una opción robusta cuando se trata de series temporales con

componente estacional. Si la temperatura diaria exhibe una estacionalidad clara (variaciones regulares a lo largo del año) el modelo SARIMA es capaz de capturar y modelar esas variaciones, así como las tendencias y los patrones autorregresivos en los datos. Como ejemplo, en [9], se modelan series temporales utilizando SARIMA para precipitaciones y temperaturas mensuales en los países del sur de Asia.

4. Método de K vecinos cercanos (KNN, de sus siglas en inglés, *K Nearest Neighbours*) y Máquinas de vector soporte (SVM, de sus siglas en inglés, *Support Vector Machines*): Son modelos de aprendizaje automático que son útiles cuando la serie temporal está influenciada por múltiples variables o presenta patrones no lineales como es el caso de la temperatura. KNN puede encontrar patrones similares en los datos históricos y utilizarlos para hacer predicciones, mientras que SVM es eficaz para trazar relaciones complejas entre variables. Son opciones muy adecuadas si se considera otro tipo de metodologías más flexibles que puedan adaptarse a relaciones complejas. Ejemplos de estas metodologías se pueden encontrar en [10], donde se predicen series temporales meteorológicas usando SVM; o en [11], donde utilizan KNN para simular precipitaciones diarias y otras variables climáticas.

Objetivos

El objetivo principal de este trabajo es realizar una comparativa entre modelos clásicos de predicción de series temporales y métodos de *Machine Learning* con el fin de encontrar la mejor manera de predecir la temperatura media diaria de la ciudad de Gijón. Dentro de estas metodologías, el estudio se ha centrado en, por un lado, el método de *Seasonal Naive* y los modelos de suavizado exponencial de Holt-Winters y SARIMA; y por otro lado los métodos de *Machine Learning* KNN y SVM. Después de introducirlos de manera teórica, se procede a ajustar los modelos o entrenar los métodos para luego realizar las predicciones. Finalmente, se procede a evaluar las diferentes opciones descritas, discutir los distintos resultados obtenidos y realizar predicciones con el que mejor desempeño muestra.

El trabajo se estructura de la siguiente manera:

1. En primer lugar, se realiza una revisión bibliográfica sobre la teoría de series temporales incluyendo los métodos clásicos de predicción y los modelos de *Machine Learning*.
2. Se presenta la base de datos climatológicos diarios, recopilada de la Agencia Estatal de Meteorología (AEMET), de la ciudad de Gijón, desarrollando cómo se lleva a cabo la limpieza y preprocesamiento de este tipo de datos para una mayor calidad de estos (con las variables más relevantes temperatura, insolación, precipitación y presión atmosférica). Se realiza además un primer estudio descriptivo de esta base de datos con las variables ya tratadas.
3. Se presenta un análisis de la serie temporal de la variable temperatura media utilizando las técnicas de análisis de series temporales para la identificación de sus componentes, estructura y, en base a ello, escoger los modelos apropiados.
4. Se entrenan los métodos y se ajustan los modelos. Después, se presentan los resultados obtenidos de sus predicciones. Además, se evalúa el rendimiento de estos utilizando distintas métricas de error comparando los resultados obtenidos y clasificando cuál de los métodos o modelos comete menos error según el criterio de cada métrica.
5. Se analizan los resultados extrayendo conclusiones sobre estos en base al funcionamiento de cada uno de los modelos y se selecciona cuál es la opción más adecuada para predecir la temperatura media en Gijón.
6. Finalmente, como muestra de aplicación práctica de la comparativa de modelos, se realizaron predicciones para los siguientes meses utilizando el modelo seleccionado.

Capítulo 2

Marco teórico

2.1. Definición de serie temporal

Para comenzar, se empezará definiendo el concepto de serie temporal y sus propiedades básicas.

Una serie temporal es el resultado de observar los valores de una variable a lo largo del tiempo en intervalos temporales regulares. Dichas variables son denotadas como $\{X_t\}$, donde $t \in \mathbb{Z}^+$. Esto se traduce en que una serie temporal es una colección de observaciones de una variable realizadas de forma secuencial en el tiempo, en las que el orden es importante. Los valores de una serie temporal van asociados a instantes de tiempo, de manera que el análisis de esta será el de un conjunto de dos variables, la propia variable de estudio y la variable tiempo que actúa como un índice. Si estos valores están distribuidos en intervalos de tiempo regulares, esto aporta a la serie lo que se denomina periodicidad, que depende de la frecuencia en la que están recogidas estas observaciones, esta puede ser anual, semestral, semanal, diaria, etc. [12].

El hecho de que las series temporales sean variables observadas en intervalos equidistantes en el tiempo está muy presente en el tratamiento práctico de los datos, por ejemplo, a la hora de manejar datos ausentes o datos que no son válidos. En el caso de que una serie temporal contenga datos desconocidos (o faltantes), se debe considerar emplear algún procedimiento para estimar dichos datos, tal y como se realizará posteriormente en este trabajo. En ningún caso habría que obviarlos dado que, en ese caso, se perdería la estructura temporal del conjunto de datos [13].

Las técnicas estadísticas utilizadas en el análisis de series temporales permiten, además de analizar y modelar el comportamiento de un fenómeno que cambia a lo largo del tiempo, hacer predicciones sobre los valores futuros de este. Dicho análisis tendría como objetivo extraer regularidades que caracterizan el comportamiento pasado de la variable, lo que sería conocer el mecanismo que la genera para poder tener un mejor conocimiento de esta. Es este mecanismo y el supuesto de que las características observadas permanecen constantes en el tiempo, lo que permitiría realizar esas predicciones. En otras palabras, cómo será su comportamiento en el futuro de una variable basándose en su pasado.

2.2. Componentes de una serie temporal

El primer paso a la hora de estudiar una serie temporal es hacer un análisis visual de esta. Las series temporales se representan generalmente en un gráfico de secuencia como el de la Figura 1, donde se puede ver representada la producción total trimestral de cerveza en Australia (en megalitros) desde 1956 (primer trimestre) a 2010 (segundo trimestre) (los datos utilizados se obtuvieron de [14]). En estos, se presentan las distintas observaciones a lo largo del tiempo unidas mediante líneas, estando el tiempo representado en el eje de abscisas y la variable que evoluciona en el eje de ordenadas.

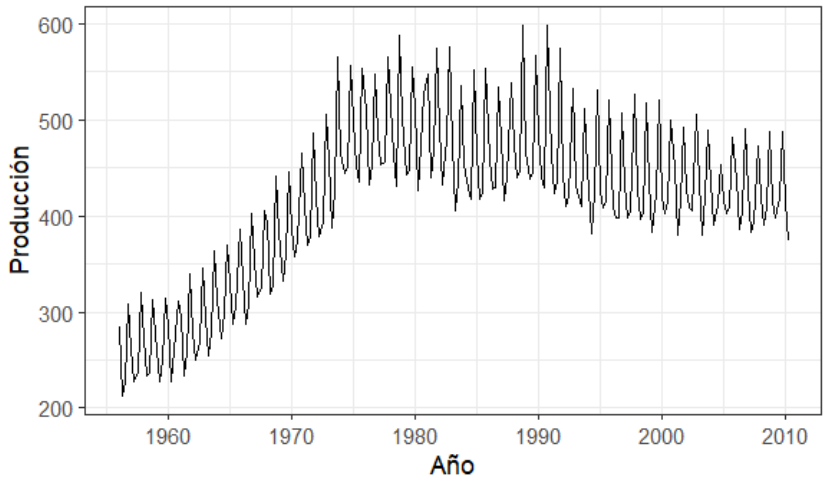

Figura 1. Ejemplo de gráfico de secuencia.

Además del gráfico secuencial, para analizar la serie temporal, se descompone en las componentes descritas a continuación.

2.2.1. Tendencia

La tendencia de una serie temporal se refiere a la dirección en la que se dirigen los datos a medida que avanza el tiempo, es decir, si los valores de la serie temporal están aumentando, disminuyendo o permanecen constantes a lo largo del tiempo (en este último caso se diría que no hay tendencia) [15]. Dos ejemplos de distintas direcciones de tendencia se pueden observar en la Figura 2, donde, en el gráfico de la izquierda, se encuentra representado el PIB per cápita de Australia trimestral desde el primer trimestre de 1971 al primero de 1998 (tendencia ascendente) y, en el de la derecha, las medidas anuales del nivel en pies del lago Huron desde 1875 a 1972 (tendencia descendente) (datos obtenidos de [14]).

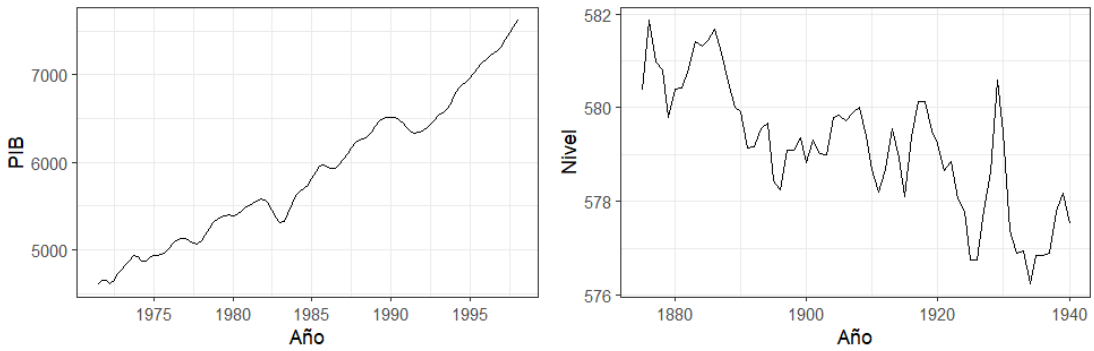

Figura 2. Ejemplo de dos series con tendencia ascendente (derecha) y descendiente (izquierda).

La tendencia puede ser local, que se refiere a los patrones o cambios de corto plazo que ocurren dentro de la serie; o global, que se refiere a la dirección general o patrón de largo plazo en una serie temporal. En una misma serie, es posible observar ambas formas de tendencia. Un ejemplo de esto lo podemos observar en los datos históricos del lago Huron de la Figura 2, donde podemos observar una tendencia global descendiente, pero donde se pueden también observar ciertos tendencias locales ascendentes y descendientes en determinados momentos. Además, las tendencias pueden ser lineales o no lineales. Las tendencias lineales se caracterizan por ser incrementos o decrecimientos aditivos (positivos o negativos respectivamente) en el nivel de la serie mientras que, por otro lado, las tendencias no lineales, suele ser multiplicativas [12].

2.2.2. Estacionalidad

Se puede decir que una serie es estacional cuando en ella observamos un patrón sistemático de fluctuación que se repite periódicamente a lo largo del tiempo y en el mismo momento. En muchas ocasiones, si una serie tiene un comportamiento estacional muy claro, se puede detectar a simple vista al representarla [16]. Es el caso de la serie temporal de la Figura 3, donde están representadas las noches realizadas por turistas internacionales trimestrales (en millones) en Australia desde 1999-2015 (datos obtenidos de [14]). En el gráfico se puede ver como en los meses cercanos al cambio de año los turistas se hospedan más noches (debido al verano en el hemisferio sur).

La identificación de patrones estacionales en una serie temporal es importante para comprender los factores que influyen en el comportamiento de la variable en cuestión, y puede ser útil para realizar predicciones.

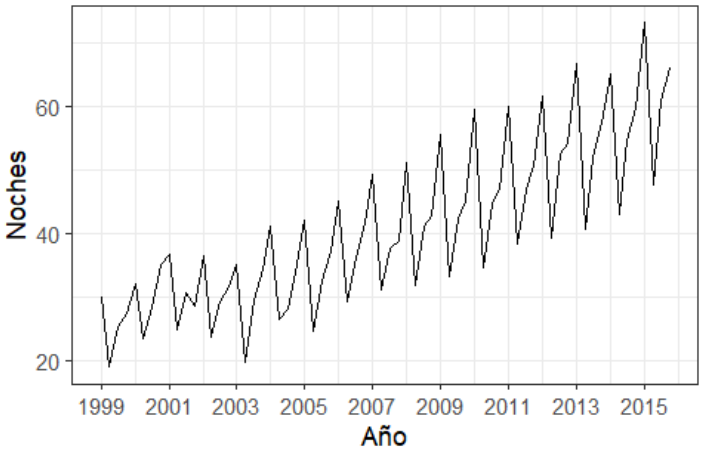

Figura 3. Ejemplo de serie temporal con estacionalidad.

2.2.3. Ciclos

Otro tipo de patrones repetitivos que podemos encontrar, que habría que diferenciarlos con la estacionalidad, son los ciclos. Los ciclos en una serie temporal se refieren a patrones recurrentes y periódicos que se observan a lo largo del tiempo. Pueden tener una duración variable y suelen repetirse en intervalos más o menos regulares. La diferencia principal entre ciclos y estacionalidad radica en la duración y la naturaleza de los patrones. Los ciclos son más largos y pueden variar en duración, mientras que la estacionalidad se repite de manera consistente dentro de un período de tiempo más corto y predefinido [13]. Un ejemplo de este tipo de patrones se encuentra en la Figura 4, que representa el número de linces anuales atrapados en la región de Mackenzie, Canadá, durante el periodo de 1821 a 1934 (datos obtenidos de [14]). En la gráfica se observan los diferentes ciclos, oscilaciones de aproximadamente 10 años, los cuales se producen debido a factores como la disponibilidad de presas o la competencia intraespecífica.

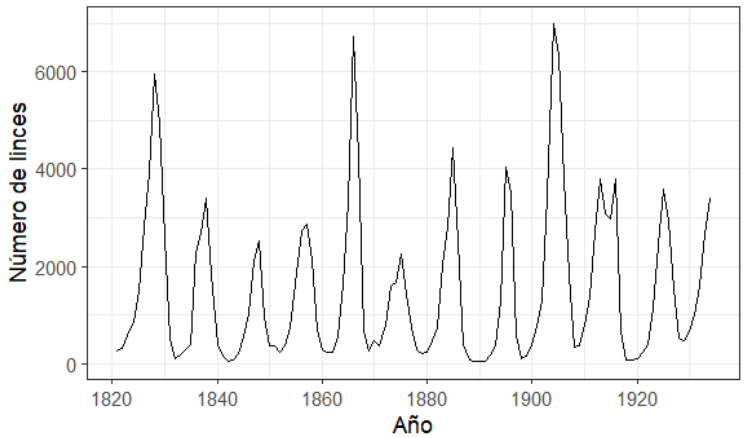

Figura 4. Ejemplo de serie temporal con patrones cíclicos.

2.2.4. Estacionariedad

Una serie temporal se considera estacionaria si sus propiedades estadísticas no cambian con el tiempo. En otras palabras, una serie temporal es estacionaria si su media y varianza son constantes a lo largo del tiempo. Esto implica entonces que, las series con tendencia y/o estacionalidad, no son estacionarias. Se pueden ver como ejemplos las representaciones incluidas en la Figura 5. El gráfico de la izquierda, muestra los totales mensuales en miles de pasajeros de aerolíneas internacionales de 1949 a 1960 (no estacionaria) y, el de la derecha, las mediciones anuales en millones de metros cúbicos del caudal del rio Nilo en Asuán entre 1890 y 1970 (estacionaria) (datos obtenidos de [14]).

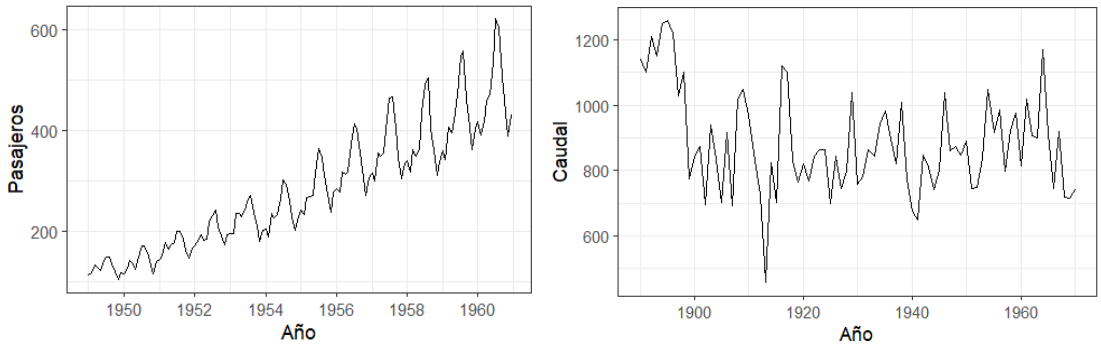

Figura 5. Ejemplos de una serie no estacionaria (izquierda) y de otra estacionaria (derecha).

Es importante destacar que la estacionariedad es una propiedad deseable en el análisis de series temporales, ya que algunos modelos y métodos de predicción clásicos requieren que la serie sea estacionaria para poder aplicarse correctamente. Sin embargo, en otros casos, su metodología requiere de estos patrones que hacen que la serie no sea estacionaria para un funcionamiento, como es el caso de métodos como *Seasonal Naive* y Holt-Winters, los cuales dependen de la estacionalidad y la tendencia. Por otro lado, otros métodos incorporan componentes que transforman la serie para su análisis, como es el caso de SARIMA [15].

2.2.5. Componente aleatoria o error

En el análisis de series temporales, es común encontrar también una componente aleatoria o error que no puede ser explicada por las tendencias, ciclos o estacionalidad presentes en los datos. Esta componente aleatoria, también conocida como residuo o error,

representa las variaciones no sistemáticas o impredecibles que no pueden ser atribuidas a ninguna otra estructura identificable en la serie temporal.

La componente aleatoria surge debido a diversas razones, como la presencia de factores no medidos o desconocidos, influencias aleatorias externas, errores de medición o limitaciones en la capacidad de los modelos para capturar todos los factores relevantes. En muchos casos, el error se asume como una variable aleatoria independiente e idénticamente distribuida, lo que implica que los valores del error no están correlacionados y tienen la misma distribución de probabilidad en todo el dominio temporal [15].

2.2.6. Funciones de autocorrelación y autocorrelación parcial

A la hora de realizar un análisis de una serie temporal, también son relevantes las funciones de autocorrelación y de autocorrelación parcial.

La función de autocorrelación mide la dependencia de una serie temporal consigo misma a lo largo del tiempo. Aunque cabe esperar, debido a la naturaleza secuencial, que para retardos [1] pequeños esta correlación sea mayor que para *lags* más grandes, no debe tomarse como aplicable a todas las series temporales y se deberá estudiar.

Por otro lado, la autocorrelación parcial, estudia la relación que hay entre observaciones en distintos instantes no contiguos. Para calcular la autocorrelación parcial, se utiliza un modelo de regresión lineal. Este incluye los retardos intermedios como variables independientes para controlar su influencia en la relación entre los valores en distintos retardos. Al examinar el coeficiente de correlación parcial de un retardo específico, se puede obtener una medida de la relación directa entre ese retardo y la observación actual, eliminando el efecto indirecto de los retardos intermedios.

Para representar estas funciones, normalmente se utilizan los correlogramas. En estas gráficas se representa la autocorrelación o autocorrelación parcial entre los instantes de una serie en función de la distancia que los separa. La línea vertical indica el valor que toma la función representada. También se suelen incluir dos líneas horizontales discontinuas que encierran los valores para los que la autocorrelación no es estadísticamente significativa (normalmente a un nivel de 0.05) [16].

[1] Desplazamientos hacia atrás en el tiempo normalmente denominados *lags*, en inglés.

En la Figura 6, se muestran dos correlogramas, uno para la autocorrelación simple (izquierda) y otro para la autocorrelación parcial (derecha). En estos, en el eje de abscisas viene representado el lag o retardo y en el eje de ordenadas la autocorrelación (ACF, de sus siglas en ingles *Autocorrelation Function*) o autocorrelación parcial (PACF, de sus siglas en ingles *Partial Autocorrelation Function*). En el ACF existe un comportamiento ondulatorio con todos los valores positivos indicando una correlación positiva en múltiples retardos. Esto indica que la serie tiene una dependencia con sus valores pasados, sugiriendo, el patrón estacional de la serie temporal. Por otro lado, en el PACF, las líneas verticales que sobresalen el umbral de significancia indican correlaciones parciales significativas en esos retardos específicos. En otras palabras, indica que las correlaciones entre las observaciones en la serie temporal, considerando y eliminando la influencia de los retrasos anteriores, son estadísticamente significativas en los momentos señalados por las líneas verticales en el gráfico del PACF.

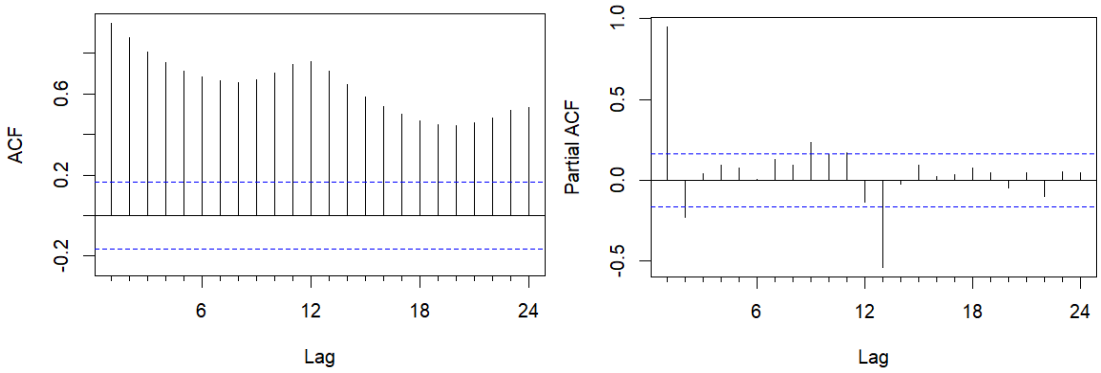

Figura 6. Ejemplo de correlogramas utilizando la serie de pasajeros de líneas internacionales.

2.3. Métodos y modelos de series temporales

Los métodos y los modelos son herramientas que permiten predecir valores futuros en base a los valores pasados de una serie temporal. La selección del método o modelo adecuado depende del comportamiento de la serie en cuestión, teniendo en cuenta sus componentes (tendencia, estacionalidad, patrones de autocorrelación). Estos métodos y modelos pueden ser todo lo complejos que se desee. Por lo general, a mayor complejidad, mayor exactitud a la hora de describir la evolución de una serie, aunque también puede llevar a un sobreajuste de los datos perdiendo la capacidad de extrapolación de los resultados. Para que un método o un modelo sea más manejable tanto teórica como computacionalmente, se suele preferir algo más simplificado, asumiendo un error razonable.

En esta sección, se presentan diferentes métodos y modelos utilizados en la predicción de series temporales.

2.3.1. Método de *Seasonal Naive*

La metodología *Naive*, también conocida como "ingenua" es una de las formas más simples y comunes para predecir series temporales. Se basa en la suposición de que el valor futuro de la serie será igual al último valor observado de la serie, por lo que, para realizar el pronóstico del siguiente periodo, simplemente estaría utilizando este último valor. El método Naive ha sido utilizado durante mucho tiempo como un punto de referencia básico para comparar la efectividad de otros modelos más sofisticados. Aunque puede parecer simple, en algunos casos, el método Naive puede proporcionar resultados bastante precisos, especialmente cuando la serie temporal no muestra patrones claros o cuando se trata de predecir un horizonte a corto plazo [17].

Basándose en esta metodología, se encuentra el método Naive estacional (conocido como *Seasonal Naive*, en inglés), este método también escoge un último valor para realizar la predicción, pero, al contrario que el Naive simple, no sería el último de la serie, si no el último valor correspondiente al periodo anterior. Es decir, si estamos pronosticando las ventas de un producto para el mes de diciembre de 2023, utilizaríamos el valor de las ventas del mes de diciembre de 2022 como pronóstico. De esta manera, se tiene en cuenta la estacionalidad de la serie, lo que es útil si se tiene series con patrones estacionales. En la Figura 7 se muestra un ejemplo de predicción mediante el método *Seasonal Naive,* donde se utilizan los datos mensuales tomados entre 1920 y 1940 de la temperatura del castillo de Nottingham en grados Fahrenheit (datos tomados de [14]). La serie temporal está representada en negro y las predicciones para los siguientes dos años en rojo, coincidiendo con los valores del último año.

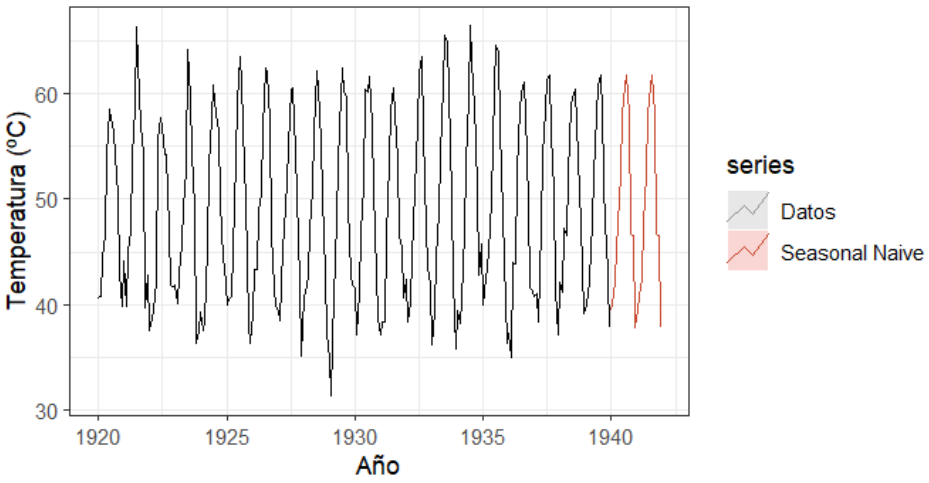

Figura 7. Predicción utilizando Seasonal Naive.

2.3.2. Modelos de suavizado exponencial

El suavizado exponencial se originó en el trabajo de Robert G. Brown a principios de la década de 1950. En 1956, Brown presentó su trabajo sobre el suavizado exponencial de las demandas de inventario en una conferencia de la *Operations Research Society of America*. Esta presentación formó la base del primer libro de Brown, *Pronósticos estadísticos para el control de inventario* [18]. Su segundo libro, *Smoothing, Forecasting, and Prediction of Discrete Time Series* [19], desarrolló la metodología general de suavización exponencial.

Durante la década de 1950, Charles C. Holt trabajó independientemente de Brown para desarrollar un método similar para suavizar exponencialmente las tendencias aditivas y un método completamente diferente para suavizar datos estacionales. El trabajo original de Holt fue documentado en un memorándum de la Oficina de Investigación Naval de Estados Unidos [20].

Los modelos de suavizado exponencial se basan en realizar los pronósticos producidos con promedios ponderados de las observaciones pasadas en una serie temporal. En estos modelos, se asignan pesos a cada observación pasada, los cuales decaen exponencialmente a medida que las observaciones envejecen. En otras palabras, cuanto más reciente sea la observación, mayor será el peso asociado proporcionando más información sobre la serie temporal que las observaciones antiguas. Al tener en cuenta estos pesos, se logra un equilibrio entre dar importancia a todas las observaciones de la

serie y considerar que las observaciones más recientes como las más relevantes [21]. Dentro de estos, existen varios modelos explicados a continuación.

2.3.2.1. Suavizado exponencial simple

El modelo de Suavizado Exponencial Simple (SES) es una técnica de pronóstico que se basa únicamente en la suma ponderada de los valores pasados para hacer predicciones. Este modelo es adecuado cuando los datos no presentan una tendencia clara ni tampoco una estacionalidad.

El modelo SES es considerado el método más sencillo de todos los modelos de suavizado exponencial. Al no tener en cuenta factores como la tendencia o la estacionalidad, es fácil de implementar y no requiere una gran cantidad de datos ni un procesamiento complejo. Sin embargo, su simplicidad también puede limitar en gran medida su capacidad para hacer predicciones de una serie temporal. Es por ello por lo que, es importante evaluar si el SES es adecuado para los datos que se quieren predecir [17].

2.3.2.2. Modelo de Holt

El modelo de Holt es una mejora del modelo SES. Mientras que el modelo SES solo tiene en cuenta los valores pasados, el modelo de Holt utiliza tanto los valores históricos como la información sobre la tendencia para generar pronósticos [20].

Al incluir un término para la tendencia, en este caso lineal, el modelo de Holt es capaz de adaptarse mejor a cambios graduales en la serie temporal en comparación con el SES. Esto le permite capturar las variaciones a largo plazo y generar pronósticos más precisos para series temporales con esta componente.

2.3.2.3. Modelo de Holt-Winters

El modelo de Holt-Winters es, a su vez, una mejora del modelo de Holt. Al igual que el modelo de Holt, el modelo de Holt-Winters incluye el término para captar la tendencia de la serie temporal (tanto aditiva como multiplicativa). Sin embargo, va un paso más allá, ya que Holt [20] y Winters [22] ampliaron el método de Holt para captar la estacionalidad.

Este modelo de suavizado exponencial es más adecuado para series que exhiben patrones de estacionalidad, tanto si es aditiva como multiplicativa. Al tener en cuenta estas dos componentes, se mejora la capacidad del modelo para adaptarse a la estructura de la serie, realizando pronósticos más precisos [21].

En la Figura 8 se muestran los resultados para la predicción a un año para la serie de datos mensuales de pasajeros de aerolíneas internacionales (encontrados en [14]). En esta figura se encuentra representada la serie temporal en negro y los resultados de utilizar los diferentes métodos de suavizado exponencial para hacer las predicciones. En rojo está representado el método de Holt, en verde el método de Holt-Winters y en azul el método SES.

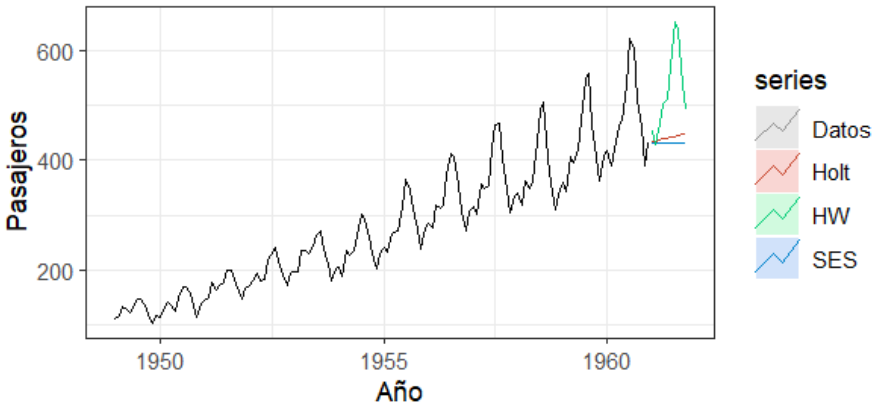

Figura 8. Predicciones para la serie de utilizando los métodos de Holt, Holt-Winters y suavizado exponencial simple.

2.3.3. Modelos SARIMA

El modelo ARIMA o método de Box-Jenkins fue introducido por Box et al [15]. Este método se centra en la autocorrelación entre las observaciones, describiendo cada valor como una función lineal de datos previos y errores debidos al azar, pudiendo incluir un componente cíclico o estacional. El acrónimo ARIMA significa media móvil integrada autorregresiva y es una generalización de un modelo de media móvil autorregresiva (ARMA). Para comprender los modelos SARIMA, conocidos así por las siglas del inglés, *Seasonal Autoregressive Integrated Moving Average*, es necesario entender sus partes por separado.

2.3.3.1. Modelos ARMA

Los modelos ARMA son una familia de modelos para series temporales estacionarias que provienen de la combinación de dos tipos de modelos más sencillos: los modelos autorregresivos (AR) y los modelos de media móvil (MA).

Los modelos AR (*AutoRegressive*) son una clase de modelos para series temporales estacionarias que utilizan términos autorregresivos para modelar la serie. En estos modelos, la observación en un momento dado se explica como una combinación lineal de observaciones previas (regresa a sí misma en el tiempo). A la hora de describir una serie mediante este tipo de modelos, se especifica un orden (p), que hace referencia al número de observaciones anteriores que este utilizarán [23].

Por otro lado, los modelos MA (*Moving Average*) utilizan la media móvil para modelar la dinámica de la serie. Se denomina así porque utilizan una media ponderada de los errores pasados para predecir los valores futuros. En otras palabras, la observación en un momento dado se explica como una combinación lineal de estos errores en lugar de las observaciones. Estos errores son los obtenidos al ajustar un modelo de regresión lineal a los datos observados de la serie temporal. Para este caso también se especifica un orden (q) que indica cuántos términos de media móvil se incluyen en el modelo [24].

2.3.3.2. Modelos ARIMA

Los modelos ARIMA, por otro lado, son adecuados para series temporales no estacionarias debido a que incluyen un término integrado (I) que toma la diferencia entre los valores observados de la serie temporal y los valores de la serie temporal desplazados en el tiempo. A este término también se le asigna un orden (d) que se refiere al número de veces que se debe de tomar la diferencia para hacer que la serie temporal sea estacionaria. Los órdenes para los modelos ARIMA serían entonces tres (p,d,q), siendo, respectivamente, observaciones pasadas usadas para AR, diferenciaciones y los términos de MA [17].

2.3.3.3. Modelos SARIMA

Finalmente se encuentran los modelos SARIMA que son una extensión de los modelos ARIMA y permiten modelar series temporales con patrones de estacionalidad.

Estos modelos combinan los componentes descritos anteriormente, así como también términos de estacionalidad para modelar la serie [25].

Los modelos SARIMA incluyen los tres parámetros principales: el orden autorregresivo (p), el orden integrado (d), y el orden de media móvil (q). A estos tres, se les añaden tres parámetros que son similares a los anteriores pero aplicados a las diferencias aplicadas a la estacionalidad: el orden del término autorregresivo estacional (P), el orden del término de media móvil estacional (Q) y término de diferencias estacionales (D). Estos tres nuevos órdenes además tendrían un parámetro relacionado con la longitud del período de la estacionalidad (S) [21].

2.3.3.4. Ejemplo

Un ejemplo de ajuste de un modelo SARIMA, es decir, incluyendo la componente estacional y uso del modelo ajustado para predecir una serie temporal se muestra en la Figura 9. Los datos representan las subvenciones mensuales (en millones de dólares) de un medicamento en Australia de 1991 a 2008, tomados de [14]. En negro se observan los datos de la serie temporal y en rojo las predicciones para los siguientes dos años. Los parámetros del modelo son los siguientes: orden autorregresivo p = 4, orden integrado d = 1, orden de media móvil q = 1, orden autorregresivo estacional P = 0, orden de diferencias estacionales D = 1, orden de media móvil estacional Q = 2, periodo de estacionalidad S = 12. Por lo tanto, el modelo ajustado es un SARIMA(4,1,1)(0,1,2)[12], según la nomenclatura que se utiliza para los modelos SARIMA [19].

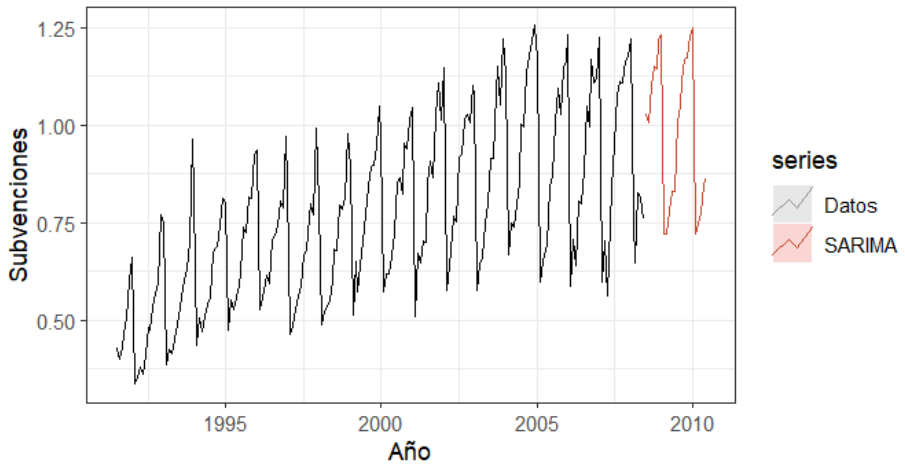

Figura 9. Ejemplo de predicción utilizando SARIMA.

2.4. Métodos de *Machine Learning*

Los modelos estadísticos y el aprendizaje automático (*Machine Learning*) tienen mucho en común, aunque la principal diferencia seria la suposición sobre los datos. En *Machine Learning*, normalmente se asume que los datos han sido generados por algún proceso desconocido de generación de datos y se usan distintos algoritmos de aprendizaje para aproximarse a él. En los modelos estadísticos, por otro lado, se asume que los datos están sujetos a un modelo siendo este el proceso de generación de los datos y se intenta estimar los parámetros de este modelo.

Dos de estas técnicas de desarrollo de algoritmos de aprendizaje son KNN y SVM. Ambas técnicas se utilizan para la predicción en muchos campos, pero últimamente se vienen utilizando en series temporales por sus habilidades para formar sistemas complejos no lineales a partir de una muestra de datos [26].

2.4.1. KNN

Sus orígenes se remontan a la década de 1950, cuando se desarrollaron las bases de la clasificación o reconocimiento de patrones, y, desde entonces, ha sido objeto de diversas contribuciones a lo largo del tiempo. Aunque el algoritmo KNN ha sido utilizado durante décadas, su popularidad ha aumentado considerablemente con el crecimiento de la disponibilidad de datos y el desarrollo de técnicas de procesamiento más potentes. A medida que se ha vuelto más fácil y eficiente almacenar grandes conjuntos de datos y calcular distancias en entornos computacionales, el KNN se ha convertido en una opción atractiva en el campo del aprendizaje automático [27].

KNN es un método utilizado tanto para clasificación como regresión. Consiste en almacenar una colección de instancias (o ejemplos). Cada instancia está formada por un vector de características y su clase asociada (para clasificación) o valor numérico asociado (para regresión). Si se presenta una nueva instancia, KNN encontrará sus k instancias más similares o cercanas (llamadas vecinos cercanos o *nearest neighbours* en inglés), en base a una distancia (como puede ser la distancia Euclidea), y predice su clase con la más repetida en sus vecinos cercanos o, en el caso de la regresión, predice su valor como una agregación de los valores asociados a sus vecinos cercanos.

Para series temporales funciona de igual manera a la de regresión. Si se tiene una serie temporal hay que determinar cómo se forman las instancias, es decir, cuáles son las

características y cuál es el valor asociado a cada instancia. El valor de una instancia será un valor de la serie temporal y sus características serán valores retardados de este valor [28].

Si se toma como ejemplo la serie temporal t = {1,2,3,4,5,6,7,8} y se quiere predecir un nuevo valor, tomando k = 2 (los dos vecinos más próximos) los valores para el vector de las características asociadas al nuevo valor serán las dos últimas instancias, es decir, {7,8}. Estas dos instancias, que son las más similares (más cercanas en el tiempo) de este nuevo valor, tiene como vector de características {5,6} y {6,7} respectivamente y sus valores {7,8} se promediarán para calcular la predicción (7.5). En la Figura 10 se puede ver un ejemplo práctico de esta metodología, utilizando los datos de las muertes en accidentes de tráfico mensuales de EE.UU. entre 1973 y 1979 (tomados de [28]). Se muestra la serie temporal en negro y en rojo las predicciones para el siguiente año.

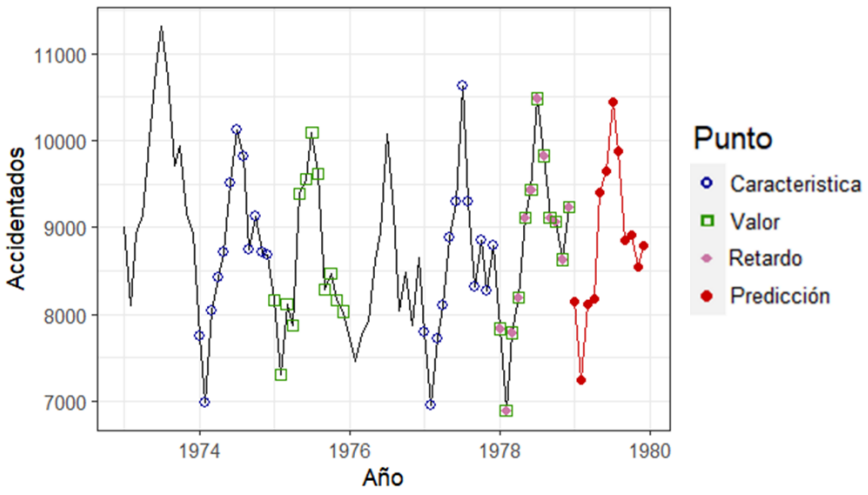

Figura 10. Ejemplo de predicción utilizando KNN.

2.4.2. SVM

Este método es un algoritmo de aprendizaje supervisado que se utiliza para problemas de clasificación y regresión. Su desarrollo se atribuye principalmente a Vladimir Vapnik y su equipo a fines de la década de 1990 en los laboratorios AT&T Bell.

Vladimir Vapnik, junto con Alexey Chervonenkis, trabajaron en el desarrollo de la teoría del aprendizaje estadístico y propusieron el concepto de SVM como una técnica para la clasificación de patrones. En 1995, Vapnik publicó un libro llamado *The Nature of Statistical Learning Theory* donde se presentaron los fundamentos teóricos de SVM [29].

SVM es un método de aprendizaje supervisado que generalmente se utiliza para problemas de clasificación, pero también se puede usar para regresión, normalmente denominado SVR (de sus siglas en ingles Support Vector Regression).

La idea de SVM es encontrar un hiperplano que separe los datos en diferentes clases y maximizar el margen de esta separación. Sin embargo, cuando los datos no son linealmente separables, se utiliza una kernel que es una función matemática utilizada para medir la similitud entre pares de vectores en un espacio de más dimensiones donde si fueran separables, para encontrar en este el hiperplano de separación [26].

Una forma común de utilizar SVM en la predicción de series temporales es considerar el problema como un problema de regresión, donde una máquina de aprendizaje lineal aprende una función no lineal en un espacio de dimensiones superiores generado por un kernel. Es esta función la se utilizará posteriormente para realizar predicciones de valores futuros.

Generalmente, los datos deberán ser procesados para crear un conjunto de datos de entrada que contengan información del instante de tiempo de la observación. Esto puede implicar, por ejemplo, la creación de variables retardadas, donde cada variable representa el valor de la serie temporal en un paso de tiempo anterior. También se pueden crear unas variables *dummy* por cada opción que haya en el conjunto de datos, es decir, si se trata de datos mensuales, una de estas nuevas variables será “Enero” y tomará el valor 1 cuando la observación corresponda a dicho mes y 0 en el resto de los datos (lo mismo para el resto de los meses). Este último método resume muy bien el carácter estacional de la serie, pero no serviría para series con tendencia [30].

En la Figura 11 se encuentra un ejemplo de la predicción para los siguientes dos años de la serie de datos mensuales de la temperatura del castillo de Nottingham. En este caso, al ser datos mensuales, la variable *dummy* utilizada son los meses. En negro se encuentran los datos originales y en rojo la predicción.

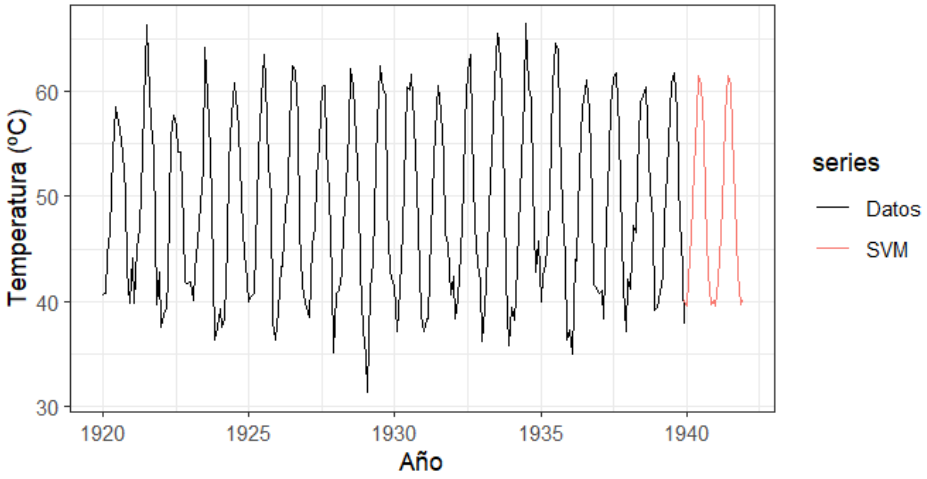

Figura 11. Ejemplo de predicción utilizando SVM.

Capítulo 3

Estudio de la temperatura en Gijón

El objetivo principal de este trabajo es realizar un estudio comparativo entre modelos de series temporales clásicos y métodos de *Machine Learning* para predecir la temperatura media diaria en Gijón, una ciudad costera ubicada en el norte de España.

La elección de que el estudio comparativo se haga con una variable climatológica se debe a que la predicción de estos valores en una región puede tener una gran repercusión para distintos sectores como la agricultura, el turismo y la gestión de los recursos naturales. No solo para la planificación y toma de decisiones de estos, si no que, además, es un campo que influye en nuestro día a día de manera directa. Por lo tanto, poder hacer predicciones precisas sobre estas variables resulta de gran utilidad.

Por otro lado, la elección de Gijón como la región de estudio, se debe a que el autor del trabajo es de esta ciudad. Esto no solo aporta un valor personal, si no que gracias a ello se tiene un mejor conocimiento del clima de la región (en comparación a una región arbitraria), lo cual facilita la interpretación de los resultados.

3.1. Origen de los datos

Los datos usados para este estudio provienen de una única fuente, la Agencia Estatal de Meteorología (AEMET) [31]. En concreto, con el software libre de R [14] se accedió a los datos gracias a que la agencia permite descargarlos en formato JSON a través del portal de OpenData utilizando una API-REST [31].

Los datos meteorológicos utilizados son diarios del periodo entre el 1 de enero de 2002 y el 31 de diciembre de 2022 (ambos inclusive). Debido a que la base de datos consultada comprendía entre el año 2002 hasta el día en el que se consultaron (en este caso 22 de febrero de 2023), se decidió empezar por la fecha de inicio de la base de datos y terminar con el último año completo que estuviera disponible para trabajar con años completos. El indicativo de la central meteorológica que tomaba los datos es 1208H, correspondiente a la zona del Puerto de Gijón, Asturias.

En estos también venían incluidos otras características como el indicativo, el nombre, la provincia o la altitud, que por ser comunes a todos los datos no se consideraron en el estudio.

3.2. Preparación de los datos

Con el objetivo de tener una base de datos de mayor calidad para un mejor estudio y análisis, una vez recopilados, se realizó una limpieza y preprocesamiento o depuración de estos. Este proceso consiste en identificar y corregir o eliminar los incorrectos, incompletos, duplicados o irrelevantes. Este es un paso crítico en el análisis de datos, ya que una mala calidad de los mismos puede afectar negativamente a los resultados y las conclusiones obtenidas [32]. Este procedimiento se realizó para todas las variables ya que se consideró importante debido al aporte de información que pudieran tener en un posterior uso.

En esta sección, se describirá el preprocesamiento realizado a los datos, como se imputaron los valores faltantes y finalmente con qué conjunto de datos se continuó el análisis.

3.2.1. Preprocesamiento de las variables del conjunto de datos

En la Tabla 1, se presentan las variables con sus unidades y junto con el tipo de dato que es.

Tabla 1. Presentación de las variables incluidas en los datos.

id	Descripción	Tipo datos	unidad
fecha	fecha del día (AAAA-MM-DD)	Cadena de caracteres	*#N/D*
tmed	Temperatura media diaria	Numérico	*°C*
prec	Precipitación diaria de 07 a 07	Numérico	*mm*
tmin	Temperatura Mínima del día	Numérico	*°C*
tmax	Temperatura Máxima del día	Numérico	*°C*
sol	Insolación	Numérico	*horas*
presmax	Presión máxima al nivel de referencia de la estación	Numérico	*hPa*
presmin	Presión mínima al nivel de referencia de la estación	Numérico	*hPa*

Un ejemplo de los datos obtenidos para un día se muestra en la Tabla 2.

Tabla 2. Extracto de la base de datos meteorológicos diarios de la ciudad de Gijón.

fecha	nombre	provincia	tmed	prec	tmin	tmax	sol	presMax	presMin
01/01/2002	GIJÓN, PUERTO	ASTURIAS	9,6	1,9	7,3	11,8	1	1020,8	1017

A continuación, se detallan los ajustes realizados al formato de las variables:

- Las fechas estaban en formato de caracteres por lo que se cambiaron a tipo de dato de fecha. Esto se hizo por la manera que tienen las funciones y los paquetes de manipular los datos y permitir una mejor interpretación de estos.
- A las variables que están definidas por medio de números, se les cambió la coma por un punto como representación del número decimal.

3.2.2. Tratamiento de datos faltantes

Los datos faltantes (también conocidos como valores perdidos o valores ausentes) son valores que no se encuentran presentes en un conjunto de datos. Es decir, son valores que no se han registrado o que se han perdido durante la recopilación, el procesamiento o la transferencia de los datos. En climatología es común encontrarse con este tipo de problemas debido a fallos en la instrumentación, problemas en la transmisión de datos o cambios en la configuración de las estaciones meteorológicas [33].

En la serie de datos de la variable temperatura, no se encontraron de datos faltantes, pero si en las variables de precipitación, insolación y presión. Debido a que estas se consideraron relevantes para la calidad de la base de datos, se hizo un tratamiento de estas variables.

Hacer este tipo de tratamiento siempre aporta un valor añadido, para obtener mayor precisión y fiabilidad en los resultados y en la construcción de modelos, ya que los datos faltantes pueden afectar negativamente, especialmente si hay una gran cantidad (para muchos modelos y metodologías, incluso es estrictamente necesario debido a que, de no ser así, no se podría procesar la información). Por un lado, los datos faltantes pueden introducir sesgos en el análisis, lo que puede llevar a conclusiones erróneas y, por otro lado, los datos faltantes pueden reducir la precisión de las predicciones, ya que estos

modelos y metodologías necesitan una cantidad significativa de datos para hacer predicciones precisas.

Aunque existen distintas formas de tratar los datos faltantes (ver, por ejemplo, [34]), para este trabajo se han utilizado las siguientes técnicas:

1. Interpolación lineal: implica que, para un valor faltante en un día determinado, se estima como el punto medio entre el valor del día anterior y el siguiente. Si hay varios valores faltantes consecutivos, se estima un conjunto de puntos igualmente espaciados a lo largo de una línea recta que conecta el último valor registrado y el siguiente.
2. Última observación llevada a cabo (LOCF, de sus siglas en inglés, *Last Observation Carried Forward*): la imputación se hace sustituyendo el valor faltante por el de la observación inmediatamente anterior. Aunque es una técnica sencilla, la información que puede estar dando es errónea si los datos no se comportan de una manera en la que se puedan encontrar varios días consecutivos el mismo valor (por ejemplo, si dos días consecutivos no suele llover lo mismo).
3. Valor medio: implica imputar los valores faltantes con la media de todos los valores observados. Es una técnica muy simple y rápida pero su problema es que desprecia la varianza y puede alterar la relación entre variables.
4. Media mensual: es también una media, pero esta se basa en sustituir el valor faltante por la media mensual. Usar un valor más acotado como la media mensual, en vez de la media global, suele ser adecuado para datos climatológicos ya que proporciona un valor más ajustado a ese momento del tiempo o estación. Por ejemplo, para meses en los que llueve más, dará un valor de precipitación más cerca de la real. Cabe destacar que esta aproximación se puede hacer si los datos tienen periodicidad menor a la mensual (datos diarios, semanales, bimensuales…), en el caso, por ejemplo, de ser datos trimestrales, se podría utilizar la media anual.
5. Descomposición estacional: esta metodología divide la serie temporal en temporadas y después realiza la imputación por separado para cada uno de los conjuntos de datos de series temporales resultantes (cada una contiene los datos de una estación específica) calculando el valor por interpolación lineal.

Para tratar los datos faltantes se utilizó la librería 'imputeTS' de R [35], la cual permite visualizar y tratar datos faltantes de manera eficiente.

3.2.2.1. Datos Faltantes de la variable precipitación

A continuación, se describe cómo se realizó el procedimiento para tratar los datos faltantes de la variable precipitación.

En primer lugar, se visualizó cuál era la distribución de los datos faltantes con las gráficas (Figura 12). Como se puede observar, la distribución es similar en todos los años sin tener ningún año con ausencia de datos mayor al 10 %, por lo que se optó por seleccionar una ventana para observar el resultado de la aplicación de los distintos métodos para tratar los datos faltantes. Esta ventana fue de 1 año (se escogió de manera arbitraria del 1 de septiembre de 2005 al 1 de septiembre de 2006). Con esta se podría observar cómo estarían representados los datos imputados en distintas épocas del año para evaluar si pudiera ser una manera correcta de estimarlos (Figura 13).

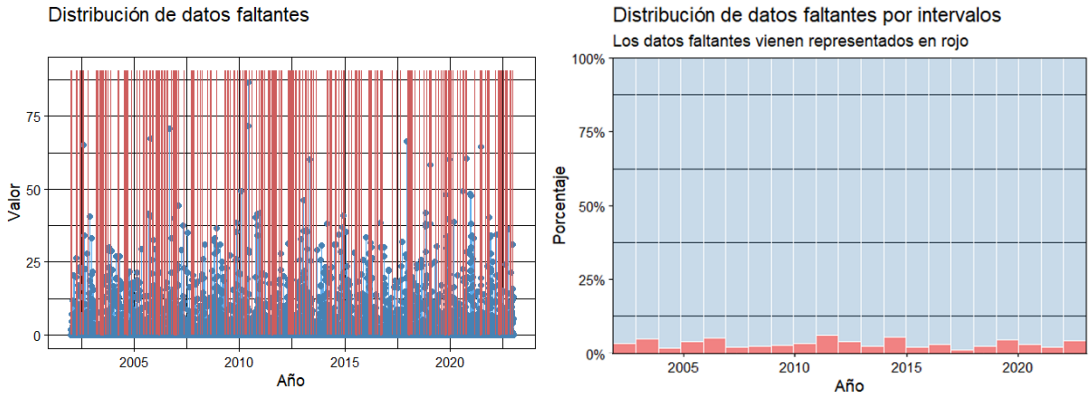

Figura 12. Distribución de los datos faltantes de la variable precipitación, donde las líneas rojas representa la ausencia de un dato (izquierda) y por intervalos de un año donde, por cada uno, se encuentra la sección roja representando el porcentaje de datos ausentes en ese intervalo (derecha).

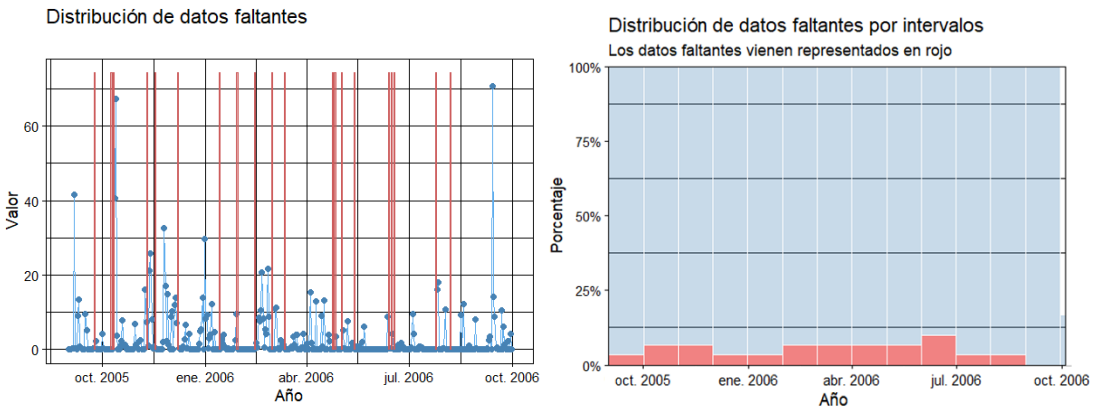

Figura 13. Distribución de los datos faltantes de la variable precipitación para la ventana de tiempo seleccionada (1 de septiembre de 2005 al 1 de septiembre de 2006), donde las líneas rojas representan la ausencia de un dato (izquierda); y por intervalos de 30 días donde, por cada uno, se encuentra la sección roja representando el porcentaje de datos ausentes en ese intervalo (derecha).

1. Interpolación lineal. En la Figura 14 se muestran los resultados de esta metodología. Esta deja valores intermedios entre los dos adyacentes, lo que a priori parece una buena aproximación, pero se prefirió explorar otras metodologías que reflejaran mejor el comportamiento de esta variable.

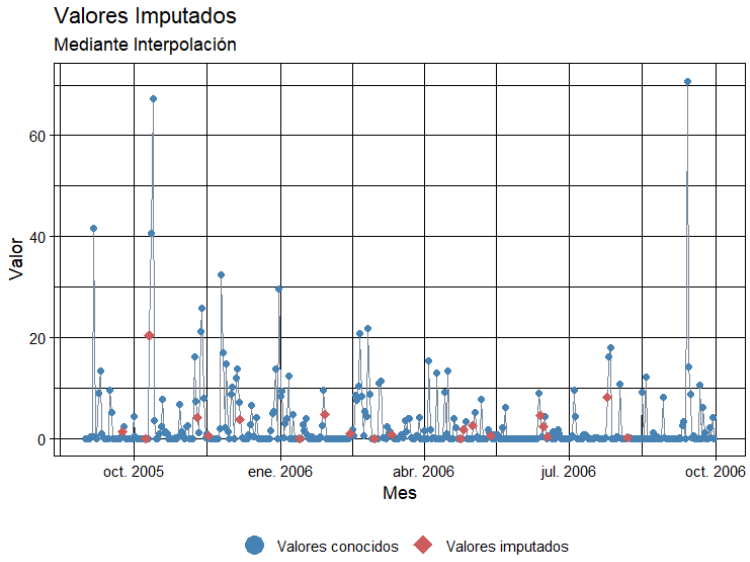

Figura 14. Datos imputados de la variable precipitación mediante interpolación lineal.

2. LOCF. Como se puede observar en la Figura 15, al ser el dato imputado una sustitución usando la última observación, esto deja unos resultados muy irreales para la precipitación usual en Gijón dado que es poco habitual que dos días seguidos llueva exactamente la misma cantidad.

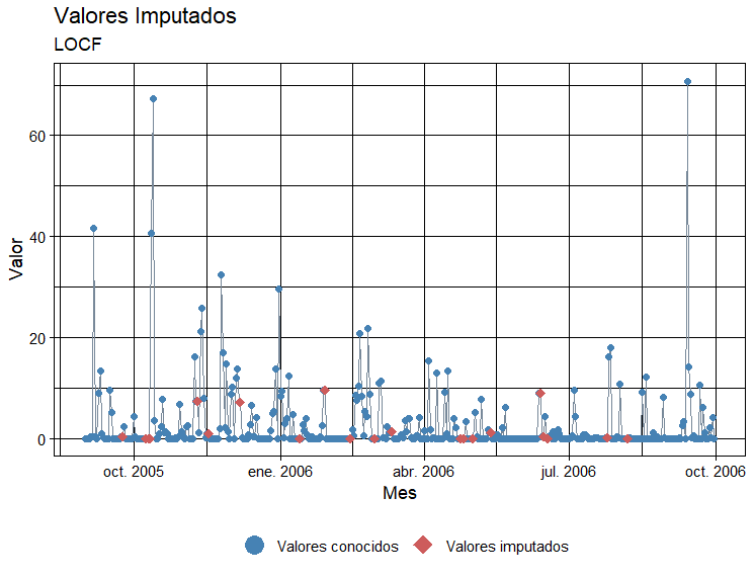

Figura 15. Datos imputados de la variable precipitación mediante LOCF.

3. Valor medio. Como se ha comentado en la descripción de las distintas metodologías y se ve representado en la Figura 16, el valor medio en este caso no parece una opción conveniente dado que los datos imputados no estarían influenciados por la estación del año o las condiciones de ese momento, por lo que, para este estudio, no se consideró como una imputación adecuada.

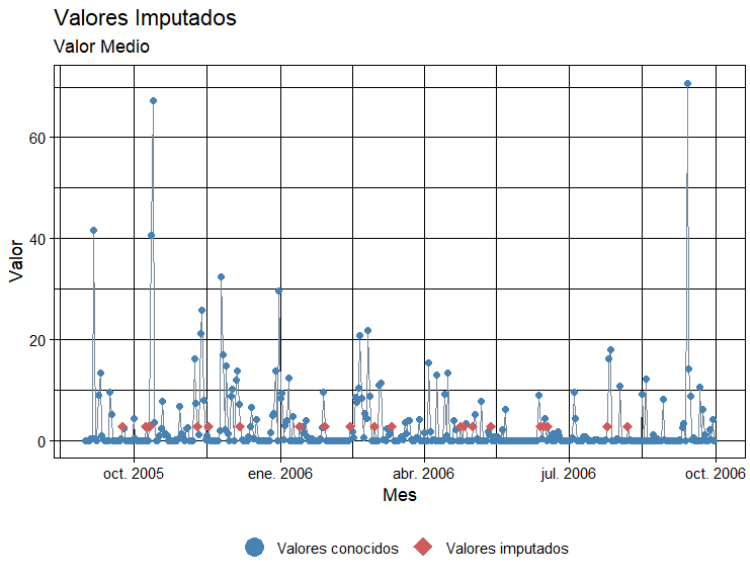

Figura 16. Datos imputados de la variable precipitación mediante valor medio.

4. Valor medio mensual. Este valor se consideró la mejor opción para imputar los datos meteorológicos de la precipitación porque ofrece una estimación de la cantidad de lluvia esperada durante un mes determinado como se puede observar

en la Figura 17. Esta opción parece proporcionar una medida más factible al tener en cuenta el momento del año en el que se encuentra, además, tendría en cuenta si ese mes coincide con una escasez o abundancia anormal de lluvias.

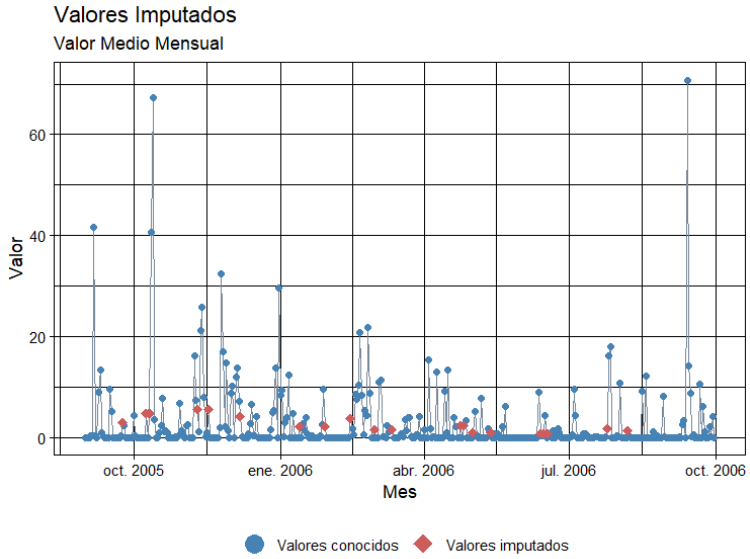

Figura 17. Datos imputados de la variable precipitación mediante valor medio mensual.

5. Descomposición estacional: como se puede observar en la Figura 18, la descomposición estacional resulta en una imputación muy similar al de la interpolación lineal (el cual no parecía apropiado), por lo que no se consideró adecuada.

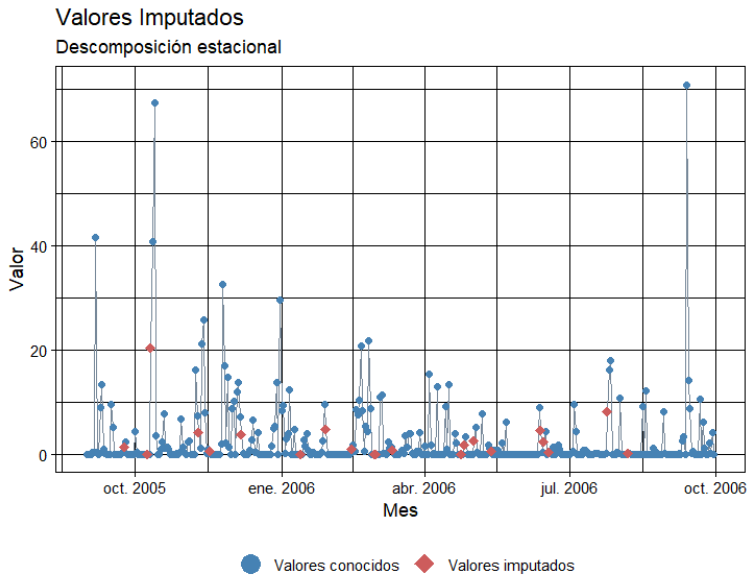

Figura 18. Datos imputados de la variable precipitación mediante descomposición estacional.

Finalmente, teniendo en cuenta lo expuesto previamente, se escoge el valor medio mensual para realizar la imputación de los datos de la variable precipitación.

3.2.2.2. Datos Faltantes de la variable insolación

A la hora de imputar los datos faltantes de la variable insolación, se encontró una cantidad importante de estos a partir de 2020 (Figura 19), llegando incluso a tener periodos del 100% de datos faltantes, es decir, años completos sin datos (Figura 20). Debido a esta ausencia de información, se decidió no considerar los datos correspondientes a los años 2021 y 2022, en espera a una posible actualización de esta por parte de la AEMET.

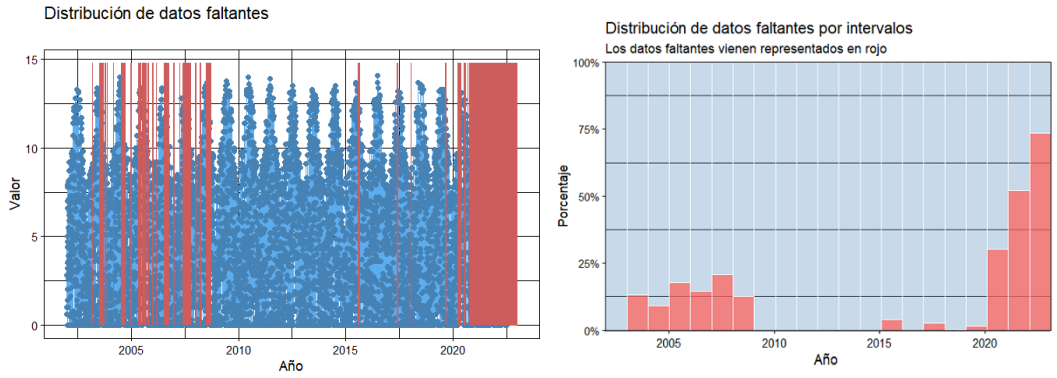

Figura 19. Distribución de los datos faltantes de la variable insolación, donde las líneas rojas representa la ausencia de un dato (izquierda) y por intervalos de un año donde, por cada uno, se encuentra la sección roja representando el porcentaje de datos ausentes en ese intervalo (derecha).

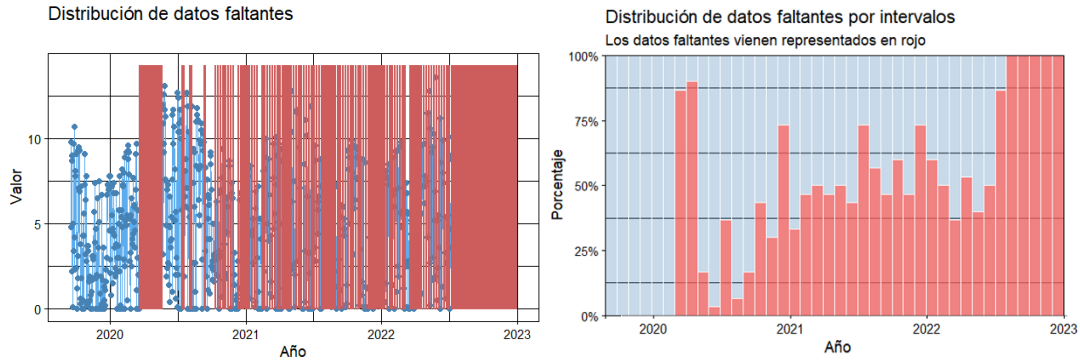

Figura 20. Distribución de los datos faltantes de la variable insolación a partir de 2020, donde las líneas rojas representa la ausencia de un dato (izquierda); y por intervalos de 30 días donde, por cada uno, se encuentra la sección roja representando el porcentaje de datos ausentes en ese intervalo (derecha).

Al igual que para los datos de precipitación, se utilizó una ventana para visualizar de mejor manera los datos imputados. En este caso, al tener una mayor cantidad de datos faltantes, la ventana fue de 2 años para incluir más ejemplos (del 1 de septiembre de 2004

al 1 de septiembre de 2006). De esta manera se logra visualizar mejor la forma de imputar los datos que tienen las distintas metodologías y comprobar si encaja con un comportamiento natural de la variable (Figura 21).

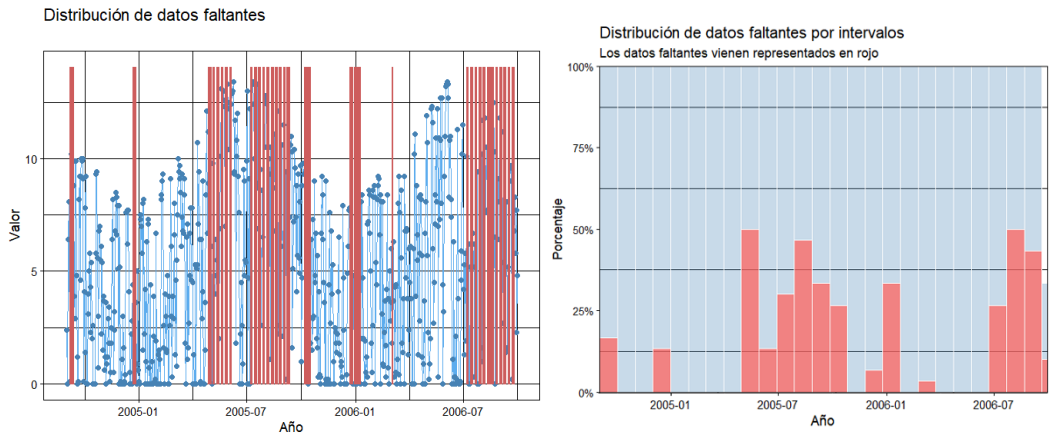

Figura 21. Distribución de los datos faltantes de la variable insolación en la ventana de tiempo seleccionada (del 1 de septiembre de 2004 al 1 de septiembre de 2006) (izquierda) y por intervalos de 30 días (derecha).

1. Interpolación lineal. Como se puede ver en Figura 22, la interpolación lineal deja valores fijados a una recta. Por lo tanto, para momentos en los que hay una cantidad de valores faltantes seguidos, no parece ser un comportamiento natural de la variable en comparación a las otras observaciones.

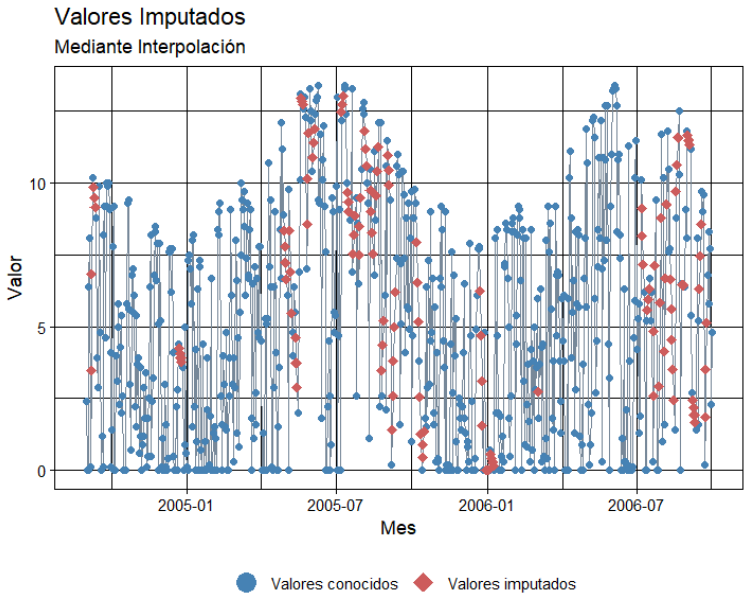

Figura 22. Datos imputados de la variable insolación mediante interpolación lineal.

2. LOCF, valor medio y valor medio mensual. Como se puede observar en los gráficos de la Figura 23, estas tres opciones resultan en una representación no realista dado que una recta horizontal. Esto, en momentos en los que existan una serie de datos faltantes consecutivos, no sería un reflejo de un comportamiento natural de las horas de sol.

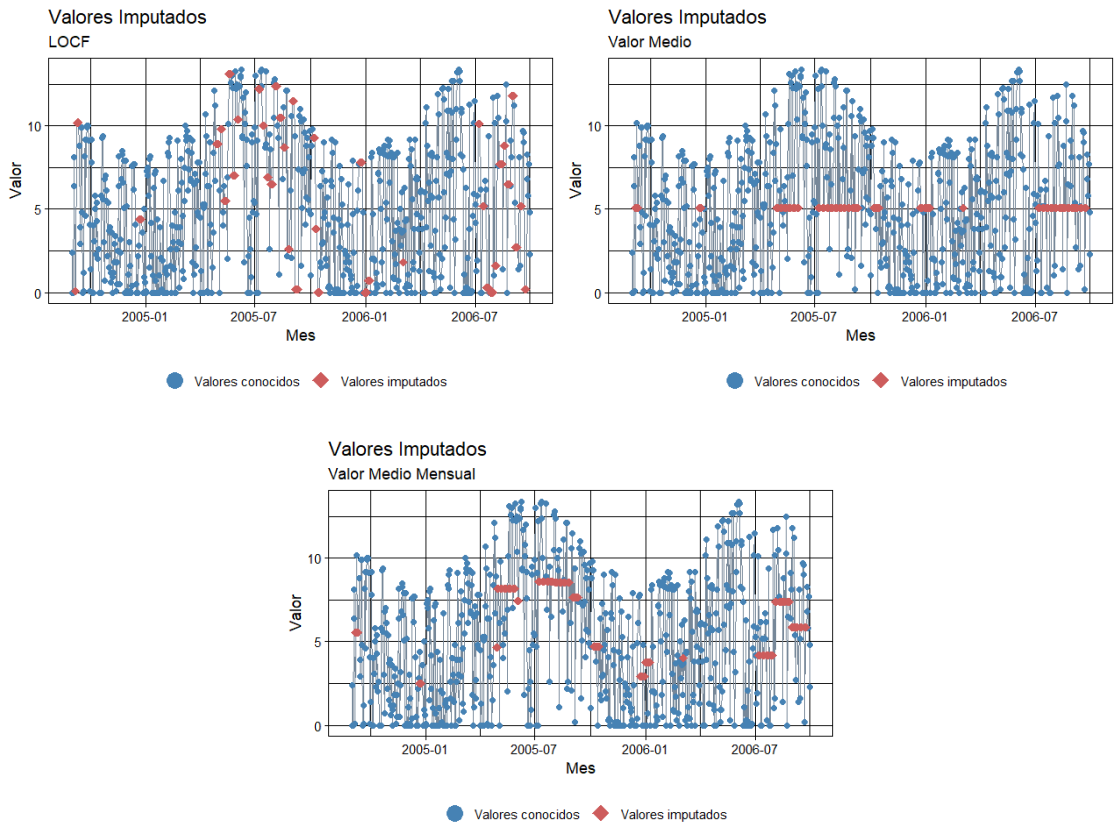

Figura 23. Datos imputados de la variable insolación mediante LOCF (arriba izquierda), valor medio (arriba derecha) y valor medio mensual (abajo).

3. Descomposición estacional. Al realizar una partición estacional de la serie temporal y después realizar la imputación de manera separada, esta metodología, al contrario que las anteriores, permite generar distintos puntos para un conjunto de datos faltantes consecutivos. Esto soluciona el problema, con respecto a las técnicas anteriores, de generar una línea recta ante varios datos faltantes.

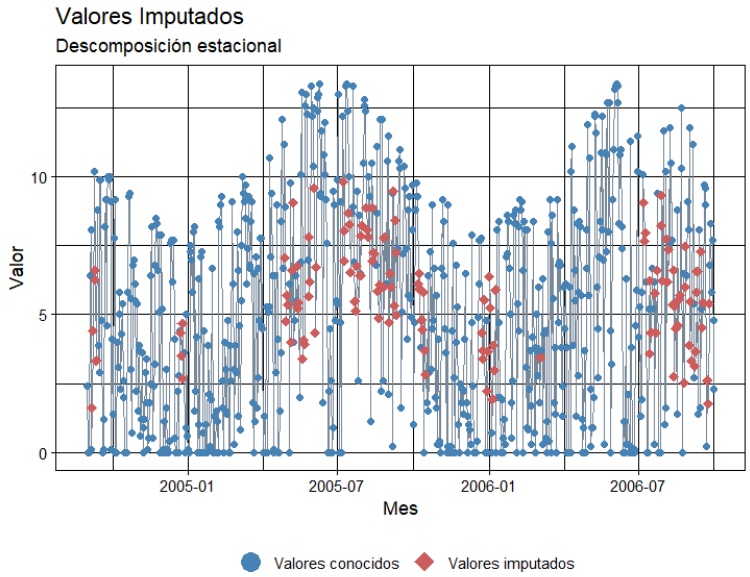

Figura 24. Datos imputados de la variable insolación mediante descomposición estacional (mostrando la ventana seleccionada para una mejor visualización).

Observando la naturaleza de los datos y el comportamiento de los datos imputados, tanto la interpolación lineal como LOCF, valor medio y valor medio mensual, se decidió que no eran opciones adecuadas ya que se creyó que arrojaban resultados poco realistas. Por ese motivo, se decidió que la opción más adecuada era la de descomposición estacional.

3.2.2.3. Datos Faltantes de las variables presión mínima y presión máxima

En cuanto a los datos de presión mínima y presión máxima, solo existía un dato ausente por lo que fue sustituido por el valor medio. Además, se calculó la variable de presión media (haciendo la media para todos los días entre su valor máximo y su mínimo).

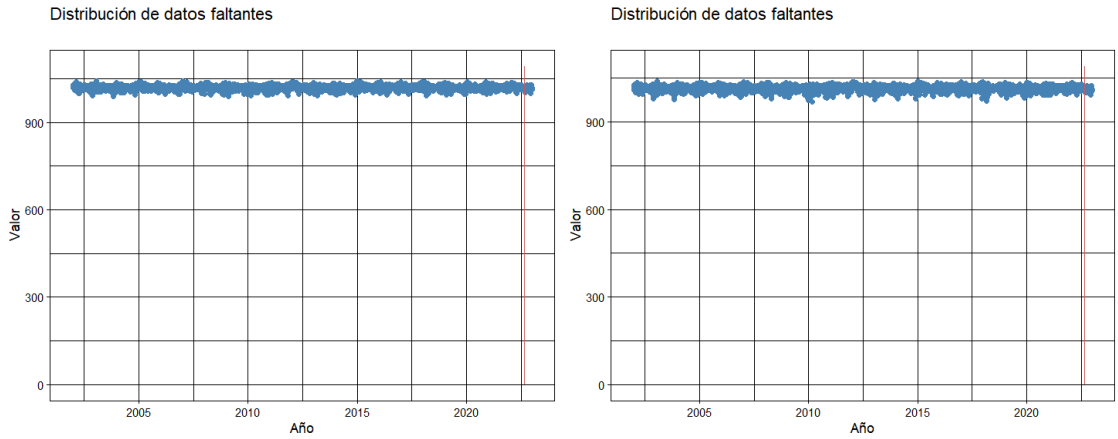

Figura 25. Datos faltantes de la variable presión máxima (izquierda) y presión mínima (derecha).

3.2.2.4. Decisión sobre la base de datos

Finalmente se continuó con los datos imputados con los métodos escogidos y sin tener en cuenta aquellos a partir del 31 de diciembre de 2019 debido a la ausencia de datos de la variable insolación.

Capítulo 4

Análisis de los datos

4.1. Análisis exploratorio de los datos

Una vez limpiados y procesados los datos se realizó un análisis exploratorio para entender mejor su naturaleza. Este análisis se hizo de dos maneras: mediante un estudio estadístico descriptivo y mediante una exploración gráfica de los datos.

4.1.1. Estudio descriptivo

Se realizó un estudio descriptivo de las variables calculando distintos valores estadísticos, recopilados en la Tabla 3. En esta tabla se incluyen el valor medio, la desviación estándar, la mediana, el rango (o diferencia entre el valor mínimo y máximo), el sesgo o asimetría de la distribución (una distribución con sesgo cero sería simétrica, mientras que una distribución con sesgo positivo tendría más datos hacia la derecha de la media y una con sesgo negativo, más datos hacia la izquierda) y la curtosis o forma de la distribución (describe cómo de aplanada es la curva, valores altos indiciaría que los valores están más centrados en la media; mientras que valores menores que cero indicarían que la distribución es más aplanada).

Tabla 3. Resumen estadístico de las variables

	Media	sd	Mediana	Min.	Máx.	Rango	Sesgo	Curtosis
Tª media	15,00	4,19	14,90	2,10	26,00	23,90	-0,09	-0,85
Precipitación	2,75	6,15	0,10	0,00	86,50	86,50	4,31	27,72
Insolación	5,10	3,88	4,90	0,00	14,10	14,10	0,27	-1,06
Presión media	1017,17	7,79	1017,70	981,10	1039,95	58,85	-0,51	0,86

Basándose en los valores obtenidos, se puede decir que en las distribuciones de la temperatura media (Tª media), la insolación y la presión media (P media) el sesgo no es muy pronunciado, dado que el valor de este es cercano a cero, lo que puede indicar que las distribuciones son simétricas en torno a la media. La variable precipitación, por otro lado, tiene un sesgo positivo muy pronunciado, indicando que la mayoría de las observaciones están inclinadas hacia la derecha de la media, con valores altos en la cola

derecha, algo que se puede observar claramente en las representaciones gráficas de la Figura 26.

Los valores de curtosis indican que la variable insolación es relativamente plana y tiene menos concentración de datos en el centro en comparación con una distribución normal (cuya curtosis es 3) algo que se puede ver en la Figura 26(e); junto con la variable temperatura que no tendría tampoco un pico agudo de los datos. La variable de presión media ya tendría una forma similar a una distribución normal, siendo su curtosis positiva cercana a 1; y la variable precipitación, cuyo valor cercano a 28 describe el pico que se encuentra en su correspondiente histograma en la Figura 26(f).

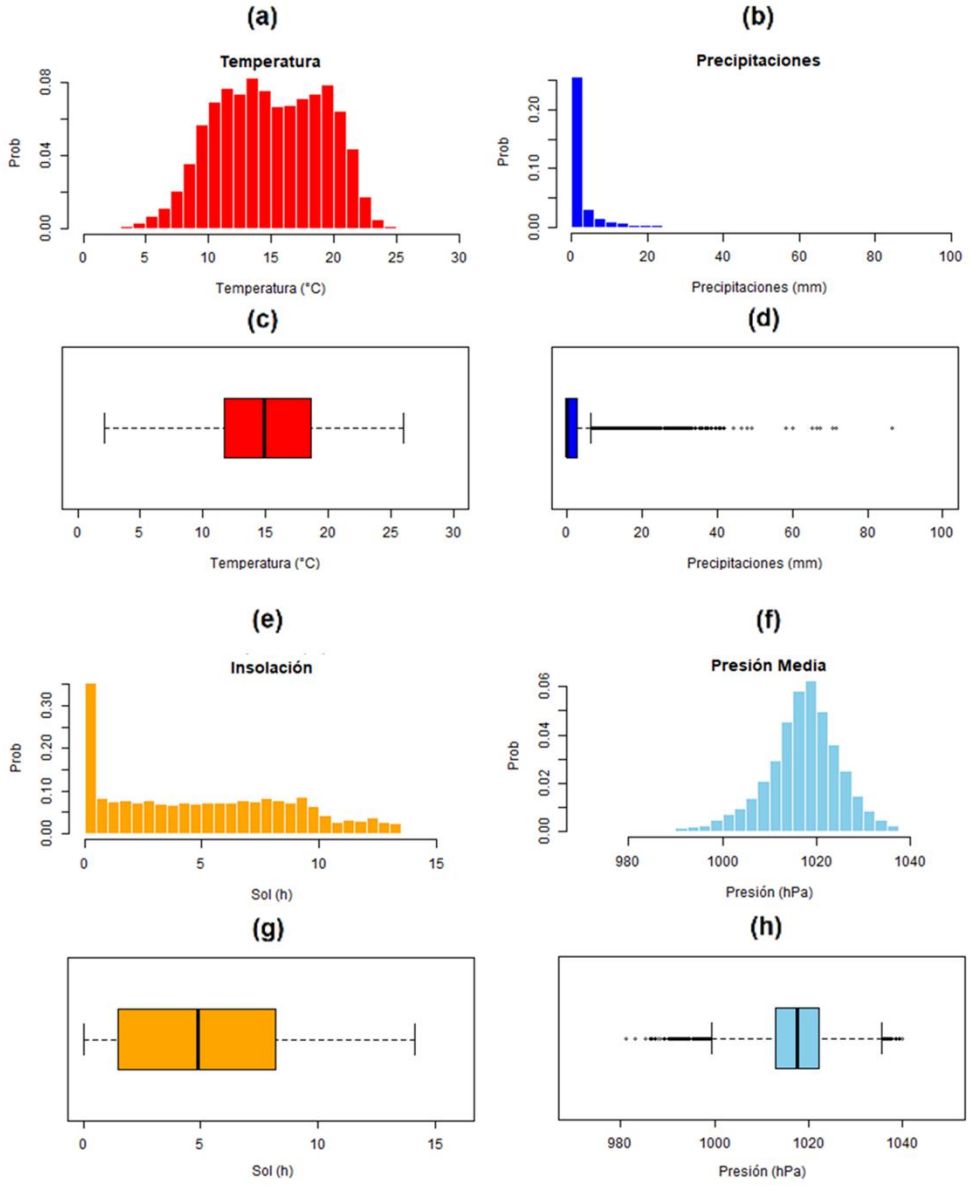

Figura 26. Distribuciones de las 4 variables principales representadas en histogramas y diagramas de cajas.

4.1.2. Correlaciones

Se estudiaron los niveles de correlación entre las distintas variables cuyos valores se pueden ver en la Figura 27. Para ello se utilizó el coeficiente de correlación de Pearson, una medida estadística que se utiliza para evaluar la relación lineal entre dos variables continuas. Este coeficiente se mide en una escala que va desde -1 a 1, siendo 1 una correlación positiva perfecta (cuando una variable aumenta, la otra también aumenta en la misma proporción) y -1 una correlación negativa perfecta (cuando una variable aumenta, la otra disminuye en la misma proporción). Un valor de 0 indicaría que no hay correlación entre las variables [36].

El estudio de correlaciones lo que proporciona es una visión general de las relaciones lineales entre las variables, ayudando a identificar relaciones importantes entre los datos. Además, que exista correlación no implica causalidad. En algunas ocasiones se puede usar alguna de las variables para explicar parte del comportamiento de otra, lo que puede ser especialmente útil en el diseño de futuros estudios o experimentos.

Figura 27. Representación gráfica de la matriz de correlaciones de las principales 4 variables.

En esta matriz de la Figura 27 se puede observar que la temperatura media (tmed) está correlacionada positivamente con la cantidad de horas de sol diarias (solImp) (r = 0.26). Esto tiene sentido ya que indica que cuando la cantidad de horas de sol aumenta, también tiende a aumentar la temperatura (de forma lineal). Por otro lado, se puede ver que la precipitación (precImp) está correlacionada negativamente con la presión atmosférica (presMean) (r = -0.31) y con la cantidad de horas de sol diarias (solImp) (r =

-0.35). Esto significaría que cuando hay mayor precipitación, tiende a haber menor presión atmosférica y/o menor cantidad de horas de sol. Esta relación es acorde a lo que pasa en la realidad, ya que altas presiones tienden a significar un mejor clima, más seco y soleado y, por otro lado, que haya más horas de sol indica que ha habido menos nubosidad, es decir, menos probabilidad de que llueva.

En general, la matriz de correlaciones puede ayudar a entender las relaciones lineales entre las diferentes variables climáticas medidas en una región y puede proporcionar información útil para futuros análisis y modelos climáticos [37].

4.1.3. Discusión

Realizar un estudio descriptivo de una base de datos que incluya múltiples variables, como la presión, la temperatura, la precipitación y la insolación, es importante por varias razones. En primer lugar, ayuda a comprender la estructura y características de la base de datos en su totalidad, lo que se podría aprovechar a la hora de identificar patrones y relaciones que pueden ser relevantes entre las variables. Además, un análisis descriptivo permite identificar posibles problemas con la calidad de los datos, como la presencia de valores faltantes, lo que puede afectar la interpretación de los resultados.

Aunque sólo se va a utilizar la temperatura para el estudio predictivo, tener información detallada y completa de las otras variables también contextualiza los resultados de los modelos predictivos. Por ejemplo, la información sobre la precipitación puede ayudar a entender cómo las variaciones en la temperatura pueden afectar al tiempo atmosférico de una región en particular. Además, si se observa una fuerte correlación entre la temperatura y otra variable, como la presión o la insolación, esto puede ser útil para hacer predicciones más precisas y confiables. En resumen, aunque sólo se vaya a utilizar una variable para un estudio predictivo, en este caso la temperatura, tener un conocimiento profundo y completo de las demás variables mejorara la calidad y confiabilidad de los resultados del estudio.

4.2. Análisis de la serie temporal de la temperatura media

Desde este punto se estudia la serie temporal de la variable temperatura media para continuar con el proyecto por las razones explicadas a continuación. En primer lugar, la temperatura es una variable relativamente fácil de medir y monitorizar, lo que significa que los datos disponibles suelen ser bastante precisos, lo que hace que el análisis sea más robusto y fiable. Otra razón considerada de peso es que la temperatura es una variable intuitiva. No solo porque los resultados puedan ser fácilmente visualizados y explicados a través de gráficos, sino porque sus resultados son muy fáciles de interpretar. Por ejemplo, no sólo es fácil entender que en verano hará más calor que en invierno y por ello tendremos estacionalidad, sino porque, las temperaturas predichas para el verano sean de 100 ºC y las del invierno de -50ºC para la ciudad en Gijón, indican que el modelo no estaría ajustándose correctamente a los datos.

Por último, los modelos de series temporales y de *Machine Learning* escogidos para este proyecto son univariantes. Aunque los modelos de *Machine Learning* pueden ser multivariantes, una de las premisas de este proyecto era realizar la comparativa de las predicciones partiendo de la base que todos los modelos debían tener la misma información inicial.

4.2.1. Serie temporal

Como se dijo previamente, la serie temporal considerada se extiende desde el 1 de enero del 2002 hasta el 31 de diciembre de 2019. Usando la librería “forecast” de R [38] se cambió el formato de datos a serie temporal, pero dado que este acepta frecuencias de 365 y no admite años bisiestos, se eliminaron los datos de los cuatro 29 de febrero que existían en la serie.

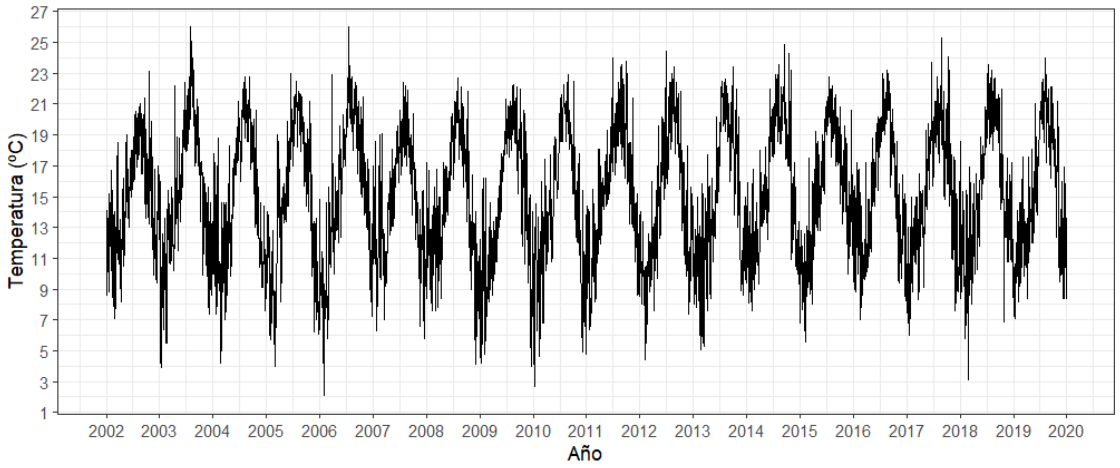

Figura 28. Gráfica de la serie temporal de la variable temperatura

Se procedió a descomponer la serie temporal para recopilar sus principales componentes (tendencia, estacionalidad y componte aleatorio) en un gráfico. Esta función utiliza el método de descomposición clásico, donde se asume que las componentes se combinan de forma aditiva. Descomponer una serie temporal es útil para poder analizar la estructura temporal y a partir de ello modelar la serie. En la Figura 29 se pueden observar, de arriba hacia abajo, la serie de datos original, la tendencia, la estacionalidad y la variación aleatoria.

La variación aleatoria hace referencia a la variabilidad residual que tendría la temperatura que no se puede explicar por la tendencia o la estacionalidad. En este caso, indica que los valores de la variación aleatoria de la temperatura diaria están dentro de un rango de -7 y 10 aproximadamente. Esto significa que, en cualquier punto de la serie, la variación aleatoria puede estar dentro de ese rango, es decir, que si un día en específico el valor de la variación aleatoria es de -3, esto indica que ese día el valor de la temperatura observada tuvo una fluctuación aleatoria de -3 ºC.

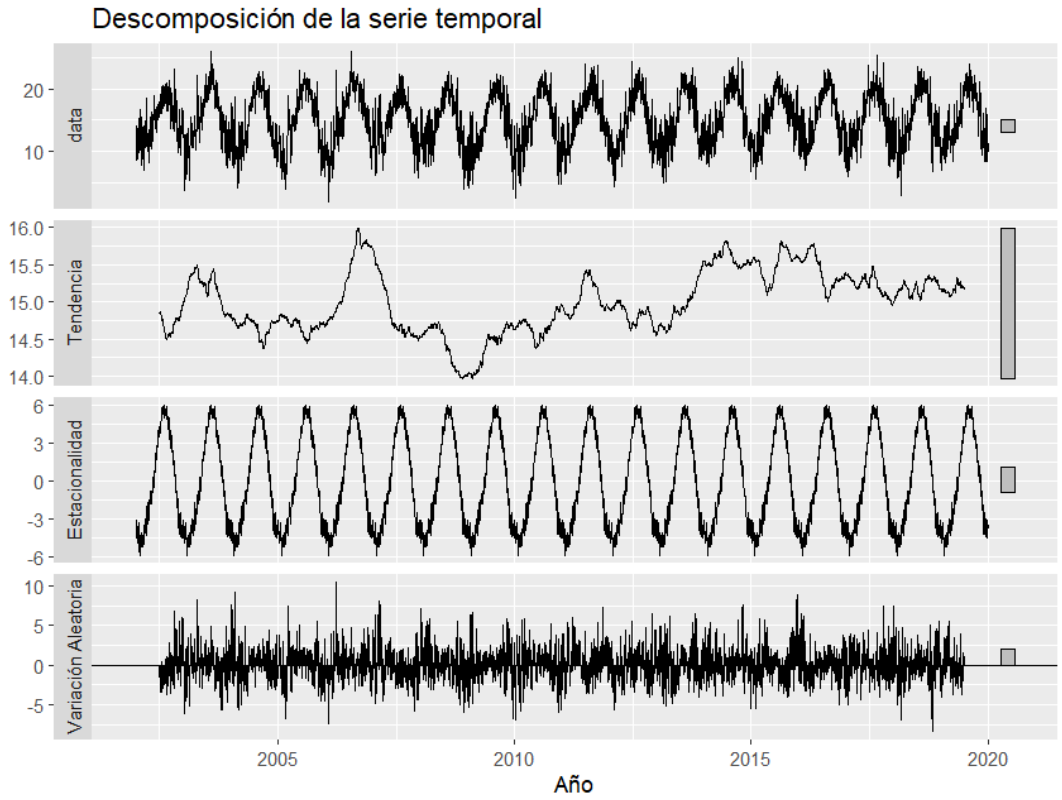

Figura 29. Representación gráfica de las distintas componentes.

4.2.1.1. Tendencia

La línea de tendencia de la temperatura muestra tendencias locales ascendentes y descendentes en distintos puntos, pero con una tendencia global ascendente poco marcada. Esto sugiere que, pese a las fluctuaciones, la temperatura media podría estar aumentando con el tiempo. Sin embargo, si se incorpora la línea de tendencia a la serie temporal como se puede observar en la Figura 30, esta tendencia no es tan perceptible.

Este factor es importante tenerlo en cuenta a la hora de determinar situaciones como el incremento de las temperaturas de una región, si se observa la línea de tendencia en un periodo de 10 años incorporada a la serie temporal, es posible que no se observe una tendencia clara, pero si se observa la serie a 30 años y se observa únicamente la línea de tendencia, se puede percibir un incremento gradual de la temperatura media.

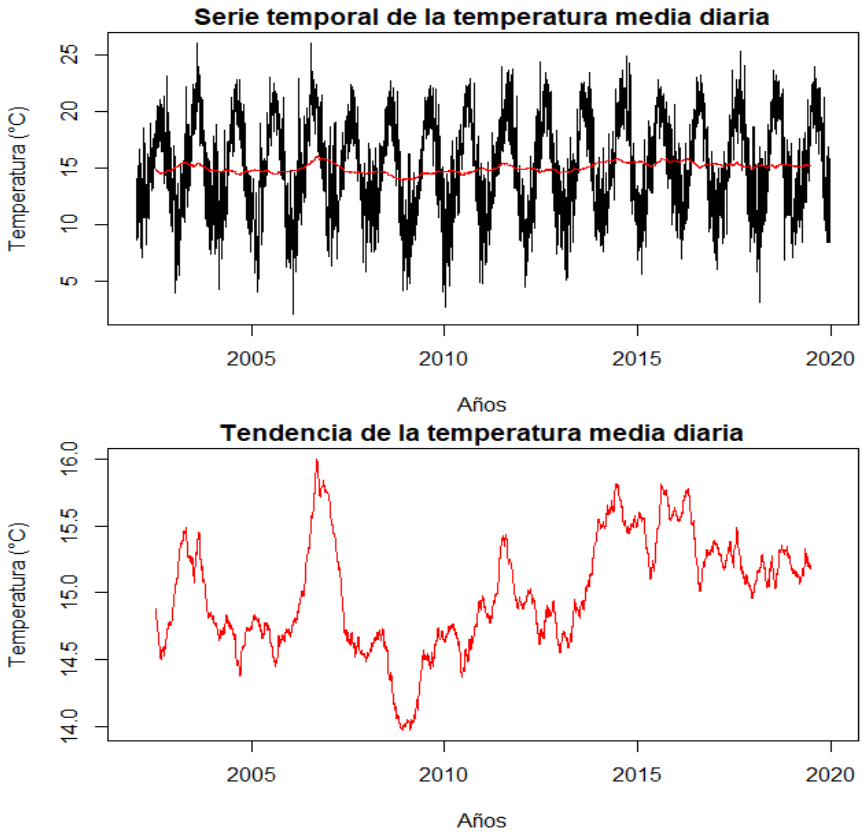

Figura 30. Gráfico de tendencia de la temperatura diaria, incorporada a la serie temporal donde la línea roja es la tendencia y la negra la serie temporal (arriba), gráfico de la tendencia aislada (abajo).

4.2.1.2. Estacionalidad

Se estudió la estacionalidad dado que ayuda a comprender las variaciones regulares que se producen en la serie a lo largo del año. Esto permite que se escojan de mejor manera los modelos al permitir seleccionar los parámetros adecuados, haciendo predicciones más precisas.

Como era de esperar, debido a las estaciones, existen patrones regulares en los datos de temperatura que se repiten cada año en un momento especifico. Se puede observar que la temperatura diaria media es más alta en los meses de verano (junio, julio y agosto) y más baja en los meses de invierno (diciembre, enero, febrero). Si se toman los primeros tres años de la serie, como se muestra en la Figura 31, se puede observar dicho patrón fácilmente.

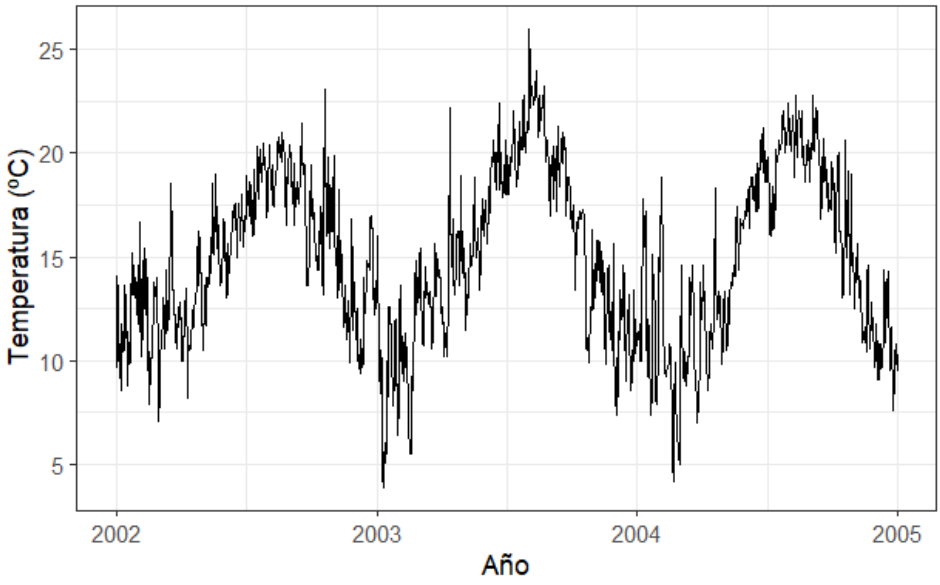

Figura 31. Ejemplo de los tres primeros años de la serie temporal.

Contrario a la tendencia general, la estacionalidad ejerce un impacto significativo en la serie, lo cual tiene repercusiones en la elección de los modelos apropiados, especialmente aquellos que consideran esta componente de manera relevante. Por esta razón, entre las opciones basadas en la teoría de series temporales, se consideraron apropiadas el método de *Seasonal Naive*, y los modelos de suavizado exponencial de Holt-Winters y SARIMA, ya que estas metodologías tienen la capacidad de poder modelar o captar la estacionalidad lo que permite realizar predicciones más precisas. En cuanto a los modelos de aprendizaje automático, dado que operan en un contexto diferente, se seleccionaron KNN y SVM como alternativas que podrían adaptarse a esta situación, debido a su capacidad para identificar patrones, incluyendo la estacionalidad.

4.2.1.3. Estacionariedad

Otra característica de la serie temporal que es necesaria estudiar es si una serie temporal es estacionaria o no porque muchos modelos de series temporales requieren que la serie sea estacionaria para que los resultados sean precisos.

Esta característica implica que las propiedades estadísticas de la serie temporal, como la media y la varianza, no cambian en el tiempo. Esto significaría que las observaciones en diferentes momentos tienen la misma distribución de probabilidad y que cualquier patrón que se observe en un momento, también será observable en otro.

Para saber si una serie temporal es estacionaria se pueden utilizar diferentes técnicas, pero dado que ninguna puede garantizar por si sola que una serie sea estacionaria, se utilizaron dos distintas para confirmarlo [16]. Las que se utilizaron son:

1. Análisis visual de la serie: se observa si hay una tendencia a largo plazo, una variación en la varianza a lo largo del tiempo, o si la serie presentan algún patrón de estacionalidad. La serie fluctúa alrededor de una media constante, pero con una ligera tendencia ascendente y, además, tiene una clara componente estacional, por lo que mediante este análisis no sería estacionaria.

2. Prueba de Dickey-Fuller aumentada (ADF): Es una prueba estadística donde se utiliza un contraste de hipótesis para evaluar si una serie temporal tiene una raíz unitaria o no, lo que indicaría si la serie es estacionaria. En esta prueba la hipótesis nula sería que no es estacionaria y la hipótesis alternativa que es estacionaria. El resultado obtenido se muestra en Tabla 4:

   Tabla 4. Test de Dickey- Fuller para la serie de la temperatura diaria.

Valor del estadístico Dickey-Fuller	p-valor
-2,1497	0,515

   Por lo que para un valor de significación de 0.05, no se tiene evidencia para rechazar la hipótesis nula por lo que se asume que la serie no es estacionaria.

   Ya que se sabe que es el patrón estacional lo que hace que la serie no sea estacionaria, se realiza una diferencia estacional para comprobar si, después de esta, es estacionaria. Los resultados obtenidos se muestran en la Tabla 5.

   Tabla 5. Test de Dickey- Fuller para la serie de la temperatura diaria diferenciada estacionalmente

Valor del estadístico Dickey-Fuller	p-valor
-4,2781	< 0,01

   Para un valor de significación de 0,05, se tiene evidencia para no rechazar la hipótesis nula por lo que se asume que la serie diferenciada es estacionaria.

Dado que las dos pruebas dieron como resultado que la serie no es estacionaria, se determinó que la serie no lo es. No obstante, se comprobó que era estacionaria al realizar una diferencia estacional, lo cual sirve para comprobar qué modelos que tienen esta diferenciación integrada (SARIMA) estarían usando esta transformación (1 diferenciación estacional).

Capítulo 5

Resultados

El objetivo de este capítulo es presentar la comparativa de los métodos y modelos, introducidos en el capítulo 2, evaluando su capacidad predictiva para la variable temperatura. Para ello se usó la metodología y métricas explicadas a continuación. Los métodos utilizados fueron los siguientes: *Seasonal Naive*, suavizado exponencial, SARIMA, KNN y SVM.

5.1. Metodología

La metodología utilizada para evaluar los métodos y poder realizar la comparativa fue la de dividir la base de datos en una parte de entrenamiento y otra de prueba. Esta técnica es utilizada muy a menudo tanto en modelos de series temporales como de *Machine Learning* para evaluar el rendimiento de un modelo [39].

Por un lado, la parte de entrenamiento es la porción de los datos que se utiliza para ajustar los modelos o entrenar los métodos, estimar los parámetros que modelizan estos datos o aprender las relaciones entre las variables. Una vez realizado este paso, con estos datos, se realizan las predicciones.

Por otro lado, los datos de prueba son la porción que se utiliza para evaluar el rendimiento del modelo, es decir, el error cometido, utilizando las predicciones realizadas por los datos de entrenamiento. Esto se hace midiendo el error que hay entre los valores reales (valores de los datos de prueba) y las predicciones del conjunto de entrenamiento. La idea es que los datos de prueba no hayan sido utilizados para entrenar el método o ajustar el modelo y, de esta manera, que estos no tengan información previa de estos datos para poder evaluar la capacidad predictora.

La proporción de cada parte varía, pero normalmente se encuentra en torno al 80% para los datos de entrenamiento y 20% para los datos de prueba. Así, el modelo tendrá información suficiente para entrenarse y datos suficientes para ser probado. En la Figura 32 se puede ver la división realizada de datos de entrenamiento y datos de prueba (la fecha del fin de los datos de entrenamiento es el 26 de mayo de 2016).

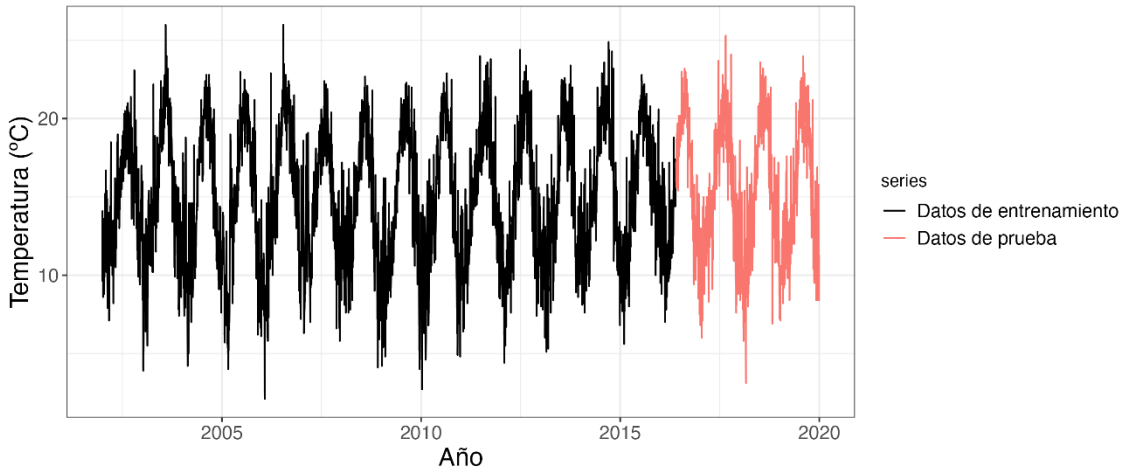

Figura 32. División de datos de entrenamiento y datos de prueba de la serie de datos.

5.2. Métricas usadas

Al construir un modelo de series temporales, es importante evaluar su capacidad para hacer predicciones. Para ello, se utilizan diferentes métricas de error que miden la diferencia entre los valores predichos con los datos de entrenamiento y los datos de prueba [40]. Algunos de estas métricas son las siguientes:

- Error medio (ME, por sus siglas en inglés, *Mean Error*): es la media aritmética de las diferencias entre los valores predichos y los valores reales de la serie. Se expresa en las mismas unidades que la serie temporal original por lo que facilita la interpretación. El inconveniente de esta métrica es que los errores positivos y negativos se compensan, afectando a la evaluación del error al contrarrestar discrepancias entre valores predichos y reales.

- Error absoluto medio (MAE, por sus siglas en inglés, *Mean Absolute Error*): es similar al ME, pero en este caso se utiliza la media de las diferencias en valor absoluto. Esto lo que consigue es evitar el inconveniente del ME, al hacer la media de valores absolutos.

- Error cuadrático medio (MSE, por sus siglas en inglés, *Mean Squared Error*): es la media aritmética de la diferencia entre los valores predichos y los reales al cuadrado. El MSE pondera más los errores grandes que los pequeños, lo que es útil si se desea penalizar más los errores graves. Un inconveniente de esta métrica es que las unidades estarán al cuadrado.

- Raíz del error cuadrático medio (RMSE, por sus siglas en inglés, *Root Mean Squared Error*): es la raíz cuadrada del MSE, y también se expresa en las mismas unidades que la serie temporal original. El RMSE es bastante popular ya que combina las ventajas del MAE y el MSE (que esté el error en las mismas unidades y que penalicen más los errores graves).
- Error porcentual medio (MPE, por sus siglas en inglés, *Mean Percentage Error*): es una medida relativa que escala el ME para que esté en unidades porcentuales en vez de las unidades de la variable. La principal ventaja es que le permite comparar varianzas entre datos de diferentes escalas, pero un inconveniente es que, al igual que el ME, se compensan errores positivos y negativos.
- Error porcentual absoluto medio (MAPE, por sus siglas en inglés, *Mean Absolute Percentage Error*): es una métrica relativa que transforma el MAE en unidades de porcentaje en lugar de unidades de la variable. La principal ventaja del MAPE es que utiliza valores absolutos para evitar la compensación entre errores positivos y negativos. Además, dado que el MAPE es una medida expresada en porcentaje, puede resultar más comprensible que otras métricas de error. Por ejemplo, si el MAPE es de 5, significa que, en promedio, el pronóstico se desvía en un 5%.

Estos errores son solo algunos ejemplos de las métricas que se pueden utilizar para evaluar modelos de series temporales y con ello poder compararlos y seleccionar el más adecuado. Para más información ver [40].

5.3. Desarrollo de los métodos y los modelos y predicciones realizadas

En este apartado se explicarán los modelos obtenidos junto con la representación gráfica de las predicciones realizadas para los datos de prueba. Para una mejor visualización de los datos se consideró mostrar el mismo número de ellos para la serie de datos de entrenamiento que para la serie de datos de prueba junto con las predicciones, es por eso por lo que los gráficos muestran una serie más reducida.

5.3.1. *Seasonal Naive*

Como se puede ver en la Figura 33, se puede observar como la metodología empleada para estas predicciones es simplemente repetir los valores del último periodo, es decir, los valores diarios del último año, ya que la estacionalidad de esta serie de datos es de 365. De esta manera, la predicción realizada por este modelo fue simplemente realizar "copias" del último año.

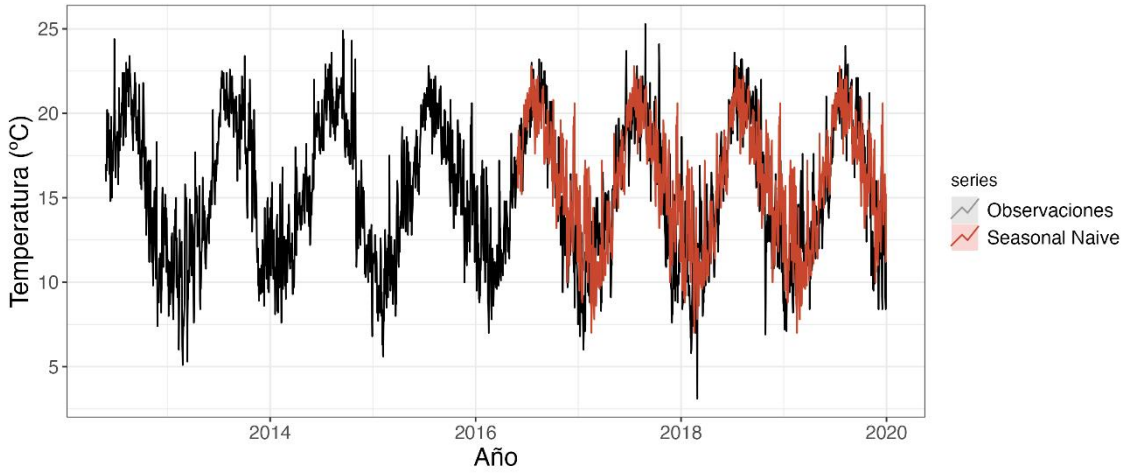

Figura 33. Predicciones realizadas con el método de Seasonal Naive.

5.3.2. Modelos de suavizado exponencial

Para estos métodos, al tener los datos una clara estacionalidad y ser esta una componente muy relevante en la serie, solo se ha tenido en cuenta el método de Holt-Winters, al ser este el único de los tres introducidos (suavizado exponencial simple, Holt y Holt-Winters) que cuenta con un parámetro para incorporarla en el modelo.

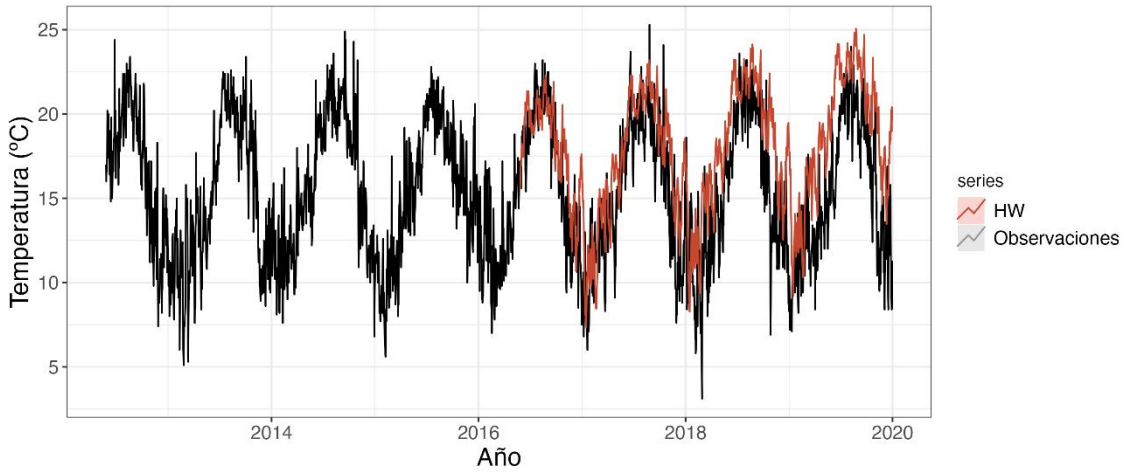

Figura 34. Predicciones realizadas con los distintos modelos de suavizado exponencial.

En la Figura 34 se muestran las predicciones realizadas por el modelo Holt-Winters. En estas se pueden ver como el modelo incorpora la tendencia (en este caso aditiva, al ser lineal), y que, al tener más en cuenta los últimos años, (más pronunciada en comparación a los primeros años de la serie como se observó en el gráfico de la tendencia en las componentes de la serie temporal) la pendiente es más ascendente que la de la serie de datos original. También, el modelo, se puede observar que es capaz de captar la estacionalidad presente en la serie temporal.

5.3.3. SARIMA

El modelo obtenido que mejor se ajusta a la serie de datos fue el siguiente: ARIMA(1,0,0)(0,1,0)[365].

Los valores obtenidos para los componentes del modelo SARIMA indican que solo se está utilizando un orden autorregresivo, es decir, que solo se ha utilizado un término en el modelo para modelar la autocorrelación de la serie. Además, se realiza una diferencia estacional de orden 1, lo cual era de esperar debido a que los datos son estacionales, con el periodo que muestra en el modelo (S = 365). Las predicciones obtenidas por este modelo se muestran en la Figura 35.

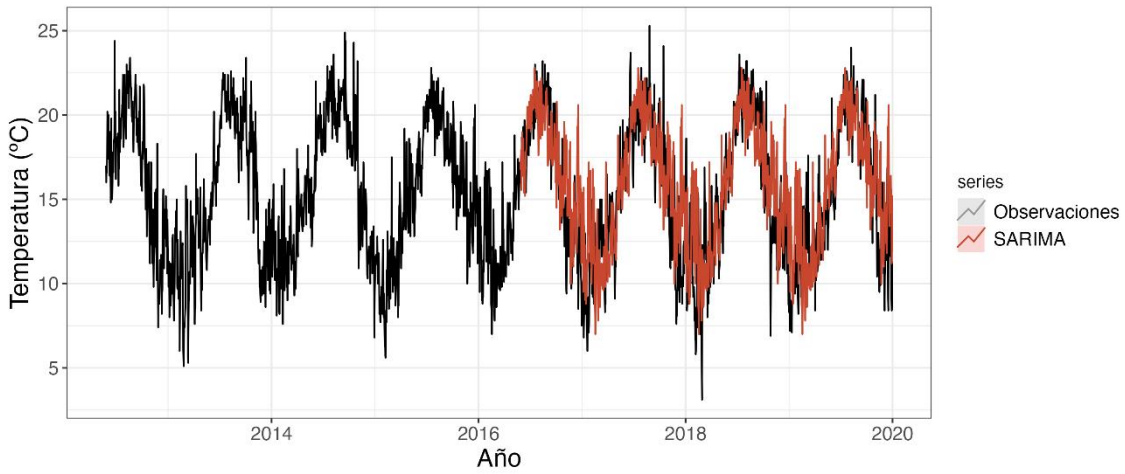

Figura 35. Predicciones realizadas por el modelo obtenido de SARIMA.

5.3.4. KNN

Para determinar el número óptimo de vecinos (k) que se utilizarían en el modelo de KNN se utilizó la técnica del codo (conocida como *elbow method* en inglés) [41]. Esta técnica consiste en trazar en una gráfica el error cometido en función de k. Con esta, se intenta encontrar el equilibrio entre el coste computacional innecesario (añadir vecinos que no aportan información y que pueden, incluso, llevar a sobreajuste) y el error (para este caso fueron utilizados ME, RMSE, MAE y MAPE). El k escogido fue 29 ya que como muestra la Figura 36, minimiza el error en ME y para RMSE, MAE y MAPE es una buena opción ya que tiene un error lo suficientemente bajo y aumentando el número de vecinos no se encuentra una diferencia sustancial (y aumentarlo conllevaría un mayor coste computacional). Por otro lado, el número de retardos utilizado fue el de la frecuencia, es decir 365.

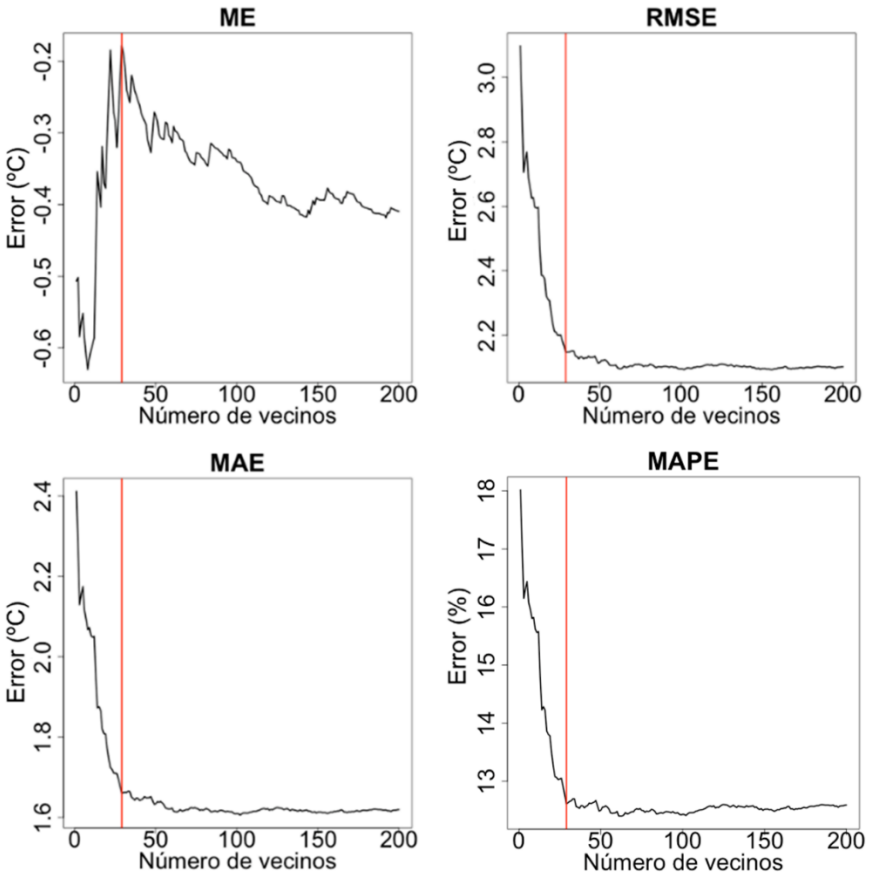

Figura 36. Optimización del número de vecinos para la elección de k en el modelo.

Las predicciones realizadas con este modelo se pueden observar en la Figura 37 que, sí se ajustan al patrón de estacionalidad, pero que parecen producir predicciones más suavizadas y estables, además de no enfocarse en capturar variaciones diarias específicas, al contrario que los modelos usados hasta el momento. Esto puede ser algo positivo

porque estaría haciendo una predicción con un patrón general de como varían los datos, en vez de una predicción diaria en base a la temperatura exacta que puede hacer un día, lo cual puede cometer un menor error medio.

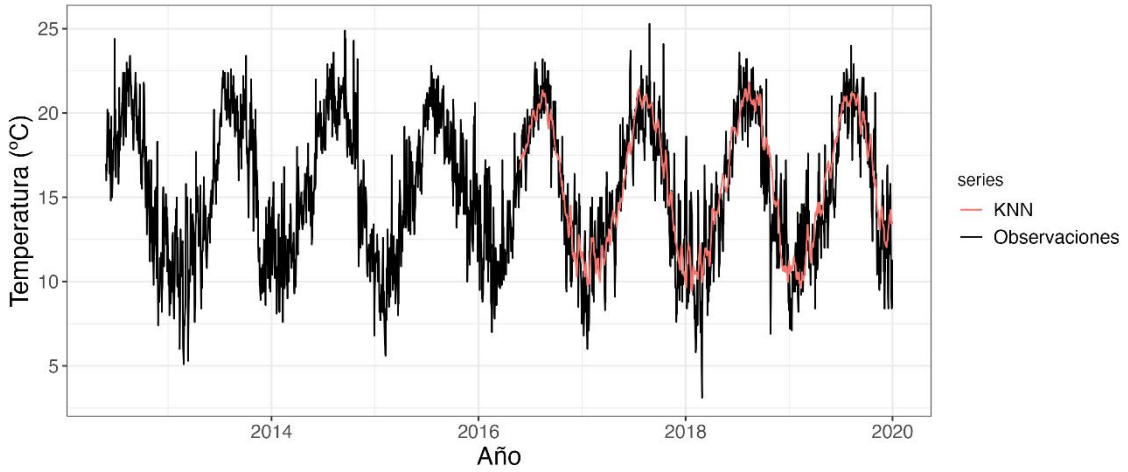

Figura 37. Predicciones realizadas por el modelo obtenido de KNN.

5.3.5. SVM

Para el modelo de SVM se siguió la metodología descrita en el capítulo 2 de construir variables *dummy*. Estas contendrían información de cada momento del tiempo y así poder reducir el problema a un caso de SVM para regresión para un mejor funcionamiento. En este caso se crearon variables *dummy* para los meses y para los días del año, es decir, si es enero, la variable vale 1 y el resto de las variables *dummy* para los meses cero, si es febrero, la variable *dummy* para febrero vale 1 y el resto 0 y así sucesivamente. Del mismo modo con las variables de los días del año, si es 1 de enero, la variable 1 de enero vale 1 y el resto de 364 variables para los días valen cero.

Como se puede ver en la Figura 38, las predicciones realizadas con SVM, al igual que en KNN, se ajusta a la estacionalidad de la serie y también tiende a realizar las predicciones más suavizadas y estables y tampoco parece enfocarse en capturar las variaciones diarias. Como se comentó con KNN, al estar prediciendo los datos de una manera más genérica siguiendo patrones que el modelo haya encontrado, también este podría estar generando menos error medio.

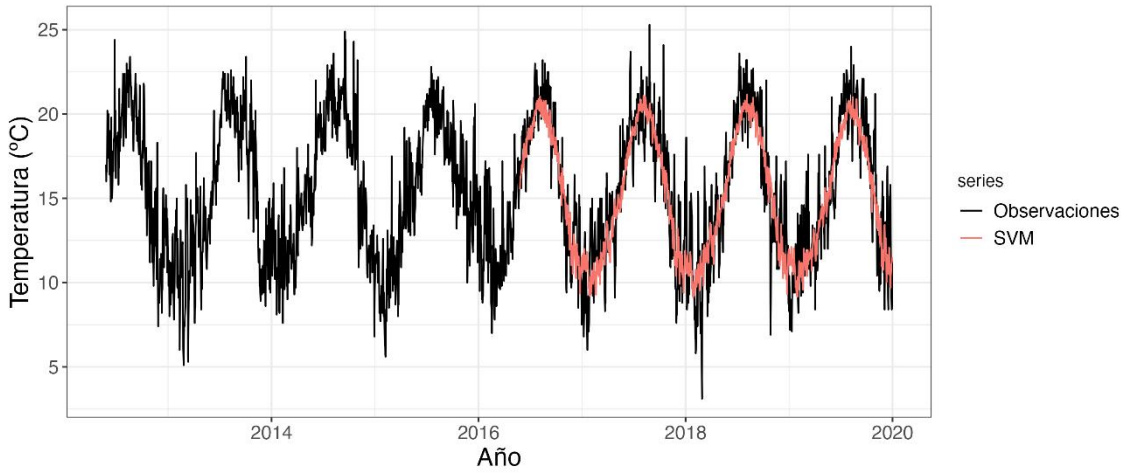

Figura 38. Predicciones realizadas por el modelo obtenido de SVM.

5.4. Comparativa de los métodos según error

En este apartado se mostrará la comparativa de los modelos, pero desde un punto de vista numérico. Para ello, se usaron las métricas descritas anteriormente (ME, MAE, RMSE y MAPE) mencionadas para las predicciones de cada modelo calculadas con respecto a los datos de prueba. Es necesario indicar que las métricas siguen todas la misma evaluación que es, cuanto más cerca esté de cero el error, el modelo se considera que funciona mejor, que es más preciso.

Tabla 6. Tabla de los errores calculados para los distintos modelos

	ME	RMSE	MAE	MAPE
Seasonal Naive	-0,373	2,969	2,242	17,559
SARIMA	-0,370	2,969	2,241	17,552
Holt-Winters	-2,379	3,659	2,911	23,090
KNN	-0,179	2,147	1,661	12,609
SVM	0,324	2,077	1,606	11,900

Si se observa la Tabla 6 de los resultados de los cálculos de los distintos errores para los distintos modelos, estos muestran que, si se tiene en cuenta el ME, el modelo que estaría cometiendo menor error sería el de KNN, mientras que para los otros tres el modelo que estaría haciendo una predicción con menor error medio sería el de SVM. Esto coincide con lo que se comentó en el apartado de las predicciones, para un conjunto de una gran cantidad de datos como es el de la temperatura media diaria de una región, podría ser más interesante utilizar metodologías que identifiquen patrones en vez de intentar predecir el dato exacto de cada día, si lo que se está haciendo es una predicción a largo plazo [26].

Adicionalmente, la Figura 39., muestra una representación más visual de estos errores donde las mayores diferencias se pueden detectar en el ME. Si se tiene en cuenta el significado de estos errores, el ME (que está en las mismas unidades que la serie temporal) estaría indicando que el método Holt-Winters cometería un error medio de 2.4 grados centígrados de diferencia con el valor original, mientras que el método de KNN de 0.18 grados de diferencia. Si se tiene en cuenta el valor absoluto, en MAE, el modelo de suavizado exponencial de Holt-Winters estaría cometiendo el doble de error que SVM, un error que, al estar en las mismas unidades que la serie, es de 3 grados aproximadamente. En cuanto a RMSE, el mayor error es de aproximadamente 3.6 grados por parte de Holt-Winters y, de igual manera, es algo menos que el doble del error cometido por SVM. Por otro lado, MAPE, al ser una medida relativa del error, estaría indicando que con el método de Holt-Winters se estaría cometiendo casi un 23 % más de error que con el de SVM que cometería un 10% de error.

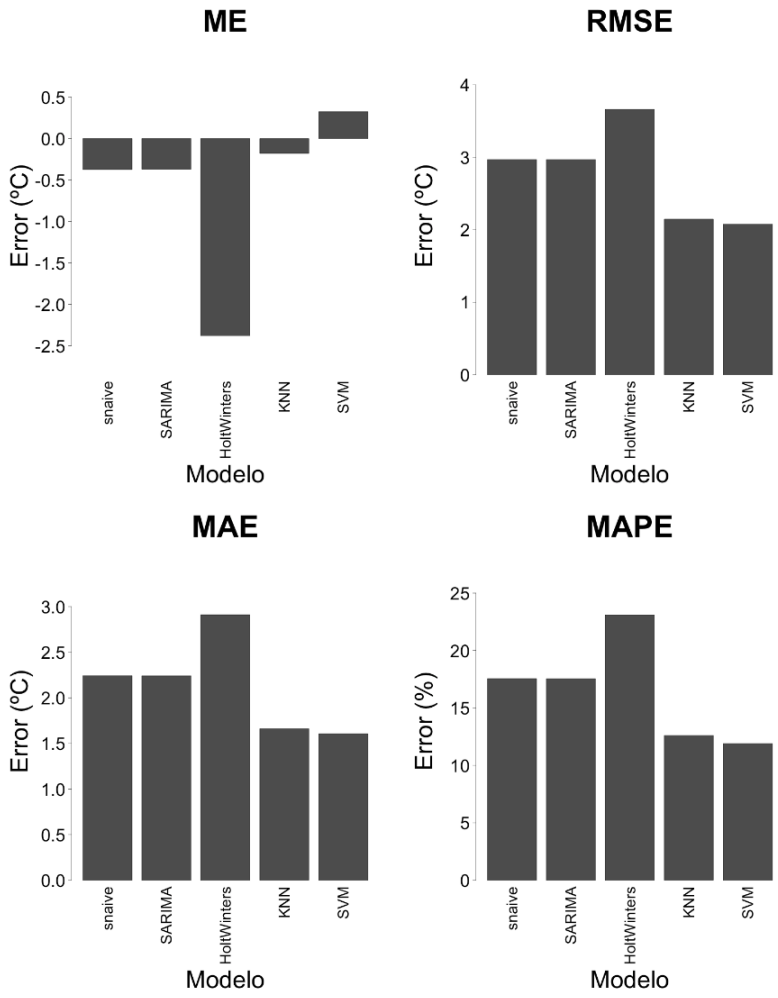

Figura 39. Representación gráfica de los errores calculados para los distintos modelos.

5.5. Discusión de resultados

Los resultados del estudio indican que los modelos de SVM y KNN tuvieron mejor desempeño en la predicción de la temperatura diaria en comparación al resto de los modelos. Seguidos de estos, los modelos *Seasonal Naive* y SARIMA, tuvieron resultado similares, aunque con un error ligeramente superior. Finalmente, el modelo Holt-Winters arrojó los errores más altos del estudio con un desempeño peor que los otros 4 modelos.

En el caso de SVM, el modelo tiene la capacidad de manejar datos no lineales y no paramétricos, es decir, que no asume una forma específica para la relación entre las variables, en lugar de esto utiliza la función kernel para construir el modelo. Esta capacidad es útil en situaciones en las que la variable a predecir tiene un comportamiento complejo y no se puede modelar fácilmente, encontrando patrones distintos a los que utilizarían los métodos clásicos [42].

Por otro lado, KNN se basa en la idea de que los puntos de datos similares tienden a estar cerca uno del otro y es esta metodología la que puede hacer que desempeñe mejores resultados en situaciones en las que los datos tienen una estructura de agrupación clara (los días del año en este caso). Además, en comparación con los modelos de series temporales, este modelo tampoco asume una estructura específica para los datos y puede ser más flexible [43].

En cuanto a los resultados de los métodos de SARIMA y *Seasonal Naive*, SARIMA por una parte es un modelo estadístico que puede manejar diferentes tipos de estacionalidad (aditiva o multiplicativa) y diferentes grados de esta además de ruido. Aunque esta cualidad puede ser más efectiva cuando la estacionalidad es más compleja o varía a lo largo del tiempo, no es de extrañar que haga predicciones bastante acertadas sobre la temperatura diaria. El método de *Seasonal Naive*, por otra parte, es un método simple que utiliza el valor observado de la misma época del año anterior para hacer la predicción, por lo que, al tener la serie de la temperatura diaria una estacionalidad bien definida y estable, y no tener una tendencia muy pronunciada, el método *Seasonal Naive* es también es buena opción [44].

En último lugar, Holt-Winters es un método que se basa en supuestos más simples acerca de la estructura de la serie temporal y puede que sea más efectivo para series temporales donde exista tendencia más relevante. Es por eso por lo que se puede ver como las predicciones adoptaron una tendencia más marcada que la existente, traduciéndose

eso en generar más error [44]. Por otro lado, se puede observar cómo capta la estacionalidad presente en la serie temporal al ser un modelo que, al igual que SARIMA, puede manejar distintos tipos de estacionalidad (multiplicativa o aditiva). Aun habiendo sido el que peor resultado ha tenido, si tomamos por ejemplo el MAPE, el error cometido es del 23% lo cual entra dentro de los parámetros de un modelo aceptable, o una diferencia de 3,6 grados en el caso del RMSE.

En resumen, los modelos de SVM y KNN pueden arrojar mejores resultados a la hora de predecir la temperatura diaria que los modelos de predicción de series temporales como *Seasonal Naive*, SARIMA y Holt-Winters en situaciones en las que la relación entre las variables es compleja o no se puede modelar fácilmente con un enfoque de series temporales. Sin embargo, es importante tener en cuenta que cada modelo tiene sus limitaciones y no es adecuado para todas las situaciones. Por lo tanto, es conveniente tener la perspectiva de varios modelos y seleccionar el que mejor se ajuste a la naturaleza del problema.

Capítulo 6

Predicción a futuro de la temperatura en Gijón

Una vez seleccionado el modelo SVM como el más adecuado para la predicción de la temperatura diaria, resulta interesante llevarlo a la práctica. Cabe destacar que, para estas predicciones, no es posible comprobar la precisión debido a la falta de datos futuros en el momento en el que se hicieron. Aun teniendo esto en cuenta, es importante resaltar que este tipo predicciones son realmente útiles.

En primer lugar, realizar predicciones a futuro nos permite tener una visión anticipada de las tendencias climáticas en los próximos meses. Aunque estas predicciones pueden tener limitaciones, como la falta de datos precisos y la exclusión de factores externos relevantes, nos brindan una idea general de lo que podríamos esperar en términos de cambios en la temperatura. Esto nos permite estar mejor preparados y adaptarnos a las posibles condiciones climáticas que puedan influir en nuestras actividades diarias [45].

Es cierto que estas predicciones son modestas en comparación con las predicciones climáticas que se realizan teniendo en cuenta más variables o con modelos más complejos. Pero, aun así, ofrecen beneficios prácticos y pueden ayudarnos a tomar decisiones informadas. Estas, ofrecen una perspectiva general de lo que podría suceder en términos de condiciones climáticas, ofreciendo una visión anticipada para los próximos meses. Esta información resulta valiosa para planificar nuestras actividades diarias, especialmente aquellas que están directamente influenciadas por la temperatura.

6.1. Entrenamiento del método y predicción

Para entrenar el método se utilizaron todos los datos disponibles en el momento de la recopilación de estos, es decir, se tomaron desde el 1 de enero de 2002 hasta la fecha más reciente de la base de datos original, 22 de febrero de 2023. Después, se realizaron predicciones para los próximos meses finalizando el 15 de septiembre de 2023. Los resultados de las predicciones están representados en la Figura 40, en esta, se encuentran representados los datos disponibles hasta el momento en negro y las predicciones en rojo. Para una mejor visualización de los datos junto con las predicciones, solo se graficaron los últimos dos años.

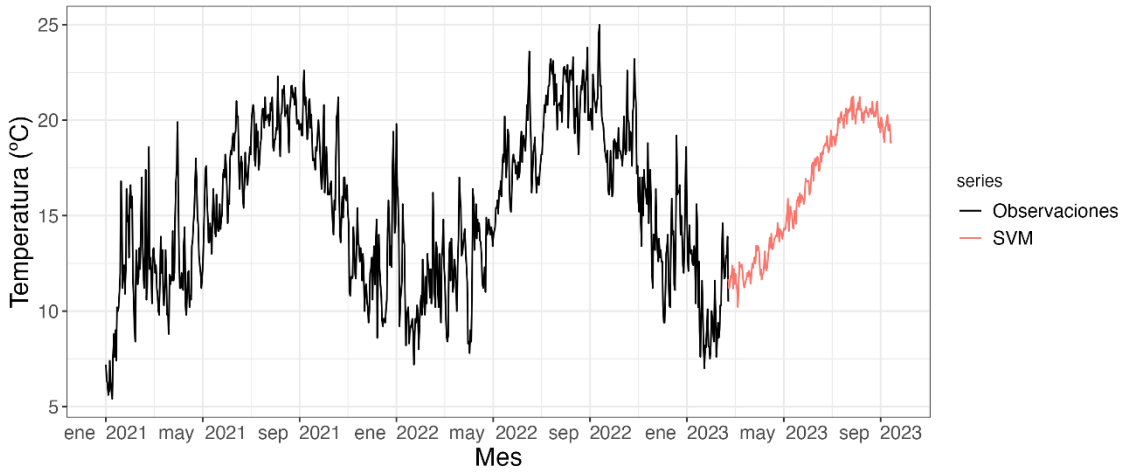

Figura 40. Resultado de las predicciones del modelo de SVM con todos los datos disponibles.

Al igual que se observaba para los datos de prueba, el modelo capta la estacionalidad adecuadamente. También, las predicciones son más suavizadas que los datos originales, siendo datos más estables en lugar de enfocarse en capturar variaciones diarias específicas.

Esto, si se toma julio y agosto como ejemplo (unos meses en los que parece mantenerse la temperatura), se puede ver que en los años anteriores se obtenía una media de 20.23ºC para los datos originales en estos meses y una media de 20.04ºC para las predicciones en el mismo marco temporal. Por otro lado, la varianza deja un valor de 1.36ºC para los datos originales y de 0.64ºC en las predicciones, valor que, aun siendo el doble para los datos originales, no resulta tan disparatado si se tienen en cuenta las variaciones diarias que puede haber en la temperatura diaria.

En resumen, SVM es una buena opción para aplicarlo en predecir la temperatura diaria de una ciudad a corto plazo. Capta la estacionalidad y realiza predicciones con menos varianza que los datos reales, pero que pueden ser igual de válidas si lo que se necesita es saber qué temperaturas va a haber en los próximos meses.

Capítulo 7

Conclusiones

En el presente trabajo se ha podido ver la versatilidad que existe a la hora de predecir la temperatura diaria de una ciudad, ya sea con teoría de series temporales o con modelos de aprendizaje automático. Gracias a las herramientas estadísticas para series temporales se puede obtener información de los datos existentes como, por ejemplo, la tendencia, la estacionalidad o la estacionariedad; y, gracias a las distintas metodologías existentes para modelizar estos datos, predecir valores futuros con una precisión aceptable. Se tratan, por este motivo, de herramientas muy útiles en un campo en continuo desarrollo, que tienen un impacto considerable en nuestras vidas.

A lo largo de este trabajo se ha recopilado literatura existente en lo relativo al campo de las series temporales. Partiendo del concepto de serie temporal, se han explicado las distintas componentes y teoría necesaria para introducir el método de *Seasonal Naive* y los modelos de suavizado exponencial y de SARIMA.

Una vez explicados los modelos apoyados en un enfoque más clásico, se introdujeron las técnicas de *Machine Learning* que abordan el problema desde un punto de vista diferente. Se estudiaron los métodos KNN y SVM, que originalmente eran aplicados a problemas de clasificación y se adaptaron manteniendo su metodología para aplicarlos a la predicción de series temporales.

Utilizando la base de datos climatológicos de la AEMET para la ciudad de Gijón, se realizó una limpieza y un preprocesamiento de sus variables (insolación, presión, temperatura y precipitaciones), junto con un tratamiento de los datos faltantes. Se procedió a realizar un análisis exploratorio de los datos, incluyendo un estudio descriptivo y de correlaciones.

Finalmente, utilizando de los datos diarios de la temperatura media en Gijón, se realizó un estudio comparando la capacidad predictora de los métodos vistos en el trabajo con esta serie temporal. Se dividió la serie en 80 % de los datos para entrenar los modelos y 20 % para evaluar las predicciones.

Los resultados de dicho estudio sugieren que, los modelos de KNN y SVM, según las métricas usadas para evaluar el rendimiento (ME, RMSE, MAE, MPE), habrían

cometido menos error a la hora de predecir los datos de prueba, seguidos de estos, estarían los modelos de *Seasonal Naive* y de SARIMA; y finalmente el método de Holt-Winters cometiendo un error superior al resto de los métodos y modelos estudiados.

Los resultados obtenidos indican que los métodos de *Machine Learning*, aunque con poca diferencia, podrían ser más adecuados que los vistos de modelos clásicos de series temporales. Esto es debido a que identifican patrones y hacen sus predicciones en base a ellos, en vez de intentar predecir la temperatura diaria exacta. Esto se traduce en una menor variabilidad diaria de la temperatura que se pudo observar en los datos.

Aunque esto de primeras puede ayudar a escoger un modelo o método, no hay que olvidar que cada metodología tiene sus ventajas y desventajas. Es importante tener en cuenta la naturaleza de cada metodología a la hora utilizar los modelos o métodos estudiados y que lo más adecuado para este tipo de problemas es combinar metodologías para tener distintos puntos de vista de cuál podría ser el valor correcto de la predicción.

Por último, con el objetivo de implementar una aplicación práctica con el modelo resultante, se tomaron todos los datos disponibles hasta el momento y se realizó una predicción para los próximos meses. Esta predicción concluyó que, SVM, también puede ser un buen modelo o método si se quiere aplicar para predecir la temperatura a corto plazo.

Futuras líneas de investigación

A pesar de los resultados prometedores obtenidos en este estudio, existen varias áreas que podrían explorarse en futuros trabajos para mejorar aún más la predicción de la temperatura diaria utilizando modelos de series temporales y métodos de aprendizaje automático. Algunas posibles direcciones para investigaciones futuras son:

1. Exploración de otras técnicas de *Machine Learning*: En este estudio, se utilizaron los algoritmos de KNN y SVM para predecir la temperatura diaria. Sin embargo, existen numerosas técnicas adicionales de aprendizaje automático que podrían considerarse, como redes neuronales, bosques aleatorios o *gradient boosting*. Comparar el rendimiento de estos algoritmos con los utilizados en este estudio podría proporcionar información adicional sobre cuál es la mejor técnica para el problema de predicción de temperatura diaria.
2. Incorporación de más variables predictoras: En este trabajo, se utilizó únicamente la serie temporal de la temperatura media diaria como variable predictora. Sin embargo, la temperatura diaria puede estar influenciada por una variedad de factores, entre otras las variables tratadas en este trabajo. Incorporar estas variables adicionales como predictoras en el modelo (ya preparadas gracias a este trabajo) podría mejorar la precisión de las predicciones.
3. Estudio de modelos híbridos: En lugar de considerar solo enfoques basados en series temporales o en aprendizaje automático, se podrían explorar modelos híbridos que combinen ambas metodologías. Por ejemplo, se podría considerar utilizar un enfoque de aprendizaje automático para capturar patrones y tendencias generales, y luego utilizar técnicas de series temporales para modelar la estacionalidad y la estacionariedad de los datos.
4. Análisis de datos de diferentes ubicaciones: Este estudio se centró en la predicción de la temperatura diaria en la ciudad en Gijón. Sin embargo, sería interesante comparar los resultados obtenidos en diferentes ubicaciones geográficas. Diferentes regiones pueden tener patrones climáticos distintos, por lo que sería relevante analizar si los modelos utilizados en este estudio mantienen su rendimiento en diferentes contextos climáticos.

En resumen, hay varias líneas abiertas para realizar futuros trabajos en el campo de la predicción de la temperatura diaria. Explorar diferentes técnicas de aprendizaje automático, incorporar más variables predictoras, desarrollar modelos híbridos, analizar datos de diferentes ubicaciones y evaluar el rendimiento en escalas de tiempo más largas pueden contribuir a un mejor entendimiento y predicción del clima diario.

Bibliografía

[1] De Gooijer, J. G., & Hyndman, R. J. (2006). 25 years of time series forecasting. *International Journal of Forecasting*, *22*(3), 443–473.

[2] Gorbatiuk, K., Hryhoruk, P., Proskurovych, O., Rizun, N., Gargasas, A., Raupelienė, A., & Munjishvili, T. (2021). Application of fuzzy time series forecasting approach for predicting an enterprise net income level. E3S Web of Conferences, 280.

[3] Chimmula, V. K. R., & Zhang, L. (2020). Time series forecasting of COVID-19 transmission in Canada using LSTM networks. Chaos, Solitons & Fractals, 135.

[4] Piwowar, J. M., & Ledrew, E. F. (2002). ARMA time series modelling of remote sensing imagery: A new approach for climate change studies. *International Journal of Remote Sensing*, *23*(24), 5225–5248.

[5] Murphy, A. H. (1993). What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Weather and Forecasting*, *8*(2), 281–293.

[6] Gneiting, T., & Raftery, A. E. (2005). Weather Forecasting with Ensemble Methods. *Science*, *310*(5746), 248–249.

[7] Joshi, H., & Tyagi, D. (2021). Forecasting and Modeling Monthly Rainfall in Bengaluru, India: An Application of Time Series Models. *International Journal of Scientific Research in Mathematical and Statistical Sciences*, *8*(1), 39–46.

[8] Zubair Khan, M., Shamshad, B., & Zara, O. (2019). Modeling and Forecasting Weather Parameters using ANN-MLP, ARIMA and ETS model: A case study for Lahore, Pakistan. *International Journal of Scientific & Engineering Research*, *10*(4), 351–366.

[9] Ray, S., Das, S. S., Mishra, P., & Al Khatib, A. M. G. (2021). Time Series SARIMA Modelling and Forecasting of Monthly Rainfall and Temperature in the South Asian Countries. *Earth Systems and Environment*, *5*(3), 531–546.

[10] Mellit, A., Pavan, A. M., & Benghanem, M. (2013). Least squares support vector machine for short-term prediction of meteorological time series. *Theoretical and Applied Climatology*, *111*(1–2), 297–307.

[11] Rajagopalan, B., & Lall, U. (1999). A *k* -nearest-neighbor simulator for daily precipitation and other weather variables. *Water Resources Research*, *35*(10), 3089–3101.

[12] Brockwell, P. J., & Davis, R. A. (2016). *Introduction to Time Series and Forecasting*. Springer International Publishing.

[13] Rodriguez Morilla, C. (2000). *Analisis de Series Temporales*. La Muralla.

[14] R Core Team. (2023). *A Language and Environment for Statistical Computing*. https://www.R-project.org.

[15] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

[16] Little, T. D. (2013). *The Oxford handbook of quantitative methods in psychology: Vol. 2: statistical analysis* (Vol. 2). OUP USA.

[17] Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

[18] Brown, R. G. (1959) Statistical forecasting for inventory control, Nueva York: McGraw-Hill

[19] Brown, R. G. (1963) Smoothing, forecasting and prediction of discrete time series, Englewood Cliffs: Prentice-Hall

[20] Holt, C. C. (1957). Forecasting seasonals and trends by exponentially weighted moving averages. *ONR Memo*, *52*, 5–10.

[21] De Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. Journal of the American Statistical Association, 106(496), 1513–1527.

[22] Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, *6*(3), 324–342.

[23] Nassar, S., SCHWARZ, K. P., EL-SHEIMY, N. A. S. E. R., & Noureldin, A. (2004). Modeling inertial sensor errors using autoregressive (AR) models. Navigation, 51(4), 259-268.

[24] Peña, D. (2010). *Análisis de series temporales*. Alianza.

[25] Vagropoulos, S. I., Chouliaras, G. I., Kardakos, E. G., Simoglou, C. K., & Bakirtzis, A. G. (2016, April). Comparison of SARIMAX, SARIMA, modified SARIMA and ANN-based models for short-term PV generation forecasting. In 2016 IEEE international energy conference (ENERGYCON) (pp. 1-6). IEEE.

[26] Fawzy, H., Rady, E. H. A., & Abdel Fattah, A. M. (2020). Comparison between support vector machines and k-nearest neighbor for time series forecasting. *J. Math. Comput. Sci.*, *10*(6), 2342–2359.

[27] Silverman, B. W., & Jones, M. C. (1989). E. fix and jl hodges (1951): An important contribution to nonparametric discriminant analysis and density estimation: Commentary on fix and hodges (1951). *International Statistical Review/Revue Internationale de Statistique*, 233-238.

[28] Martínez, F., Frías, M. P., Charte, F., & Rivera, A. J. (2019). Time Series Forecasting with KNN in R: the tsfknn Package. *R J.*, *11*(2), 229.

[29] Sain, S. R. (1996). The nature of statistical learning theory.

[30] Velásquez, J. D., Olaya, Y., & Franco, C. J. (2010). Predicción de series temporales usando máquinas de vectores de soporte. *Ingeniare. Revista Chilena de Ingeniería*, *18*(1), 64–75.

[31] Agencia Estatal de Meteorología. (n.d.). *AEMET*. Retrieved July 1, 2023, from https://www.aemet.es/es/datos_abiertos/AEMET_OpenData

[32] Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

[33] Martínez, E. J. A., & Soley, F. J. (2009). Descripción de dos métodos de rellenado de datos ausentes en series de tiempo meteorológicas. *Revista de Matemática: Teoría y Aplicaciones*, *16*(1), 60–75.

[34] Pigott, T. D. (2001). A review of methods for missing data. *Educational Research and Evaluation*, *7*(4), 353–383.

[35] Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: time series missing value imputation in R. *R J.*, *9*(1), 207.

[36] Devore, J. L. (2009). Probabilidad y estadística para ingeniería y ciencias. *Cengage Learning Editores*.

[37] Keune, J., Ohlwein, C., & Hense, A. (2014). Multivariate probabilistic analysis and predictability of medium-range ensemble weather forecasts. *Monthly Weather Review*, *142*(11), 4074–4090.

[38] Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, *27*, 1–22.

[39] Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python: a guide for data scientists*. " O'Reilly Media, Inc."

[40] Chan, K.-S., & Cryer, J. D. (2008). *Time series analysis with applications in R*. Springer.

[41] *Elbow Method in Supervised Machine Learning(Optimal K Value)*. (n.d.). Doumbia, M. Retrieved April 14, 2023, from https://medium.com/@moussadoumbia_90919/elbow-method-in-supervised-learning-optimal-k-value-99d425f229e7

[42] Woldemariam, W. (2022). A framework for transportation infrastructure cost prediction: A support vector regression approach. *Transportation Letters*, *14*(9), 997–1003.

[43] F. Chen, X. Chen y Y. Xie, A new K-nearest neighbor-based short-term traffic flow prediction method. *Journal of Ambient Intelligence and Humanized Computing,* vol. 12, nº 7, pp. 6931-6943, 2021.

[44] F. M. Tseng, Y. Y. Chen y C. W. Lin, «A comparison of time series forecasting models for hotel revenue managem ent,» *Journal of Hospitality and Tourism Management,* nº 31, pp. 141-149, 2017.

[45] Nadal, I. S., & Muñuzuri, V. P. (2006). *Fundamentos de meteorología* (Vol. 6). Univ Santiago de Compostela.