Escuela Superior
de Ingeniería y Tecnología
Universidad de La Laguna

Universidad de La Laguna

Trabajo de Fin de Grado

Grado en Ingeniería Informática

APLICACIÓN DE MACHINE LEARNING PARA LA PREDICCIÓN DE OCUPACIÓN HOTELERA

APPLICATION OF MACHINE LEARNING FOR HOTEL OCCUPANCY PREDICTION

XUEMEI LIN

La Laguna, Junio de 2024

Escuela Superior
de Ingeniería y Tecnología
Universidad de La Laguna

D. SERGIO DÍAZ GONZÁLEZ, profesor Titular de Universidad adscrito al Departamento de Ingeniería Informática y de Sistemas de la Universidad de La Laguna, como tutor.

C E R T I F I C A (N)

Que la presente memoria titulada:
"APLICACIÓN DE MACHINE LEARNING PARA LA PREDICCIÓN DE OCUPACIÓN HOTELERA"
ha sido realizada bajo su dirección por **Dña. XUEMEI LIN**, con **N.I.E.Y-1576473-K.** Y para que así conste, en cumplimiento de la legislación vigente y a los efectos oportunos firman la presente en La Laguna a 21 de noviembre de 2018.

Agradecimientos

En un abrir y cerrar de ojos, mi vida universitaria está a punto de acabar. Tras estos meses de estudio y esfuerzo, este trabajo de fin de grado también está a punto de terminar. Primero, quiero expresar mi profundo agradecimiento a mi tutor, Sergio Díaz González. Desde la primera reunión hasta la memoria final, mi tutor me ha brindado un gran apoyo y sus conocimientos para completar este proyecto de manera oportuna y exitosa. Durante el proceso de desarrollo, debido a mi falta de experiencia, mi tutor ha sido paciente al resolver todas mis dudas. Por lo tanto, quiero agradecerle otra vez.

En segundo lugar, quiero agradecer a mis padres. Gracias por vuestro constante apoyo, preocupación y ánimo en cualquier circunstancia.

También agradezco a todos los profesores que me han enseñado y ayudado durante mis años universitarios. Me habéis enseñado mucho y siempre habéis sido un modelo a seguir para mí.

Además, quiero agradecer a mis amigos por vuestra compañía y ayuda a lo largo de estos años. Espero que todos tengan un futuro brillante y que nos encontremos en la cima.

Finalmente, agradezco a la Universidad de La Laguna y al Departamento de Ingeniería Informática por proporcionar los recursos y el entorno adecuados para mi desarrollo académico y personal.

Gracias a todos por hacer posible este logro.

RESUMEN

En los últimos años, la inteligencia artificial ha avanzado a pasos agigantados, especialmente desde la última década del siglo XXI. Destaca particularmente el rápido desarrollo del aprendizaje automático (Machine Learning), es una rama de la inteligencia artificial que ha revolucionado diversos sectores gracias a su capacidad para analizar y aprender de grandes volúmenes de datos.

Este trabajo de fin de grado se centra en la utilización de diferentes técnicas de aprendizaje automático como ARIMA, SARIMAX y LSTM para predecir la ocupación diaria de un hotel en Tenerife, con el propósito de facilitar a los administradores hoteleros la toma de decisiones enfocada a la optimización de los recursos y el incremento de los beneficios del hotel.

Este trabajo se basa en datos históricos de ocupación hotelera, separando los datos en dos partes: datos de entrenamiento y datos de prueba. Estos datos, junto con variables adicionales como las temperaturas y días vacacionales, se utilizan para implementar y comparar los tres mencionados modelos de series temporales. Los modelos se evalúan con las técnicas de RMSE (Root Mean Squared Error) y Verificación de series de tiempo hacia adelante (Walk-Forward Validation).

Se desarrolla con el lenguaje de programación Python, utilizando diferentes combinaciones de parámetros. Tras la evaluación de modelos, se encuentra un modelo válido que ayuda la predicción de ocupación en el sector hotelero.

Palabras claves: Inteligencia artificial, Aprendizaje Automático (Machine Learning), Aprendizaje profundo (Deep Learning), Series temporales, Python, ARIMA, SARIMAX, LSTM, Turismo, Hoteles.

Abstract

In recent years, artificial intelligence has made significant strides, especially since the last decade of the 21st century. Particularly notable is the rapid development of machine learning, a branch of artificial intelligence that has revolutionized various sectors due to its ability to analyze and learn from large volumes of data.

This final degree project focuses on using machine learning techniques with three time series models, ARIMA, SARIMAX, and LSTM, to predict the daily occupancy of a hotel in Tenerife. The aim is to assist hotel managers in making decisions that optimize resources and increase hotel profits.

This project is based on historical hotel occupancy data, separating the data into two parts: training data and test data. These data, along with additional variables such as temperatures of Tenerife, United Kingdom and Germany, and holidays, are used to implement and compare the three mentioned time series models. The models are evaluated using RMSE (Root Mean Squared Error) and Walk-Forward Validation techniques.

The project is developed using the Python programming language, employing different combinations of parameters. After evaluating the models, find a valid model to help in predicting hotel occupancy.

key words: Artificial intelligence, Machine Learning, Deep Learning, Time series, Python, ARIMA, SARIMAX, LSTM, Tourism, Hotels.

Índice de contenidos

Índice Tablas

Índice Figuras

Capítulo 1 Introducción

En este trabajo se desarrollan y comparan diferentes modelos para predecir la ocupación diaria de un hotel, con el propósito de facilitar a los administradores hoteleros la toma de decisiones enfocada en la optimización de recursos e incremento de beneficios del hotel.

Los datos se dividirán en un 80% de datos de entrenamiento y un 20% de datos de prueba. El conjunto de datos se basará inicialmente en los datos de ocupación del hotel junto con variables adicionales como la temperatura de Tenerife, días festivos, entre otras, para mejorar la precisión de la predicción.

Se utilizarán tres técnicas de machine learning: ARIMA, SARIMAX y LSTM. Estas tres técnicas se entrenarán utilizando diferentes combinaciones de parámetros para encontrar el modelo más adecuado.

En estas tres técnicas de machine learning, se utilizará el RMSE como medida de precisión comparativa y validación de avance en el tiempo (walk forward validation).

El desarrollo se llevará a cabo utilizando el lenguaje de programación Python, empleando librerías como pandas, numpy, matplotlib, statsmodels entre otras, para el manejo de datos, implementación de modelos y visualización de datos y resultados.

1.1 Formulación del problema

La utilización de las estrategias de marketing pueden ayudar a los hoteles a ganar prestigio y obtener mayores ingresos. Una de las estrategias de marketing para el sector hotelero es la optimización de los ingresos. El marketing no solo atrae a más clientes, sino que también ayuda a maximizar los ingresos. Estrategias como la gestión de ingresos y la segmentación de mercado permiten ajustar precios y promociones según la demanda y el comportamiento del cliente.

Un estudio de McKinsey & Company muestra que la personalización en el marketing puede aumentar los ingresos en un 5-15% y la eficiencia del gasto en marketing en un 10-30%.

Para implementar esta estrategia, los gerentes de hoteles deben vender el producto correcto, al precio correcto, al cliente correcto y en el momento correcto. La base de esta

estrategia es establecer segmentos de clientes, definir precios adecuados para cada segmento y, a través de la predicción de la demanda, determinar los precios convenientes, a quién vender y cuándo hacerlo. Por ejemplo, si la predicción indica que la ocupación será alta, la estrategia probablemente será aumentar el precio. Si la ocupación es baja, seguramente se necesitará bajar el precio para promover las ventas.

Si la predicción no es correcta, existe el riesgo de vender a un precio demasiado bajo o de no realizar una venta. Por lo tanto, para evitar esta situación, es importante predecir la tasa de ocupación de manera precisa, lo que permitirá vender de manera adecuada y maximizar los beneficios del hotel.

1.2 Justificación

La industria del turismo da gran importancia para muchos países y ciudades en todo el mundo, y el sector hotelero desempeña un papel clave también. Su importante no solo radica en el ámbito del empleo, ya que promueve la creación de numerosos puestos de trabajo en áreas como recepción, mantenimiento, gestión, limpieza, entre otros, sino también en el fomento del desarrollo económico del país, beneficiando a restaurantes, bares, tiendas, lugares de entretenimiento, etc.

Por lo tanto, el uso de herramientas y estrategias adecuadas en la toma de decisiones del sector hotelero puede contribuir de manera positiva al desarrollo social y económico del país.

En el proceso de toma de decisiones, es necesario realizar pronósticos más precisos sobre la tasa de ocupación de los hoteles para poder establecer estrategias de precios adecuadas que impulsen las ventas.

En este trabajo, utilizaré diferentes modelos de series temporales y analizaré cuál es el modelo más adecuado para predecir la ocupación hotelera. De esta manera, los administradores de hoteles podrán utilizar los resultados del modelo para tomar decisiones, maximizando los ingresos y planificando los recursos, convirtiéndose así en una herramienta para mejorar la competitividad del sector en la ciudad.

1.3 Objetivos

1.3.1 General

Desarrollar y validar un modelo de Machine Learning para predecir la ocupación diaria de un hotel.

1.3.2 Específicos

1. Encontrar las técnicas de Machine Learning que pueden ser aplicadas a la predicción de series temporales.
2. Recopilar datos y realizar ingeniería de características para construir un conjunto de datos para entrenamiento y pruebas.
3. Realizar una comparativa de las diferentes técnicas de Machine Learning escogidas e identificar cuál ofrece mejores resultados en la predicción de la tasa de ocupación en el sector hotelero.

Capítulo 2 Estado del arte

2.1 Inteligencia Artificial y Aprendizaje automático

El machine learning (ML), conocido en español como aprendizaje de máquina, es una disciplina de la ciencia de la computación y una rama de la inteligencia artificial que tiene por objetivo desarrollar sistemas que aprenden automáticamente, reconocen patrones y predicen comportamientos, a partir de conjuntos de datos [1].

Los problemas que se pueden resolver mediante ML se dividen en dos clases principales:

- **Aprendizaje supervisado (Supervised Learning)**
- **Aprendizaje no supervisado (Unsupervised Learning)**

En el **aprendizaje supervisado**, disponemos de un conjunto de datos con respuestas conocidas. El objetivo es aprender cómo llegar a esas respuestas a partir de los datos. Este tipo de aprendizaje se divide en **regresión** y **clasificación**. La **regresión** predice resultados continuos, como estimar la edad de una persona a partir de una foto. La **clasificación** predice resultados discretos, como determinar si un tumor es maligno.

En el **aprendizaje no supervisado**, los datos están disponibles sólo en forma de entrada sin una variable de salida correspondiente. Estos algoritmos modelan los patrones subyacentes en los datos para aprender más sobre sus características. Una de las principales técnicas de aprendizaje no supervisado es el clustering, que descubre grupos inherentes en los datos y los utiliza para predecir salidas para entradas no vistas.

2.2 Historia de Machine Learning

La historia del Machine learning se remonta a mediados del siglo XX cuando Walter Pitts y Warren McCulloch intentaron modelar matemáticamente las redes neuronales humanas en 1943. En 1950, Arthur Samuel desarrolló un programa de damas con el algoritmo minimax. Alan Turing propuso una prueba de inteligencia para máquinas ese mismo año. Samuel introdujo el término "machine learning" en 1952, y Frank Rosenblatt creó el Perceptrón para el reconocimiento de imágenes en la década de 1950. Aunque hubo

un estancamiento, la investigación en redes neuronales y aprendizaje automático resurgió en la década de 1990.

El aprendizaje automático es una herramienta poderosa que puede ayudar a los hoteles a mejorar sus operaciones e impulsar la transformación digital. Permite a las marcas hoteleras automatizar y optimizar diversos procesos, como el análisis de datos y la toma de decisiones, lo que puede ayudarles a ser más eficientes y eficaces. Además, el aprendizaje automático puede ayudar a los hoteles a comprender mejor a sus clientes y sus necesidades, lo que puede conducir a mejores productos y servicios. Existen varios beneficios potenciales del uso del aprendizaje automático en la industria hotelera actualmente y futuro. Como por ejemplo: personalización mejorada, Seguridad mejorada, Mejora de la toma de decisiones, Chatbots, Mejorar la experiencia general de los huéspedes, etc. Por lo tanto, en general, el aprendizaje automático tiene el potencial de beneficiar enormemente a los hoteles al ayudarlos a mejorar sus operaciones, su servicio al cliente y la experiencia general de sus huéspedes.

2.3 ¿Qué es Deep Learning?

El Aprendizaje profundo o Deep Learning, es un subcampo de Machine Learning que usa una estructura jerárquica de redes neuronales artificiales, que se construyen de una forma similar a la estructura neuronal del cerebro humano, con los nodos de neuronas conectadas como una tela de araña. Esta arquitectura permite abordar el análisis de datos de forma no lineal. El aprendizaje profundo es una técnica que, al igual que otros algoritmos de aprendizaje, enseña a los ordenadores a hacer lo que es natural para los humanos: aprender con el ejemplo.[2]

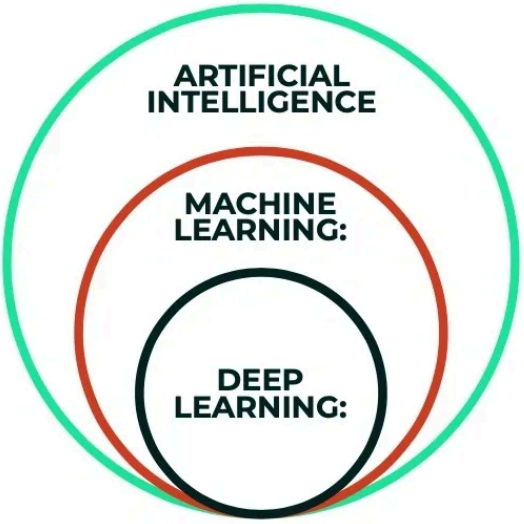

Figura 1: Jerarquía de de Inteligencia Artificial, Machine Learning, y Deep Learning

2.4 Series temporales

El análisis y la predicción de series temporales han evolucionado hasta convertirse en un método maduro para la predicción.

Una serie temporal (o simplemente una serie) es una secuencia de N observaciones (datos) ordenadas y equidistantes cronológicamente sobre una característica (serie univariante o escalar) o sobre varias características (serie multivariante o vectorial) de una unidad observable en diferentes momentos.[3]

Las series temporales se descomponen en varios componentes fundamentales que ayudan a entender y modelar su comportamiento a lo largo del tiempo:

- **Tendencia:** Representa la dirección general a largo plazo de la serie temporal, mostrando crecimiento o declinación de forma gradual y consistente debido a factores persistentes.
- **Ciclo:** Son fluctuaciones que se dan alrededor de la tendencia, generalmente de naturaleza recurrente y que no siguen patrones fijos. Un ejemplo común es el ciclo económico, que alterna entre períodos de prosperidad y recesión.
- **Variación Estacional:** Son patrones que se repiten anualmente debido a factores como el clima o costumbres estacionales. Estos patrones tienden a ser predecibles y tienen una duración fija dentro del año.

- **Fluctuaciones Irregulares:** Son movimientos impredecibles y erráticos que no siguen patrones claros. Estas fluctuaciones pueden deberse a eventos inusuales como desastres naturales o crisis económicas.

2.5 Modelos de series temporales

2.4.1 Modelo ARIMA

El modelo autorregresivo integrado de media móvil (Autoregressive Integrated Moving Average Model, ARIMA, por sus siglas en inglés) es uno de los métodos de análisis comúnmente utilizados para la predicción de series temporales.[4]

Los parámetros estructurales de ARIMA tiene 3 componentes:

ARIMA(p, d, q)

El modelo ARIMA requiere datos estacionarios y tiene un término de diferenciación (d) para hacer que los datos de series temporales sean estacionarios.

- **d**: Representa el número de diferencias necesarias para hacer que los datos sean estacionarios.

Los otros parámetros son el orden AR (p) y el orden MA (q).

- **p**: Representa el número de retardos en la serie temporal utilizados en el modelo de predicción, conocido como término AR (Auto-Regressive).
- **q**: Representa el número de retardos en los errores de predicción utilizados en el modelo, conocido como término MA (Moving Average).

2.4.2 Modelo SARIMAX

El modelo de promedio móvil integrado auto regresivo estacional con variables exógenas (SARIMAX - Seasonal Autoregressive Integrated Moving Average with eXogenous regressors) es uno de los métodos de análisis para la predicción de series temporales. La elección del modelo SARIMAX para predecir el número de pasajeros en un

hotel puesto que permite capturar la estacionalidad, manejar datos no estacionarios y considerar la correlación con datos históricos. Además, SARIMAX puede incorporar variables exógenas como la temperatura, eventos locales y vacaciones escolares, que influyen significativamente en el número de pasajeros. [5]

Los parámetros estructurales de SARIMAX consisten en siete componentes:

SARIMAX(p, d, q)×(P, D, Q)s

El modelo SARIMAX se representa como SARIMAX(p, d, q)x(P, D, Q, s), donde:

- **p:** Número de retardos en el término AR (Auto-Regressive).
- **d:** Número de diferencias no estacionales necesarias para hacer la serie estacionaria.
- **q:** Número de retardos en el término MA (Moving Average).
- **P:** Número de retardos estacionales en el término AR.
- **D:** Número de diferencias estacionales necesarias.
- **Q:** Número de retardos estacionales en el término MA.
- **s:** Periodicidad estacional de la serie temporal. Por ejemplo, si los datos de un año determinado tienen una relación estrecha con los datos del año anterior, entonces se establece una periodicidad de 12 meses.

2.4.3 Modelo LSTM

LSTM (Long Short-Term Memory) es una arquitectura de red neuronal recurrente (RNN) ampliamente utilizada en el aprendizaje profundo (Deep Learning). Se destaca por capturar dependencias a largo plazo, lo que la hace ideal para tareas de predicción de secuencias.[6]

A diferencia de las redes neuronales tradicionales, LSTM incorpora conexiones de retroalimentación, lo que le permite procesar secuencias completas de datos, no solo puntos de datos individuales. Esto la hace altamente efectiva para comprender y predecir patrones en datos secuenciales como series temporales, texto y voz.

Los parámetros principales que ayuda a realizar un modelo LSTM son los siguientes:

- **Épocas *(Epochs)*:** El número de veces que el algoritmo de aprendizaje procesa todo el conjunto de entrenamiento. Otros valores podrían ser considerados en función de

la capacidad del modelo para aprender patrones. Ajustar el número de épocas es parte del proceso de optimización del modelo y puede requerir experimentación adicional.

- **Tamaño de lote *(batch size)*:** El tamaño de lote es la cantidad de muestras que se utilizan en cada iteración de época para entrenamiento del modelo.
- **Paso de tiempo *(time Step)*:** El paso de tiempo define cuántos datos anteriores se usan para hacer predicciones.
- **Tamaño de entrada *(input_size)***: Es el número de variables independientes para predecir la variable dependiente.

2.6 Métodos de validación

2.6.1 RMSE

El Error Cuadrático Medio (RMSE, por sus siglas en inglés) es una métrica comúnmente utilizada para medir las diferencias entre los valores predichos por un modelo y los valores realmente observados. [7].

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$$

n: es el número de observaciones,
yt: representa los valores reales,
y^t: representa los valores predichos.

RMSE se utiliza ampliamente en el análisis de regresión, la previsión de series temporales y la evaluación de modelos de aprendizaje automático para evaluar qué tan bien las predicciones de un modelo coinciden con los datos observados reales.

2.6.2 Walk forward validation

Walk-Forward Validation(Verificación de series de tiempo hacia adelante): La técnica de Walk Forward Validation es necesario que la partición de los datos que se desean tomar mantenga la secuencia original de los mismos. La cual se encarga de reentrenar el

modelo con cada nuevo dato que esté disponible, evaluando su desempeño en cada nuevo punto pero más importante manteniendo la secuencia de los datos y garantizando una medición más exacta sobre los datos[8].

Capítulo 3 Metodología

3.1 CRISP-DM

La metodología que se utiliza en este trabajo de ciencia de datos es una adaptación de **CRISP-DM** *(Cross-Industry Standard Process for Data Mining)*. La metodología CRISP-DM es una de las más empleadas actualmente para el desarrollo de proyectos de minería de datos[9]. Es una metodología de facto para proyectos dedicados a extraer valor de los datos, para este proyecto, utiliza los datos resultantes para ayudar a los administradores a mejorar su toma de decisiones, por lo tanto, es una metodología adecuada para este trabajo. A continuación se muestra una descripción detallada de cada etapa de mi metodología personal junto con esta metodología aplicada al proyecto de previsión de ocupación hotelera.

Figura 2: El ciclo de la metodología CRISP-DM

1. Definición del problema

El objetivo de esta fase consiste en entender el problema que necesita abordar y los objetivos a alcanzar, por lo tanto se define el objetivo de desarrollar y validar un modelo para predecir la ocupación diaria de un hotel para optimizar la gestión de recursos y

maximizar los beneficios identificando las necesidades específicas del sector hotelero en Tenerife.

2. Recopilación de datos

La segunda fase implica recolectar información relevante para construir un dataset que pueda ser utilizado en el entrenamiento de los modelos de aprendizaje automático para la predicción de ocupación hotelera.

3. Análisis exploratorio de datos

Esta fase consiste en un análisis breve con los datos disponibles, identificando problemas de calidad y obteniendo una visión inicial de los mismos. Se realiza un análisis detallado de la estructura y distribución de estos datos, utilizando herramientas como librerías de Python (matplotlib), para visualizar tendencias en series temporales.

4. Procesamiento de datos

Durante la fase de preparación de los datos, se realiza una limpieza y transformación exhaustiva del conjunto de datos, asegurando la calidad y adecuación para el modelado. Esto incluye el manejo de valores faltantes, la normalización de datos y la creación de nuevas características relevantes para el análisis. La preparación de los datos es importante para mejorar la precisión de los modelos predictivos que se implementarán posteriormente.

5. Ingeniería de características

Esta fase tiene como objetivo crear, transformar y seleccionar características (variables) que puedan mejorar el rendimiento del modelo. Se pueden generar nuevas variables como la temperatura en Tenerife, la temperatura en otros países, días festivos, etc., que son factores que pueden afectar la predicción de series temporales. Estas variables pueden ser creadas a partir de las existentes o derivadas de fuentes relevantes. Esta fase puede retroceder a la recopilación de datos (fase 2) para mejorar y optimizar la creación de dataset.

6. Preparación de datos

Esta fase junto con modelado (fase 7) y evaluación (fase 8) pueden ser iterables para cualquier algoritmo, puesto que la fase de preparación de los datos consiste en dividirlos en un formato adecuado para su entrenamiento y evaluación. Aunque puede haber ligeras diferencias en la preparación de los datos según el modelo utilizado, pero en general, los datos se dividen en dos partes:

- Datos de Entrenamiento: Este conjunto de datos se utiliza para entrenar el modelo. Constituye la mayor parte de los datos disponibles y se emplea para que el modelo aprenda los patrones y relaciones inherentes en los datos.
- Datos de Prueba: Este conjunto de datos se separa para evaluar el rendimiento del modelo. Su objetivo es garantizar que el modelo generalice bien a datos predictivos.

7. Modelado

La fase de modelado se centra en seleccionar y aplicar técnicas de modelado adecuadas, ajustando los parámetros de los modelos y evaluando su rendimiento. Se implementan y comparan tres modelos de series temporales con diferentes parámetros: ARIMA, SARIMAX y LSTM. Cada modelo se ajusta, utilizando herramientas como statsmodels para ARIMA y SARIMAX, y keras para LSTM. El objetivo es identificar el modelo que mejor se adapte a los datos y ofrezca las predicciones más precisas.

8. Evaluación

En la fase de evaluación, se verifica si el modelo cumple con los objetivos del trabajo y se decide si es lo suficientemente bueno para ser utilizado. Como se mencionó en la fase de preparación de datos, en esta etapa también se puede evaluar o validar el modelo utilizando diferentes técnicas de validación. En este caso, se utiliza el RMSE (Root Mean Squared Error) para los modelos ARIMA y SARIMAX, y RMSE junto con la validación walk-forward para el modelo LSTM, comparando su rendimiento y precisión. Esta evaluación asegura que las predicciones sean fiables y útiles para la toma de decisiones en la gestión hotelera.

9. Análisis de resultados

En la última fase, se realiza una comparación con los objetivos establecidos inicialmente, contrastando las predicciones del modelo con los resultados reales. Se analiza posibles causas de errores o desviaciones significativas, encontrando un modelo válido para apoyar a los administradores de hoteles en la toma de decisiones efectivas, tanto en la gestión hotelera como en otros contextos relevantes.

Utilizando una combinación de la metodología CRISP-DM junto con una metodología personalizada, el trabajo sigue un enfoque estructurado y repetible para analizar y predecir la ocupación hotelera.

3.2 Herramientas y Entorno de Desarrollo

1. **Google Colab**: Entorno de desarrollo en la nube utilizado para la implementación y ejecución de los modelos.
2. **Python**: Lenguaje de programación principal, con diversas librerías especializadas en análisis de datos y aprendizaje automático.
3. **Librerías Principales**:
    a. **Pandas**: Manipulación y análisis de datos.
    b. **Numpy**: Operaciones matemáticas y manejo de matrices.
    c. **Matplotlib**: Visualización de datos.
    d. **Statsmodels**: Implementación de modelos estadísticos.
    e. **scikit-learn**: Para técnicas de preprocesamiento, evaluación de modelos, y cálculo del RMSE.
    f. **Keras**: Construcción y entrenamiento de modelos de aprendizaje profundo.

Capítulo 4 Análisis de datos

4.1 Recopilación de datos

La calidad de los datos proporcionados por los hoteles es importante para realizar análisis de datos efectivos, ya que cuanto mayor sea la calidad de los datos, más precisas serán las predicciones futuras. En este caso, se utiliza el número de huéspedes diarios de un hotel anónimo para llevar a cabo la predicción de series temporales. El hotel proporciona datos sobre las reservas y la ocupación del hotel, que abarcan desde el 6 de abril de 2022 hasta el 30 de diciembre de 2024.

	Unnamed: 0	AÑO	MES	DIA	Fecha_Foto	PAX_ESTANCIAS
0	0	2022	4	6	15-02-2024	1640
1	1	2022	4	7	15-02-2024	1732
2	2	2022	4	8	15-02-2024	1742
3	3	2022	4	9	15-02-2024	1807
4	4	2022	4	10	15-02-2024	1840
...	...	...	...	...	...	...
996	996	2024	12	27	15-02-2024	762
997	997	2024	12	28	15-02-2024	875
998	998	2024	12	29	15-02-2024	818
999	999	2024	12	30	15-02-2024	710
1000	1000	2024	12	31	15-02-2024	669

1001 rows × 6 columns

Tabla 1: Fechas, número de reservas y ocupación diaria

3.2 Análisis exploratorio de datos

Frente al problema, inspeccionar y entender los datos también es importante para aprovechar el uso del aprendizaje automático. Esto permite identificar patrones y tendencias significativas, fundamentales para entrenar modelos de Machine Learning de manera efectiva.

Se puede ver claramente en la gráfica siguiente que hubo un pico en abril de ambos años. Es evidente que después de mayo, hubo una disminución en la ocupación, seguida de un aumento, y luego otra disminución en septiembre, seguida de un aumento y una nueva disminución hasta enero.

El análisis preliminar de los datos proporcionados muestra que en abril, debido a la Semana Santa, hay un aumento significativo en los viajes. Luego, en mayo, tanto empleados como estudiantes podrían estar ocupados resumiendo planes corporativos para la primera mitad del año o preparándose para exámenes y entregas de proyectos, lo que podría reducir las tasas de turismo. Posteriormente, se sube el número de ocupantes por el comienzo de las vacaciones de verano en julio y agosto, que se extienden hasta septiembre, cuando las personas regresan al trabajo y la escuela. Más adelante, en noviembre o diciembre, comienzan las vacaciones de invierno, implica otra vez la subida de número de pasajeros. Este análisis preliminar coincide en gran medida con la temporada alta de Tenerife.

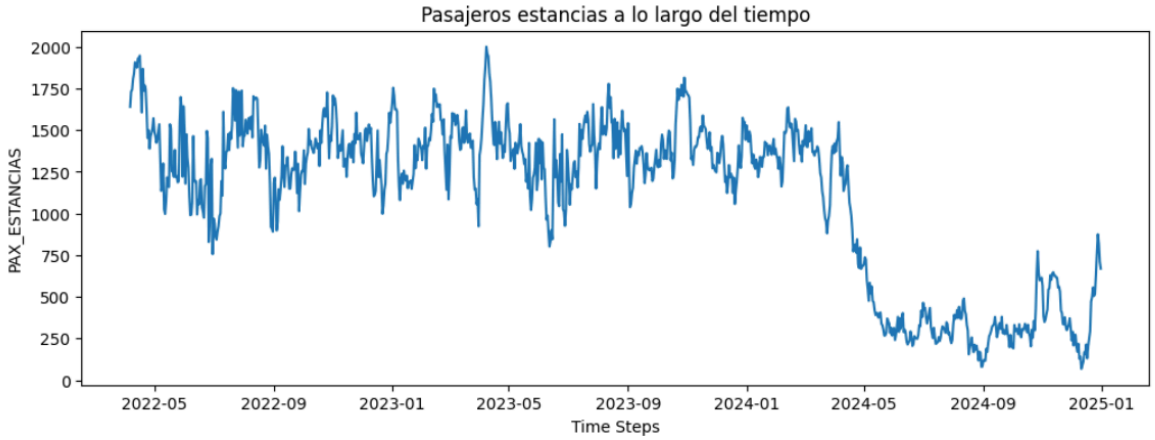

Figura 3: Número de reservas y ocupación diaria del hotel en series temporales

3.3 Procesamiento de datos

En los datos proporcionados, aunque no hay valores faltantes, es necesario convertir los datos de "pax_estancias" a enteros debido a que representan unidades de personas. (Para las variables AÑO, MES Y DÍA también).Además, como mencionamos anteriormente, los datos proporcionados por el hotel son sobre la cantidad de reservas y ocupación, abarcando desde el 6 de abril de 2022 hasta el 30 de diciembre de 2024. Sin embargo, para entrenar modelos de aprendizaje automático con datos reales, por lo tanto es necesario dividir el conjunto de datos desde el 6 de abril de 2022 hasta el 15 de febrero de 2024.

	datetime	PAX_ESTANCIAS
0	2022-04-06	1640
1	2022-04-07	1732
2	2022-04-08	1742
3	2022-04-09	1807
4	2022-04-10	1840
...	...	...
676	2024-02-11	1497
677	2024-02-12	1629
678	2024-02-13	1637
679	2024-02-14	1545
680	2024-02-15	1516

681 rows × 2 columns

Tabla 2: Fechas y ocupación diaria

Además, se utilizan pruebas de raíz unitaria, como el test aumentado de Dickey-Fuller (ADF) para determinar si la serie temporal PAX_ESTANCIAS es estacionaria, puesto que para muchas técnicas de modelado suponen los datos son estacionarios. Para ellos, se usa la librería ***statsmodels***, con la función ***adfuller.***

Resultados:

ADF Statistic:	-5.204580295756969
p-value:	8.604738749480318e-06

Tabla 3: Resultado de Dickery Fuller sobre la secuencia de pax_estancias

ADF Statistic: Este valor es el resultado numérico del test de Dickey-Fuller Aumentado (ADF). En este caso, el valor es -5.204580295756969. Este estadístico negativo indica que los datos son más estacionarios. Cuanto más negativo sea el valor del estadístico, más fuerte es la evidencia en contra de la hipótesis nula de que la serie temporal tiene una raíz unitaria (es decir, es estacionaria).

p-value: Este valor es la probabilidad asociada con el estadístico ADF. En este caso, el p-valor es 8.604738749480318e-06, que es extremadamente pequeño. Generalmente, si el p-valor es menor que un nivel de significancia (como 0.05 o 0.01), se rechaza la hipótesis nula de que la serie tiene una raíz unitaria, lo que sugiere que la serie es estacionaria. En este caso, el p-valor es mucho menor que 0.05, lo que indica una fuerte evidencia en contra de la no estacionariedad de la serie.

Dado que el valor del estadístico ADF es significativamente negativo (ADF Statistic: -5.204580295756969) y el p-valor asociado es muy bajo (p-value: 8.604738749480318e-06) , podemos concluir que los datos de la serie temporal son estacionarios en términos de media y varianza.

3.4 Ingeniería de características

La ingeniería de características puede ser un proceso importante en la predicción de series temporales, ya que permite extraer y transformar variables relevantes que mejoran la precisión de los modelos predictivos. En el contexto de la predicción de series temporales, la ingeniería de características implica la creación de nuevas variables a partir de los datos originales para capturar mejor los patrones y tendencias subyacentes.

Al recopilar datos de ingeniería de características, hay que tener en cuenta valores faltantes o valores erróneos, es necesario completarlos, puesto que esto permite garantizar que las predicciones de series de tiempo posteriores puedan ser más precisas.

3.4.1 Temperatura de Tenerife

Una de las principales ingeniería de características es la temperatura en Tenerife. Las temperaturas agradables y estables son un factor clave para muchos turistas a la hora de elegir su destino de vacaciones. Por ejemplo, lugares con climas cálidos y soleados suelen atraer a más turistas, especialmente durante las temporadas frías en otras regiones. Esto se debe a que los viajeros buscan escapar de climas adversos y disfrutar lugares con temperatura buena. La temperatura también puede influir en la programación de eventos y festividades que atraen a turistas. Por ejemplo, festivales de verano, conciertos al aire libre, deportes acuáticos y parque acuático suelen coincidir con temperaturas agradables, lo que incrementa la afluencia de turistas y la demanda hotelera.

Por lo tanto, la temperatura es una variable importante que puede influir en la precisión de los modelos de predicción de la demanda hotelera. Al incluir datos de temperatura en los modelos, se pueden identificar patrones y tendencias que ayudan a predecir mejor los picos y caídas en el número de pasajeros. Esto permite a los hoteles optimizar su gestión de recursos, ajustar precios y mejorar el servicio al cliente.Para obtener los datos de temperatura necesarios para este análisis, se utilizaron los registros proporcionados por la Agencia Estatal de Meteorología (AEMET - un organismo público español cuyo objetivo básico es la prestación de servicios meteorológicos que sean competencia del Estado.) , disponibles en el portal AEMET2013.

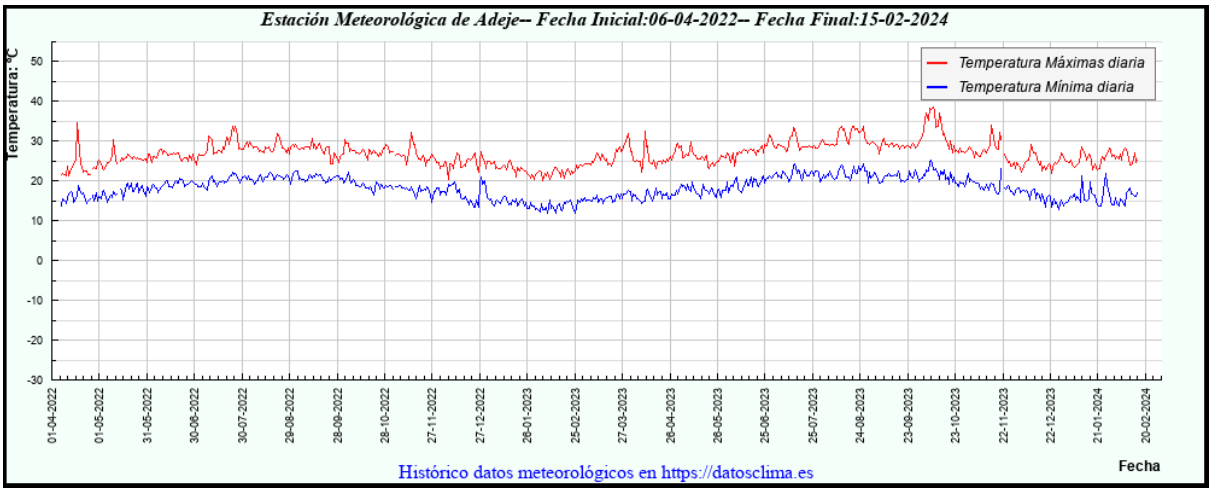

Figura 4. Datos históricos de meteorología de Adeje

Para el procesamiento de datos en relación con la temperatura de Tenerife, se aborda los datos nulos o vacíos utilizando un método de interpolación. En particular, para los días en los que faltaban valores de temperatura, excepto el primer valor del todo que toma el siguiente valor como referencia y el último valor del todo que toma el último valor como referencia. Se calcula el promedio de las temperaturas de los dos días anteriores y posteriores. Esta técnica se basa en la suposición de que la variación de la temperatura en días consecutivos no es significativa, lo que permite estimar de manera precisa y coherente los valores faltantes para asegurar la integridad y continuidad de datos.

No obstante, para poder utilizar los datos de temperatura más convenientemente en el futuro, se crea una nueva columna llamada TempAverageTNF para guardar el promedio de las temperaturas máximas y mínimas, y esta se usará como la temperatura de la isla de Tenerife.

Los datos finales de la temperatura de Tenerife son los siguientes:

	datetime	TempAverageTNF
0	2022-04-06	17.65
1	2022-04-07	17.65
2	2022-04-08	18.70
3	2022-04-09	18.25
4	2022-04-10	17.80
...	...	...
674	2024-02-11	20.30
675	2024-02-12	20.65
676	2024-02-13	21.60
677	2024-02-14	20.45
678	2024-02-15	21.30

679 rows × 2 columns

Tabla 4: Datos de temperatura de tenerife en series temporales

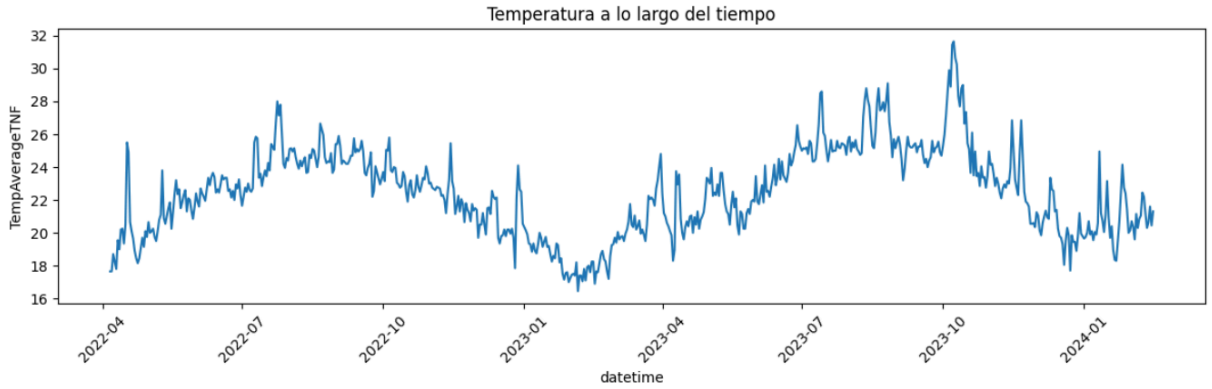

Figura 5. Temperatura media de Tenerife por día

3.4.2 Temperatura de Reino Unido y Alemania

Además de la temperatura de Tenerife, he seleccionado otras características relevantes, como la temperatura de Reino Unido y Alemania.

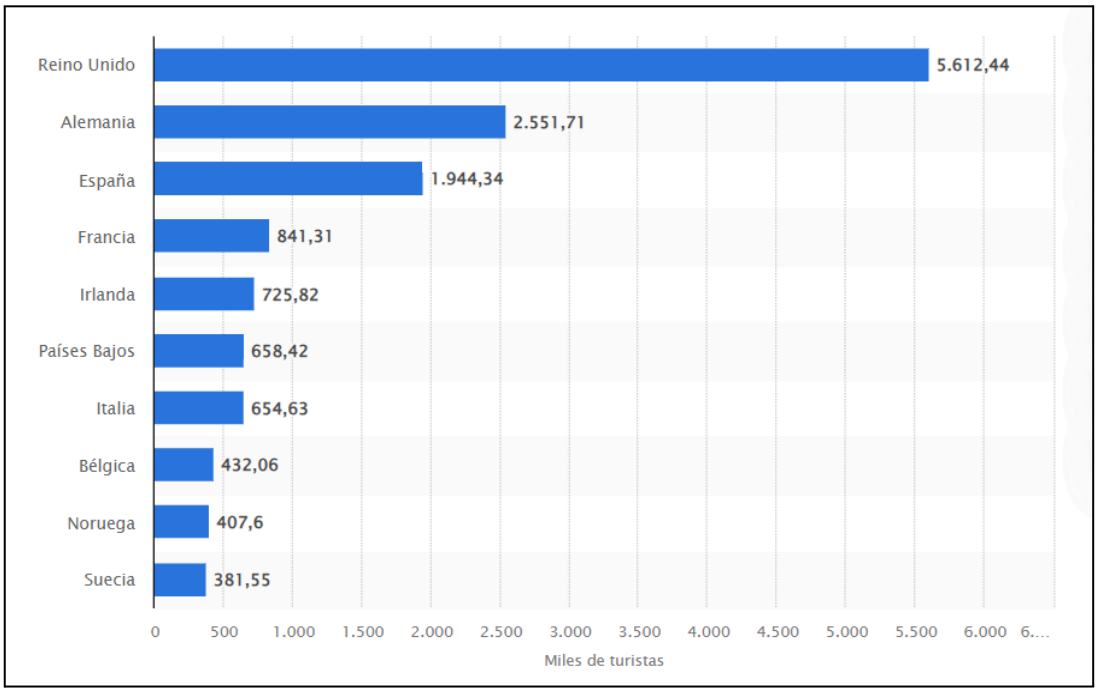

Figura 6. Principales países de origen de los turistas que visitaron Canarias en 2023 (fuente)

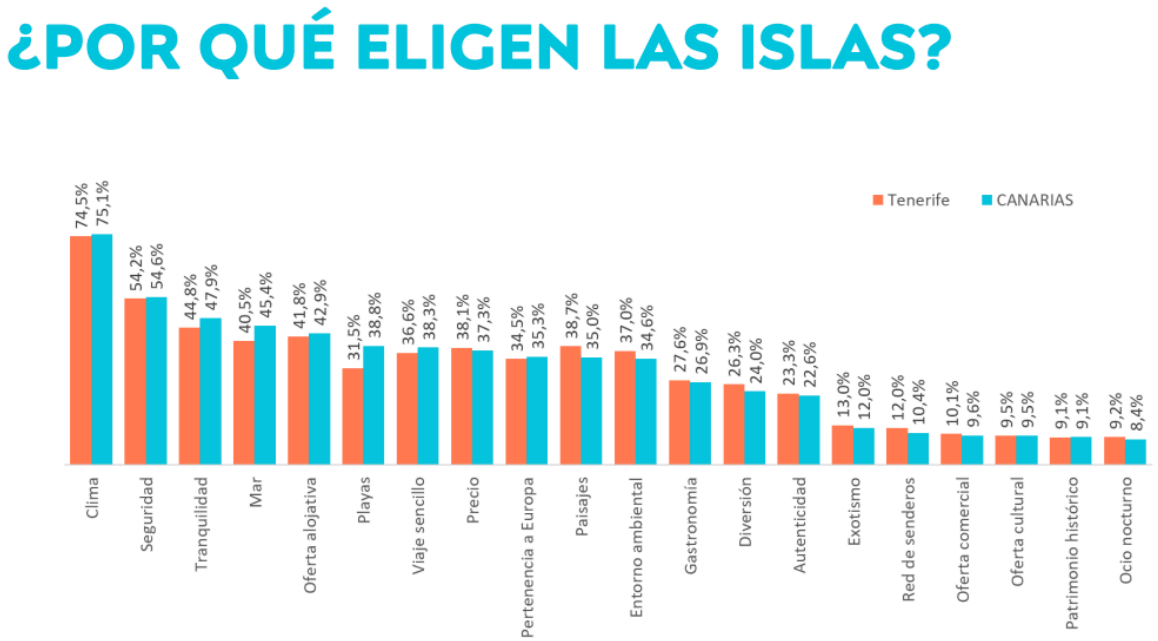

Figura 7. Motivos que eligen las islas. Perfil del turista que visita Tenerife – 2023, Turismo de Islas Canarias (fuente).

Según las dos figuras anteriores, se puede observar que, en primer lugar, el Reino Unido y Alemania son los dos países con mayor proporción de turistas que visitan Tenerife. En segundo lugar, la razón principal por la que los turistas viajan a Tenerife es su clima. Combinando estas dos razones, recopilé datos de temperatura del Reino Unido y Alemania. Al incluir las temperaturas de estos países en el análisis, se puede obtener una comprensión más detallada de cómo las condiciones climáticas en los lugares de origen de los turistas influyen en la ocupación hotelera en Tenerife.

Los datos de temperatura del Reino Unido y Alemania se han obtenido de las siguientes fuentes: los datos históricos meteorológicos de Londres provienen de la página web Weather Spark, y los datos históricos meteorológicos de Berlín se obtuvieron a través de Visual Crossing. Estas fuentes proporcionan información detallada y precisa sobre las condiciones climáticas históricas en ambas ciudades.

Debido a que no hay valores erróneos ni faltantes en los datos meteorológicos recopilados sobre el Reino Unido y Alemania, entonces se omite el paso del procesamiento de datos.

	datetime	tempC_DE	tempC_UK
0	2022-04-06	10.3	11.0
1	2022-04-07	10.4	10.2
2	2022-04-08	7.1	7.4
3	2022-04-09	5.6	7.9
4	2022-04-10	4.9	9.0
...	...	...	...
677	2024-02-12	7.2	6.7
678	2024-02-13	5.7	8.6
679	2024-02-14	5.9	12.7
680	2024-02-15	11.1	13.9
681	2024-02-16	11.9	11.8

682 rows × 3 columns

Tabla 5: Datos de temperatura Reino Unido y Alemania por día

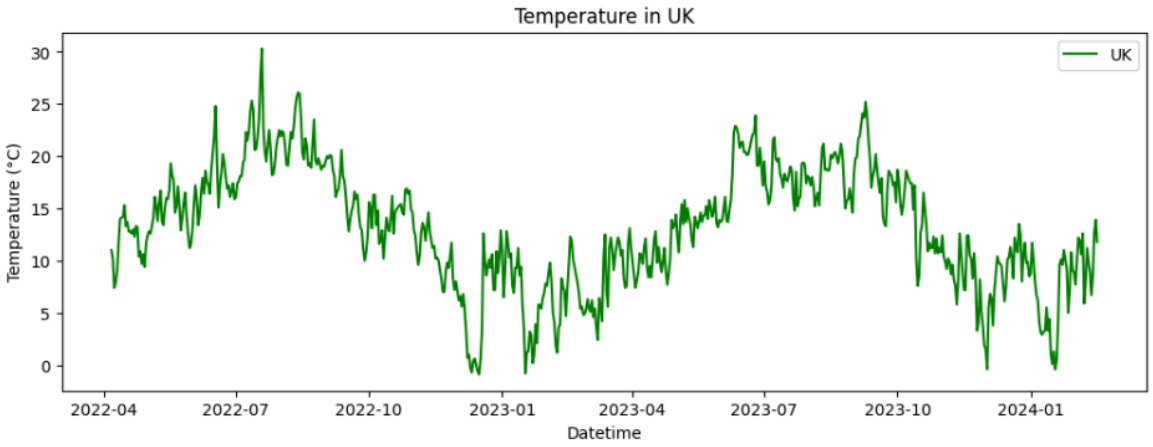

Figura 8: Temperatura media de Reino Unido por día

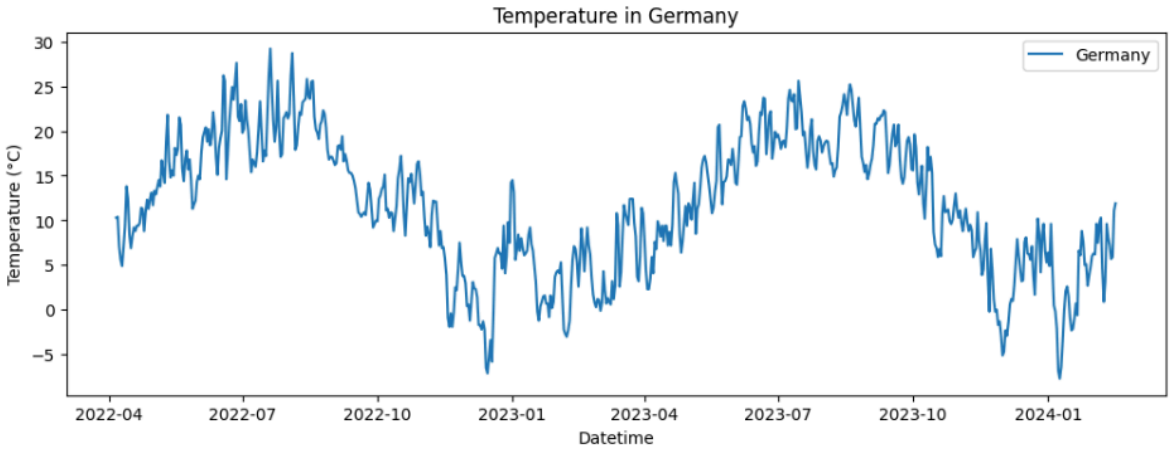

Figura 9: Temperatura media de Alemania por día

3.4.3 Días vacacionales

Otra ingeniería de características que se considera útil es los días vacacionales, tanto para trabajadores como para estudiantes. Es otra razón que puede influir en la ocupación hotelera, puesto que la gente suele viajar por relajación y descanso, la exploración de nuevos lugares y culturas, visita a familiares y amigos. En otras palabras, las vacaciones permiten a las personas dedicarse a actividades recreativas, reducir el uso de tecnología y mejorar su salud mental. Por lo tanto, al usar esta característica en el análisis, se puede obtener una visión más precisa de cómo los distintos motivos de viaje afectan la demanda hotelera. Con esta información ayuda a optimizar la gestión hotelera y mejorar las decisiones estratégicas, adaptando la oferta de servicios a las necesidades y preferencias de los turistas.

En esta ingeniería de características, se siguen los siguientes pasos:

1. **Añadir el día de la semana y determinar días vacacionales:** Para cada fecha registrada en el conjunto de datos, se determina el día de la semana correspondiente. Por ejemplo lunes corresponde con 1, martes corresponde con 2, así sucesivamente. Si la fecha cae en sábado o domingo, se considera como un día "vacacional" para estudiantes (valor 1 en la columna de días vacacionales).
2. **Recopilar los días festivos públicos del Reino Unido y Alemania:** Se recopilan los días festivos reconocidos oficialmente en el Reino Unido y Alemania. Estos días son importantes para identificar períodos en los que es probable que las escuelas estén

cerradas, los estudiantes estén de vacaciones y los días libres para los trabajadores. Los datos se obtienen de los dos sitios web siguientes:

- (United Kingdom Bank Holidays 2024)
- (Public Holidays in Germany 2024)

3. Definir los días vacacionales de estudiantes: Se consideran días vacacionales para estudiantes desde la mitad de junio hasta el día 1 de septiembre (vacaciones de verano), y desde el 20 de diciembre hasta el 7 de enero del año siguiente (vacaciones de invierno).
4. Crear una nueva columna indicativa de días festivos: Se introduce una nueva columna en el conjunto de datos para marcar con '1' los días que corresponden a días festivos (incluidos los días vacacionales definidos) y con '0' los días que no lo son. Esta columna facilita la identificación y análisis de los períodos vacacionales y festivos en los datos.

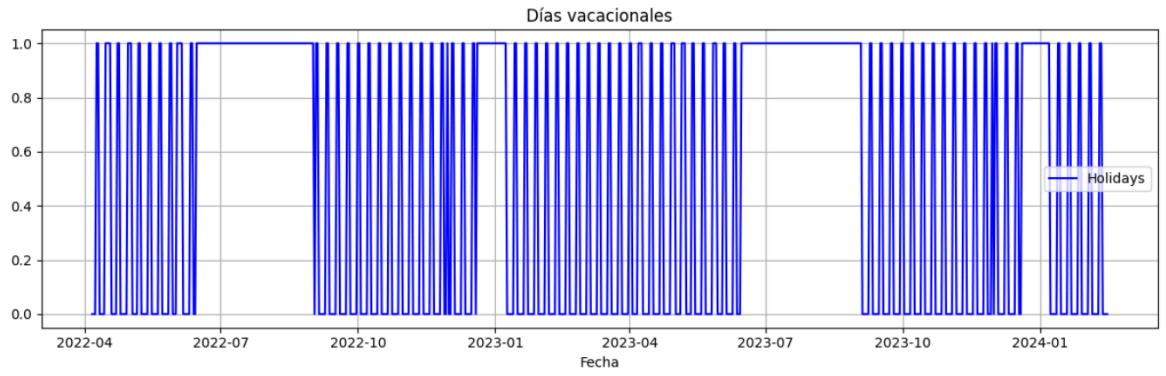

Figura 10: Días vacacionales representado por 0 (día no vacacional)y 1 (día vacacional)

Como no tiene valores faltantes, entonces no hace falta el procesamiento de datos para esta ingeniería de características.

Los datos finales se presentan de la siguiente manera:

	datetime	PAX_ESTANCIAS	TempAverageTNF	tempC_DE	tempC_UK	Week	Holidays
0	2022-04-06	1640	17.6	10.3	11.0	3	0

1	2022-04-07	1732	17.6	10.4	10.2	4	0
2	2022-04-08	1742	18.7	7.1	7.4	5	0
3	2022-04-09	1807	18.2	5.6	7.9	6	1
4	2022-04-10	1840	17.8	4.9	9.0	7	1
...	...	...	...	...	...	...	...
674	2024-02-11	1497	20.3	8.1	8.9	7	1
675	2024-02-12	1629	20.6	7.2	6.7	1	0
676	2024-02-13	1637	21.6	5.7	8.6	2	0
677	2024-02-14	1545	20.4	5.9	12.7	3	0
678	2024-02-15	1516	21.3	11.1	13.9	4	0

679 rows × 7 columns

Tabla 6: Conjunto de datos en series temporales

Capítulo 5 Construcción y Evaluación de los Modelos Predictivos

5.1 ARIMA

5.1.1 Preparación de datos

Antes del modelado de ARIMA, se realiza el proceso de preparación. Los datos se dividen en dos conjuntos: un 80% se utiliza para el entrenamiento del modelo, y el 20% restante se reserva para pruebas.

5.1.2 Modelado

5.1.2.1 Determinación de parametros de ARIMA

El primer paso en el modelado de ARIMA(p,d,q) es identificar los parámetros adecuados del modelo. Como se mencionó anteriormente, los datos de pax_estancias son datos estacionarios, por lo tanto no es necesario aplicar diferenciación adicional (d=0).

Para determinar los valores de AR(p) y MA(q), se utilizan los criterios de AIC y BIC, son dos herramientas fundamentales en estadística para la selección de modelos, especialmente en el análisis de series temporales. Además se añade variables exógenas (temperatura de Tenerife, de Reino Unido, de Alemania, y días vacacionales) para determinar valores p y q.

El AIC calcula una medida relativa de la calidad de un modelo, considerando tanto el ajuste como la complejidad del modelo. Un valor más bajo de AIC indica un modelo que proporciona un mejor ajuste con menos complejidad. Por otro lado, el BIC incorpora una penalización más fuerte por la complejidad del modelo que el AIC, favoreciendo modelos más simples cuando se comparan modelos con ajustes similares.

AR(p)	MA(q)	AIC	BIC
0	0	7376.992	7402.775
0	1	7078.036	7108.115
0	2	6970.616	7004.993
1	0	6836.744	6866.824
1	1	6825.817	6860.194
1	2	6830.317	6868.991
2	0	6827.756	6862.133
2	1	6828.903	6867.577
2	2	6831.289	6874.260

Tabla 7: Valores de AIC Y BIC según diferentes parámetros de p y q para el modelo ARIMA

Según los resultados de AIC Y BIC para el modelo de ARIMA:

El mejor modelo según AIC (Akaike Information Criterion):

- En tu caso, el mejor modelo según AIC es aquel con los parámetros p=1, d=0 (sin diferenciación) y q=1.
- El valor asociado, 6825.817 es el AIC calculado para este modelo específico.

El mejor modelo según BIC (Bayesian Information Criterion):

- Según el BIC, el mejor modelo también es aquel con los parámetros p=1 d=0 y q=1.
- El valor asociado, 6860.194 es el BIC calculado para este mismo modelo.

Ambos resultados indican que, entre todas las combinaciones de p y q, el modelo ARIMA con p=1, d=0 y q=1 es el que presenta un mejor equilibrio entre ajuste del modelo y complejidad.

Se obtiene el modelo ARIMA con los parámetros:

- **Orden de diferenciación (d): d=0**, datos estacionarios.

- **Orden del término autoregresivo AR (p): p = 1,** el modelo ARIMA considera que el valor actual de la serie temporal depende linealmente del valor de la serie temporal un paso atrás.
- **Orden del término de media móvil (q): q = 1**, el modelo ARIMA considera que el valor actual de la serie temporal depende linealmente del valor de los errores (residuos) de un paso atrás.

ARIMA(1, 0, 1)

5.1.2.1 Entrenamiento del modelo ARIMA

Una vez identificados los valores de p, d y q, se procede el modelo ARIMA para la predicción. Usa los parámetros de la función ARIMA que se utilizan:

- **train_data['PAX_ESTANCIAS']**: Esta es la serie temporal que modela, también conocida como serie endógena. Es la variable que se intenta predecir o modelar utilizando el modelo ARIMA.
- **order=(1, 0, 1):** Estos son los parámetros de orden del modelo ARIMA
- **exog=train_data[['TempAverageTNF', 'tempC_DE', 'tempC_UK', Holidays']]:** Estas son las variables exógenas que se incorporan al modelo ARIMA. Las variables exógenas son aquellas que no son parte de la serie temporal principal pero que pueden influir en ella. En este caso, las variables *TempAverageTNF, tempC_DE, tempC_UK y Holidays* se utilizan como variables exógenas para mejorar la precisión de las predicciones del modelo ARIMA.

5.1.3 Evaluación

Resultado de RMSE:

```
Root Mean Squared Error (RMSE): 162.8991
```

En este caso un RMSE de 163 unidades aproximadas (en la misma escala que PAX_ESTANCIAS) significa que, en promedio, las predicciones del modelo están desviadas en aproximadamente 163 unidades respecto a los valores reales.

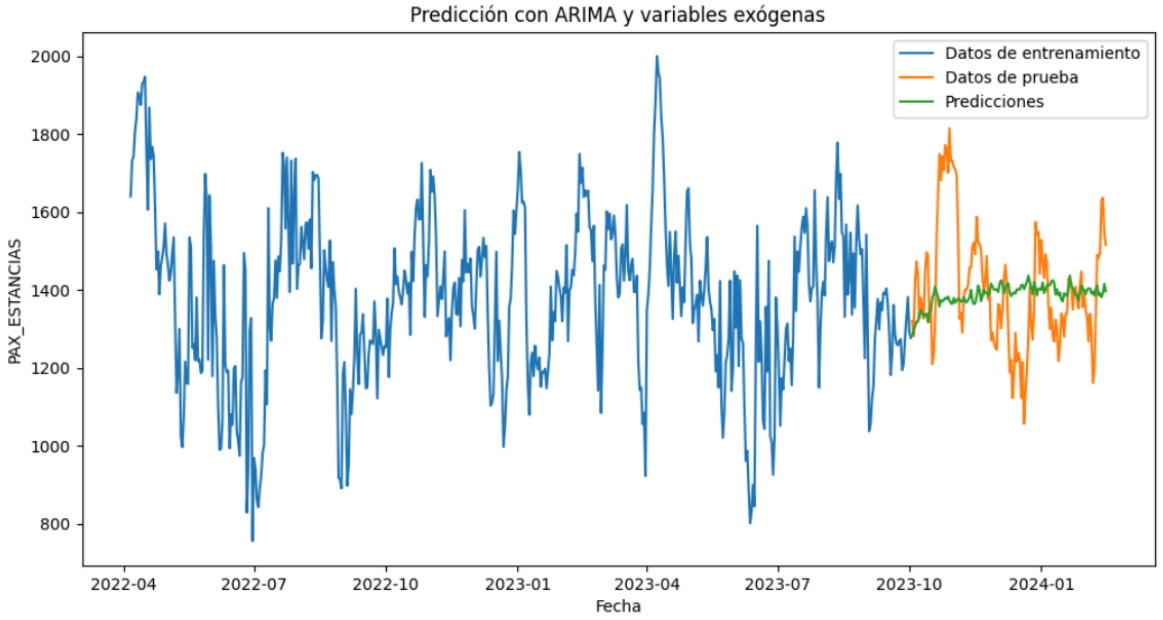

Figura 11: Predicción con el modelo ARIMA

5.2 SARIMAX

5.2.1 Preparación de datos

Igual que el modelo ARIMA, para el modelado de SARIMAX, los datos también se dividen en dos partes, por un lado, 80% de datos entrenamiento para realizar 20% de datos predictivos.

5.2.2 Modelado

5.2.2.1 Determinacion de parametros de SARIMAX

Para obtener los parámetros óptimos del modelo SARIMAX, se utilizan los criterios de información que menciona anteriormente: AIC (Akaike Information Criterion) y BIC (Bayesian Information Criterion).

Se definen los valores de p y q en el rango de 0 a 3, mientras que el valor de s se establece en 12, debido a la correlación de los datos con los últimos 12 meses. El valor de d se fija en 0, puesto que las pruebas anteriores demostraron que los datos son estacionarios. Y D = 1, puesto que los datos tienen estacionalidad, aplicar una diferencia estacional D=1 ayuda a eliminar esa estacionalidad .

SARIMAX(p, d, q)x(P, D, Q, S)	AIC	BIC
(0, 0, 0)x(0, 1, 0, 12)	9299.344	9321.850
(0, 0, 0)x(0, 1, 1, 12)	8864.933	8891.832
(0, 0, 0)x(0, 1, 2, 12)	8650.404	8681.656
(0, 0, 0)x(1, 1, 0, 12)	9019.453	9046.360
... ...	... ...	... ...
(2, 0, 2)x(1, 1, 2, 12)	8024.057	8077.595
(2, 0, 2)x(2, 1, 0, 12)	8153.765	8202.858
(2, 0, 2)x(2, 1, 1, 12)	8030.888	8084.444
(2, 0, 2)x(2, 1, 2, 12)	8024.841	8082.840

Tabla 8: Resultados de AIC y BIC para diferentes valores de (p, d, q)x(P, D, Q, S)

El mejor modelo según AIC: ***pdq(1, 0, 2)x PDQs(1, 1, 2, 12)***

- Parámetros del Modelo:
  - pdq = (1, 0, 2): Esto indica que el modelo tiene un orden de autoregresión p=1, no requiere diferenciación d=0, y un orden de media móvil q=2.
  - PDQs = (1, 1, 2, 12): Esto indica que el componente estacional del modelo tiene un orden de autoregresión P=1, requiere diferenciación estacional D=1, un orden de media móvil estacional Q=2, y una periodicidad s=12.
- Valores del Criterio:
  - AIC: 8000.208
  - BIC: 8049.284

El mejor modelo según BIC: ***pdq(1, 0, 2)x PDQs(0, 1, 2, 12)***

- Parámetros del Modelo:
  - pdq = (1, 0, 2): Igual que el mejor modelo según AIC.

    - PDQs = (0, 1, 2, 12): Esto indica que el componente estacional del modelo tiene un orden de autoregresión P=0, requiere diferenciación estacional D=1, un orden de media móvil estacional Q=2, y una periodicidad s=12.
- Valores del Criterio:
    - AIC: 8001.020
    - BIC: 8045.634

Dado que se han identificado dos posibles conjuntos de parámetros óptimos según los criterios AIC y BIC, se procederá a ajustar el modelo SARIMAX utilizando ambos conjuntos de parámetros.

SARIMAX (1, 0, 2)x(1, 1, 2, 12)

SARIMAX (1, 0, 2)x(0, 1, 2, 12)

5.2.2.1 Entrenamiento del modelo SARIMAX

Para el entrenamiento del modelo SARIMAX, se usa la librería de python statsmodels. Dado que hay dos conjuntos de parámetros seleccionados para los modelos SARIMAX, se desarrollarán y entrenarán dos modelos diferentes utilizando estos parámetros para determinar cuál ofrece el mejor rendimiento. A continuación se explican la función SARIMAX con sus parámetros:

- **train_data['PAX_ESTANCIAS']:** Esta es la serie temporal que se modela y predice, que representa el número de pasajeros en el hotel en nuestro caso.
- **order=param1:** Este parámetro especifica los órdenes del modelo ARIMA no estacional. En este caso, *param1 o param2* incluirá los valores de (p, d, q).
- **seasonal_order=seasonal_param1:** Este parámetro especifica los órdenes del componente estacional del modelo SARIMA. *seasonal_param1 o seasonal_param2* contendrá los valores de (P, D, Q, s).
- **exog=train_data[['TempAverageTNF', 'tempC_DE', 'tempC_UK','Holidays']]:** Estas son las variables exógenas que se incluyen en el modelo para mejorar la precisión de la predicción. Las variables exógenas son características o factores externos que pueden influir en la serie temporal. En este caso, *TempAverageTNF, tempC_DE, tempC_UK* representan diferentes medidas de temperatura, y *Holidays* indica si el día es festivo o no.

5.2.3 Evaluación

Para evaluar y comparar la precisión de los dos modelos SARIMAX utilizando la misma técnica de variación que ARIMA RMSE. Se obtiene los siguientes resultados:

Modelos	Parámetros	RMSE
Primer modelo	(1, 0, 2)x(1, 1, 2, 12)	`166.6479`
Segundo modelo	(1, 0, 2)x(0, 1, 2, 12)	`161.6573`

Tabla 9: Comparación entre 2 parámetros diferentes del modelo SARIMAX

El RMSE del segundo modelo (161.6573) es ligeramente menor que el RMSE del primer modelo (166.6479). Esto indica que, en promedio, las predicciones del segundo modelo están más cerca de los valores reales en comparación con las del primer modelo. Sin embargo, la diferencia entre los dos RMSE no es muy grande, lo que sugiere que ambos modelos tienen un rendimiento similar.

Un RMSE de 161.66(162 aproximadamente) o de 166.6479(167 aproximadamente) indica que, en promedio, las predicciones del modelo están desviadas por 162 unidades o 167 unidades del valor real de los pasajeros en el hotel.

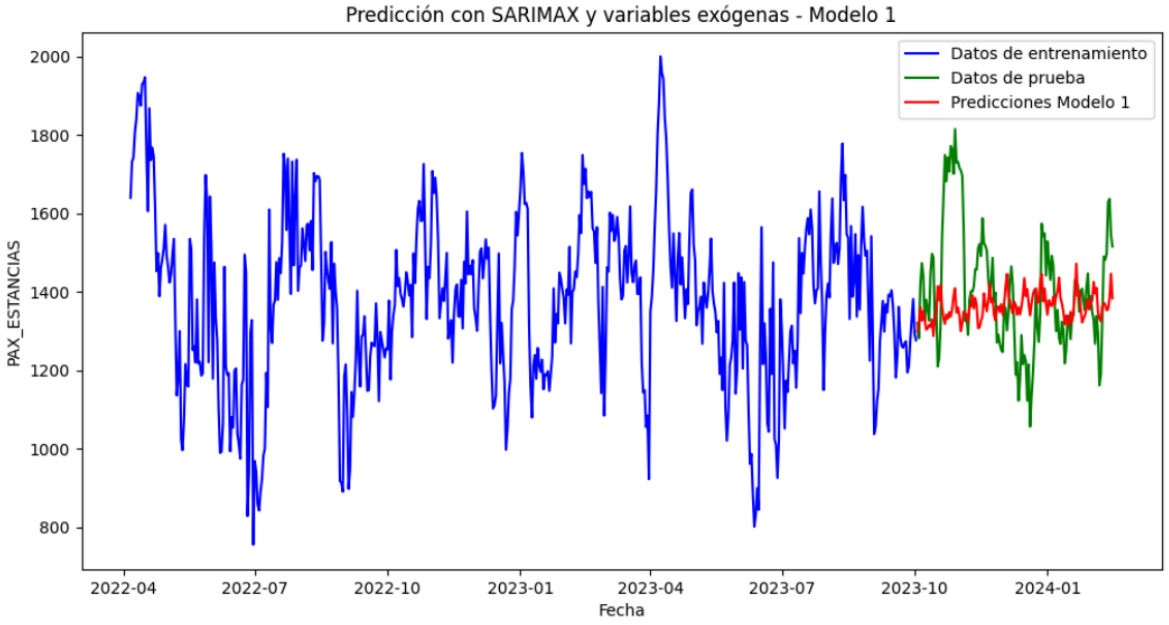

Figura 13: Predicción con parámetros SARIMAX (1, 0, 2)x(1, 1, 2, 12)

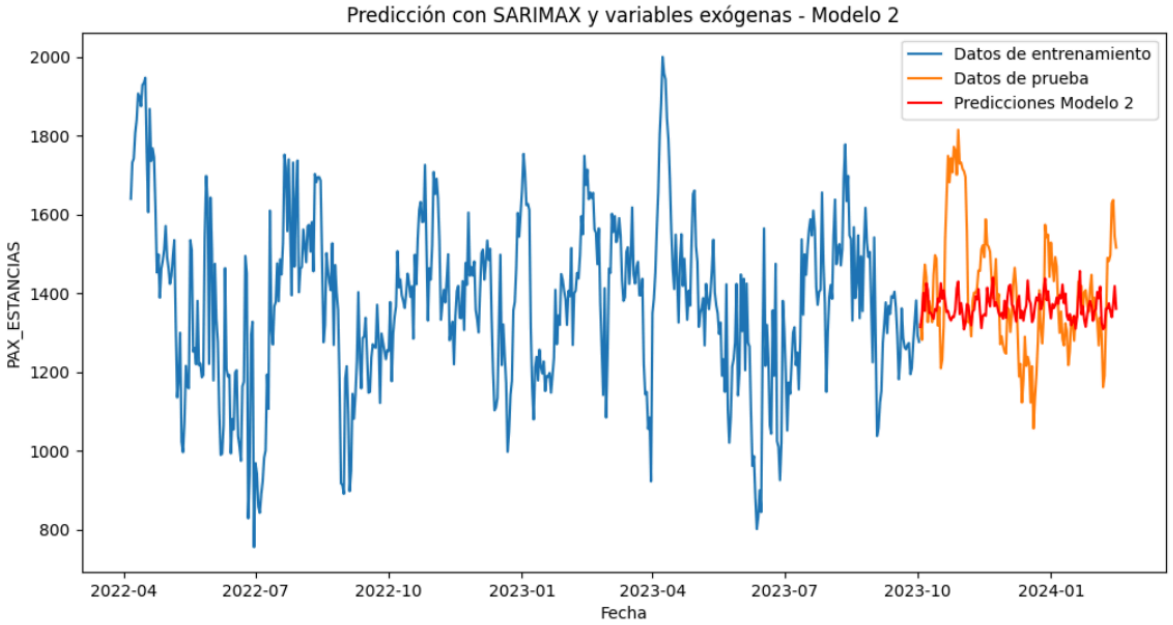

Figura 14: Predicción con parámetros SARIMAX (1, 0, 2)x(0, 1, 2, 12)

5.3 LSTM

5.3.1 Preparación de datos

Al igual que en modelos anteriores, es necesario dividir el conjunto de datos en dos partes: datos de entrenamiento y datos de prueba. Se divide el conjunto de datos en dos subconjuntos, separa los datos en entrenamiento (80%) y prueba (20%), ajusta las longitudes para que sean divisibles por 7, y reorganiza los datos en bloques semanales, puesto que se usa 7 días de datos para predecir los próximos 7 días de datos.

5.3.2 Modelado

Para el modelado de LSTM (Long Short-Term Memory) en la predicción de la ocupación hotelera, se configuraron varios parámetros clave que afectan el rendimiento y la precisión del modelo:

- **Épocas *(epochs)* = *50***: 50 ciclos que dura el entrenamiento, este valor se puede ajustar el número de etapas para optimizar el modelo.
- **Tamaño de lote *(batch_size)*:** El tamaño de lote se desarrolló en 30, eso significa que cada bloque hay 30 muestras que se utilizan en cada iteración de cada época para entrenar el modelo.

- **El paso de tiempo*(Time Step):*** El paso de tiempo se fijó en 7 días, puesto que se usa los datos de ocupación de los últimos 7 días para predecir la ocupación de los próximos 7 días.
- **Tamaño de entrada *(input_size)***: Se usan solo las variables de pasajeros (PAX_ESTANCIAS) como características. Puesto que en este modelo se centra en la variable más relevante para la predicción de la ocupación del hotel.

5.3.3 Evaluación

Utiliza la validación de series de tiempo hacia adelante para cada 7 días y así obtener el RMSE de cada día y calcula el promedio de RMSE para la evaluación final. El valor de RMSE obtenido para el modelo LSTM, es de 128.664 Este resultado indica que, en promedio, las predicciones del modelo LSTM tienen un error de aproximadamente 133 unidades aproximadamente en la misma escala que los datos originales. El valor de RMSE es mucho menor haciendo una comparación con los modelos ARIMA y SARIMAX, eso implica que los resultados predictivos son más precisos.

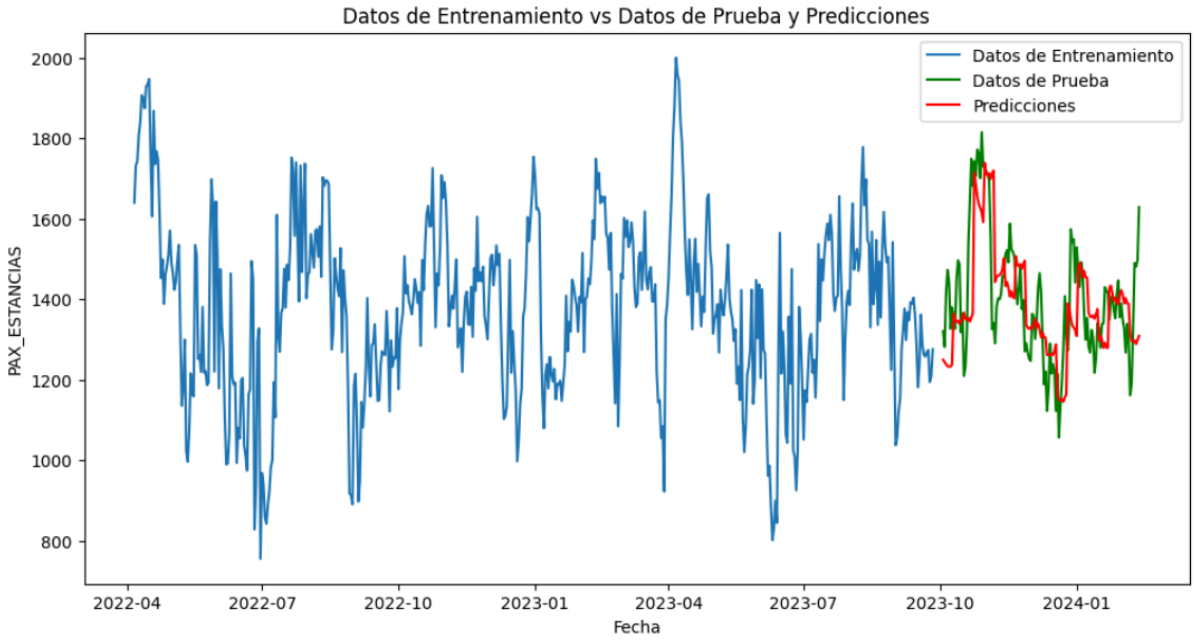

Figura 15: Predicción del modelo LSTM

Capítulo 6 Análisis de resultados

En este trabajo se compararon las técnicas de Machine Learning ARIMA, SARIMAX y LSTM con el objetivo de identificar cuál ofrece mejores resultados en la predicción de ocupación diaria en el sector hotelero. Los resultados se analizaron siguiendo la combinación de metodología CRISP-DM y personalizada, utilizando series temporales con variables adicionales como temperatura de tenerife, temperatura de reino unido y alemania, días vacacionales, y el análisis de LSTM utilizando información de ocupaciones anteriores.

	RMSE	**RMSE aproximada**
ARIMA	162.8991	163
SARIMAX	161.6573	162
LSTM	128.664	129

Tabla 10: Resultado RMSE para cada técnicas de aprendizaje automático (Machine Learning)

Respecto al modelo ARIMA, Como se puede observar en la gráfica anterior en la evaluación del modelo ARIMA y el valor de RMSE, es la técnica que tiene peor resultado de las tres, los valores predichos tienen ciertas diferencias con los valores reales, no es muy ideal para la predicción del problema, pero se considera que puede ser por:

1. Cambios en el Comportamiento Temporal: El comportamiento de reserva y estancia de los pasajeros puede evolucionar con el tiempo, afectando la relación y la correlación entre los datos históricos y los futuros.
2. Datos atípicos o fluctuaciones irregulares: Los datos atípicos pueden afectar significativamente la estimación de parámetros en ARIMA, especialmente en series de tiempo donde pueden ocurrir eventos inusuales.
3. Modelo inapropiado para la serie temporal: Para series temporales con patrones no lineales o complejas estacionalidades, como en este caso particular, ARIMA podría no ser el modelo más adecuado.

En cuanto al modelo SARIMAX, al observar las dos gráficas anteriores (apartado 5.2.3) de las predicciones frente a los valores reales y al valor de RMSE, se puede observar que el modelo SARIMAX generalmente tiene un rendimiento superior al ARIMA. Aunque existen ciertas áreas con grandes tendencias donde las predicciones no coinciden perfectamente con los datos reales, las tendencias generales son bastante similares. La razón por la cual no se detectan las tendencias a gran escala podría deberse a que, al igual que ARIMA, el modelo SARIMAX puede captar adecuadamente los patrones estacionales pero no considera la correlación entre las reservas de los pasajeros y sus estancias.

Por último, en el resultado gráfico del modelo LSTM y en el valor de RMSE, se observa que tiene un mejor resultado en la predicción de ocupación hotelera en comparación con ARIMA y SARIMAX. Esto se debe a su capacidad para capturar dependencias a largo plazo en series temporales, lo cual permite que el modelo considere la influencia acumulativa de ocupaciones pasadas en las predicciones futuras. Además, LSTM es eficaz para modelar relaciones no lineales y patrones complejos presentes en los datos de ocupación, proporcionando así una representación más precisa de las fluctuaciones y variaciones en la demanda.

Finalmente, los resultados destacan la superioridad de la técnica LSTM en la predicción de la ocupación en el sector hotelero. Además, la incorporación de variables adicionales como temperaturas y días festivos al modelo de series de tiempo mejora los pronósticos, aunque estas mejoras fueron modestas.

Capítulo 7 Conclusiones y líneas futuras

Según el análisis resultante, el pronóstico de ocupación hotelera es un problema que puede ser abordado aplicando técnicas de Machine Learning. Por lo tanto, he logrado el objetivo principal del trabajo, que es desarrollar y validar un modelo de Machine Learning para predecir la ocupación diaria de un hotel. En particular, la técnica de LSTM, que demostró ser capaz de predecir con un grado de precisión satisfactorio. Además, mediante la utilización de la metodología CRISP-DM, he alcanzado también los objetivos específicos. Esto incluye la comparación de diferentes técnicas para concluir que la técnica de LSTM es la que mejor predice la ocupación diaria en el sector hotelero.

Con el desarrollo de la tecnología, la alta capacidad de los equipos de cómputo y lenguajes especializados de programación, como Python, que ofrecen librerías de uso libre con algoritmos de Machine Learning, es más asequible la aplicación de este tipo de técnicas, consiguiendo con esto soluciones que pueden impactar de manera sustancial la competitividad en la industria hotelera.

Para futuros trabajos, se sugiere incluir en los experimentos otras variables adicionales como el crecimiento del PIB de España, la tasa de interés del BCE y el Índice de Confianza del Consumidor, ya que estas reflejan la salud económica y el comportamiento de los consumidores, influyendo en la demanda hotelera. Además, se pueden explorar otras técnicas de Machine Learning como Ridge Regression, Redes Neuronales, etc., para mejorar la interpretación de resultados y ampliar aún más las capacidades predictivas y aplicativas en la gestión hotelera.

Por último, siguiendo la metodología CRISP-DM, se puede llevar a cabo la implementación de los métodos de predicción en producción para ayudar a los administradores a tomar decisiones de manera facilitada, optimizando recursos y maximizando beneficios en los hoteles.

Capítulo 8 Summary and Conclusions

According to the resulting analysis, hotel occupancy forecasting is a problem that can be addressed by applying Machine Learning techniques. Therefore, I have achieved the main objective of the project, which is to develop and validate a Machine Learning model to predict the daily occupancy of a hotel. In particular, the LSTM technique, it proved capable of predicting with a satisfactory degree of accuracy. Additionally, by utilizing the CRISP-DM methodology, I have also achieved the specific objectives. This includes comparing different techniques to conclude that the LSTM technique best predicts daily occupancy in the hotel sector.

With the development of technology, the high capacity of computing equipment, and specialized programming languages like Python, which offer open-source libraries with Machine Learning algorithms, the application of these techniques is more accessible, providing solutions that can substantially impact competitiveness in the hotel industry.

For future work, it is suggested to include additional variables in the experiments such as the growth of Spain's GDP, the ECB interest rate, and the Consumer Confidence Index, as these reflect the economic health and consumer behavior, influencing hotel demand. Furthermore, other Machine Learning techniques such as Ridge Regression, Neural Networks, etc., can be explored to improve result interpretation and further expand predictive and applicative capabilities in hotel management.

Finally, by following the CRISP-DM methodology, the implementation of prediction methods in production can be carried out to help administrators make decisions more easily, optimizing resources and maximizing profits in hotels.

Capítulo 9 Presupuesto

Concepto	Descripción	Responsable	Herramienta	Horas aplicadas	Tasa	Costo total
Análisis de datos	Análisis descriptivo, estadísticas simples y visualizaciones de datos.	Data Scientist	Python, Google Colab	40 horas	30€/h	1200€
Recopilación de datos	Obtención de datos relevantes para el trabajo	Data Scientist	Python, Google Colab	24 horas	30€/h	720€
Procesamiento de datos	Preparación inicial de los datos.	Data Scientist	Python, Google Colab	20 horas	30€/h	600€
Modelado y Desarrollo de Modelos	Investigación de modelos e implementación y ajuste de modelos de Machine Learning o series temporales.	Data Scientist	Python, Google Colab	80 horas	30€/h	2400€
Evaluación y Validación del Modelo	Evaluación del rendimiento del modelo, ajuste de parámetros y validación	Data Scientist	Python, Google Colab	30 horas	30€/h	900€
Total				194 horas		**5820€**

Tabla 11: Presupuesto del trabajo

Capítulo 10 Referencias

[1] Shalev-Schwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. New York, NY, USA: Cambridge University Press.

[2] Cayllante Capia, J. A. (2024). *Modelo de Aprendizaje Profundo para identificar plagas en la producción de quinua. Revista Ingeniería*, *8*(20), 31–48. https://doi.org/10.33996/revistaingenieria.v8i20.116

[3] Mauricio, J. A. (2007). *Análisis de series temporales*. Universidad Complutense de Madrid.

[4] Kumar, M., & Anand, M. (2014). An application of time series ARIMA forecasting model for predicting sugarcane production in India. *Studies in Business and Economics, 9*(1), 81-94.

[5]Alharbi, F. R., & Csala, D. (2022). A seasonal autoregressive integrated moving average with exogenous factors (SARIMAX) forecasting model-based time series approach. *Inventions, 7*(4), 94. https://doi.org/10.3390/inventions7040094

[6] Yadav, A., Jha, C. K., & Sharan, A. (2019). Optimizing LSTM for time series prediction in Indian stock market. In *Proceedings of the International Conference on Computational Intelligence and Data Science (ICCIDS 2019)*. Banasthali Vidyapith, Rajasthan, India: AIM & ACT.

[7] Chai, T. and Draxler, R. R.: Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature, Geosci. Model Dev., 7, 1247–1250, https://doi.org/10.5194/gmd-7-1247-2014, 2014.

[8] Brownlee, J. (2016). *How to backtest machine learning models for time series forecasting*. Machine Learning Mastery. https://machinelearningmastery.com/backtest-machine-learning-models-time-series-forecasting/

[9] Espinosa-Zúñiga, Javier Jesús. (2020). Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. *Ingeniería, investigación y tecnología*, *21*(1), e00008. Epub 03 de agosto de 2020.https://doi.org/10.22201/fi.25940732e.2020.21n1.008

Anexo: https://github.com/XueMei-L/Trabajo-Fin-de-Grado.git