Universidad de Sonora

Facultad Interdisciplinaria de Ciencias Exactas y Naturales

Curso propedéutico de Probabilidad y Estadística

Predicción de la Demanda de Energía del Día Siguiente en la Ciudad de Hermosillo

Realizado por:

Merino Cedeño Ángel Alberto

Hermosillo, Sonora

11 de junio de 2023

Contenido

Tabla de ilustraciones

Introducción

La energía eléctrica es un bien primordial e integral, el cual ayuda al desarrollo de las actividades productivas y económicas del Estado, así como también para la transformación social, ya que interviene de forma directa en los servicios básicos para la población.

Dada la importancia e impacto que tiene la energía eléctrica en nuestras vidas cotidianas, es necesario asegurar un suministro eléctrico suficiente y confiable que permita llevar a cabo las actividades productivas de los diferentes sectores en las que interviene; como en las telecomunicaciones, el transporte, la industria, la agricultura, los comercios, los servicios, las oficinas y los hogares. Estas actividades ayudan a impulsar el crecimiento y el desarrollo económico del país.

No sólo esto, la producción de energía tiene un impacto en el ambiente. Este es un producto que se convirtió en una forma de energía complementaria y alternativa en muchos ámbitos, en un inicio un movimiento innovador. Además, hoy en día busca ser una forma de energía más versátil y limpia en comparación a otros tipos de energía, como los hidrocarburos. Irónicamente su producción tiene un impacto negativo en el ambiente, esto varía dependiendo la forma en la que se produce la energía. Ejemplos de impacto ambiental, son los siguientes:

- Emisiones de gases de efecto invernadero: La generación de energía a partir de combustibles fósiles (carbón, petróleo, gas natural, etc.) produce grandes cantidades de dióxido de carbono (CO2) y otros gases de efecto invernadero. Estos contaminantes contribuyen a la mala calidad del aire y pueden tener efectos negativos en la salud humana y en los ecosistemas.
- Impacto en los recursos hídricos: Producción energética a través de las centrales hidroeléctricas y las plantas de energía nuclear, tienen impactos significativos en los recursos hídricos. Las represas hidroeléctricas pueden alterar los ecosistemas acuáticos y afectar a las especies que dependen del agua. Por otro lado, las plantas nucleares requieren grandes cantidades de agua para el enfriamiento, lo que puede afectar la disponibilidad de agua dulce.
- Generación de residuos: La producción de energía también puede generar residuos peligrosos. Por ejemplo, las centrales eléctricas de carbón producen grandes cantidades de cenizas volantes y gases de desecho que contienen metales pesados y otros contaminantes.
- Impacto en la biodiversidad: La construcción de infraestructuras de energía, como represas hidroeléctricas y parques eólicos, puede implicar la destrucción de hábitats naturales y la fragmentación de ecosistemas.

Viendo que la producción de energía tiene muchas áreas de impacto, tanto económicas, sociales y ambientales, es necesario poder generar la cantidad, lo más precisa de energía eléctrica. Recordando

que producir energía es algo muy caro, en términos económicos y de materia prima, además, no podemos sobre producir energía dado que el almacenamiento de energía eléctrica es algo que nadie tiene en cuenta, ya que es algo casi imposible de lograr.

El objetivo dentro de este proyecto es claro. Utilizar un modelo de series temporales que nos permita predecir el consumo de energía futuro, este es un problema de series de tiempo. Para esto es necesaria conocer las variables que intervienen, o las más utilizadas, a la hora de querer hacer este tipo de modelos.

Figura 1 Parámetros que intervienen en el análisis de la predicción de consumo de energía futuro

Metodología

Los enfoques basados en datos se utilizan para estimar mejor las demandas de los consumidores, optimizar la gestión de la energía y reducir el impacto ambiental. Pasos para resolver dicha problemática son los siguientes.

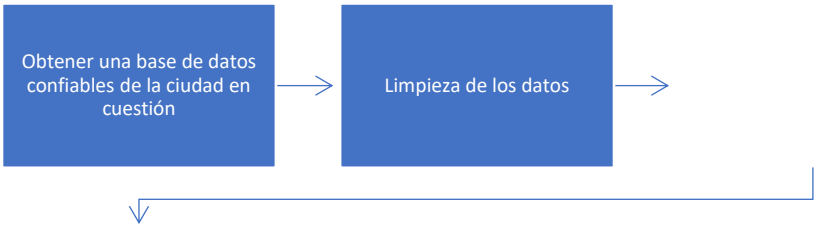

Figura 2 Diagrama de procesos para implementar un modelo de predicción

En esta metodología vamos a explicar conceptualmente lo que se hizo para crear el modelo que nos ayuda a lograr nuestro objetivo. Sin embargo, también hay una explicación más clara, explícita y detallada de todas las funciones, las gráficas, los DataFrames y más, dentro de nuestra libreta de Jupyter Notebook.

Los datos recopilados para este trabajo fueron tomados de la base de datos abiertos proporcionados por la CENACE (Centro Nacional de Control de Energía). Siendo que esta base de datos no incluye datos climatológicos ni las fechas de asueto, o festividades del país, la base de datos correspondientes al clima fue tomada de una página de la NASA, con las variables de temperatura mínima, temperatura máxima, temperatura de punto de rocío, temperatura de bulbo húmedo, humedad específica, humedad relativa, presión superficial, velocidad del viento, e índices de rayos UV. La base de datos correspondiente a los días festivos fue elaborada a "mano", basándonos en un calendario que encontramos en internet.

Con las bases de datos es importante saber si hay datos faltantes, esto ya que queremos resolver un problema de series de tiempo, y si nos faltan datos, esto puede afectar en las predicciones futuras. Con esto en mente, cada DataSet que utilizamos, con excepción del que hicimos a "mano", aplicamos una función de información, el cual nos brinda la cantidad de datos No Nulos, y el tipo (int, str, float, etc.) de datos pertenecientes a cada DataFrame. A cada DataFrame se le hizo un cambio en las columnas donde venían las fechas, esto ya que tenían el tipo *object*, pero con la librería de *pandas*, les hicimos un cambio a tipo *datetime64*. Esto es necesario ya que queremos utilizar las fechas como índices para las gráficas y demás. También, los DataFrames que utilizamos no contenían *missing data*.

Primero decidimos darles tratamiento a los datos provenientes de la CENACE. ¿Por qué hicimos esto si los datos no contenían *missing data*? Bueno, Python cuenta con una función que describe los valores (en este caso numéricos, pero también se puede para los tipo *string*), dándonos el número de datos que hay en el DataFrame, promedio, desviación estándar, el valor mínimo, los percentiles (25, 50 y 75), y el valor máximo, para las columnas de tipo numérico. Dentro de esta función, vimos que el valor máximo estaba en un orden de magnitud muy alto, comparando con el promedio y los percentiles, decidimos ver una ilustración que nos permitiera entender cómo se distribuían los datos y ver qué era lo que estaba pasando. Con esto en mente, obtuvimos la siguiente gráfica.

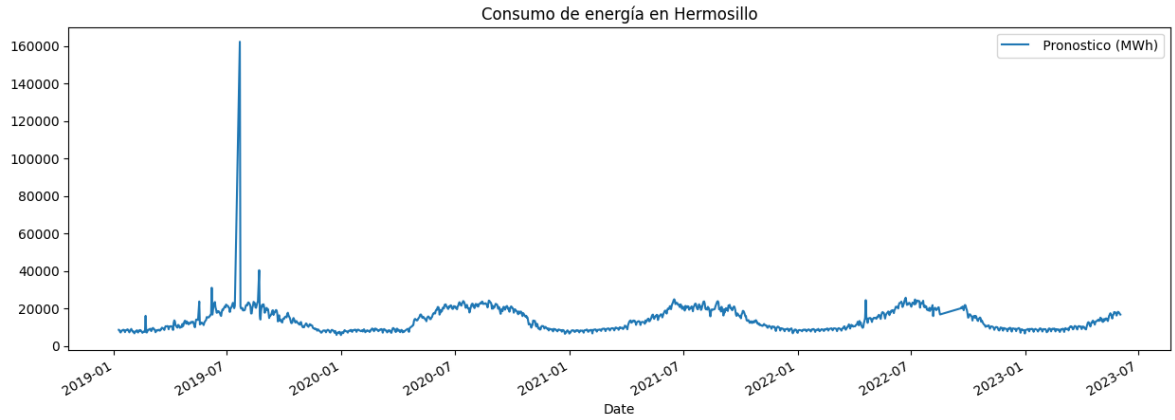

Figura 3 Consumo de energía en Hermosillo, Sonora

Esta gráfica nos permite ver que efectivamente hay unos datos que tienen un comportamiento atípico. Por eso es por lo que decidimos mejor eliminarlos. Eliminar datos no es la mejor técnica, y lógicamente mete ruido, o sesgo, en nuestro modelo. Sin embargo, creemos que es la mejor decisión dadas las circunstancias, además que sólo eran 4 valores atípicos, de más de 1500 datos. Después de esta decisión, observamos cómo quedan nuestros datos.

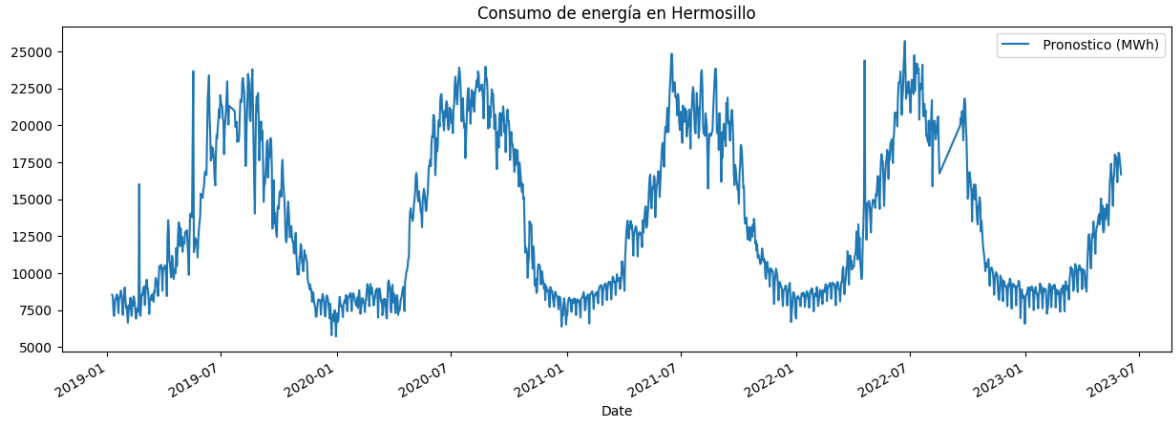

Figura 4 Consumo de energía, sin datos atípicos, en Hermosillo, Sonora

Ya con datos más consistentes, pasamos al siguiente paso. Hacer un *merge* del DataFrame de la demanda de energía con el del clima. Y siendo que existe una función que describe el DataFrame en número y que da bastante información, no hay información que sea mejor visualizada que con una gráfica. A partir del *merge*, graficamos el valor máximo y mínimo de temperatura a lo largo del

tiempo (del 9 de enero del 2019 al 2 de Junio del 2023), junto con la demanda energética histórica. Dándonos como resultado la siguiente gráfica.

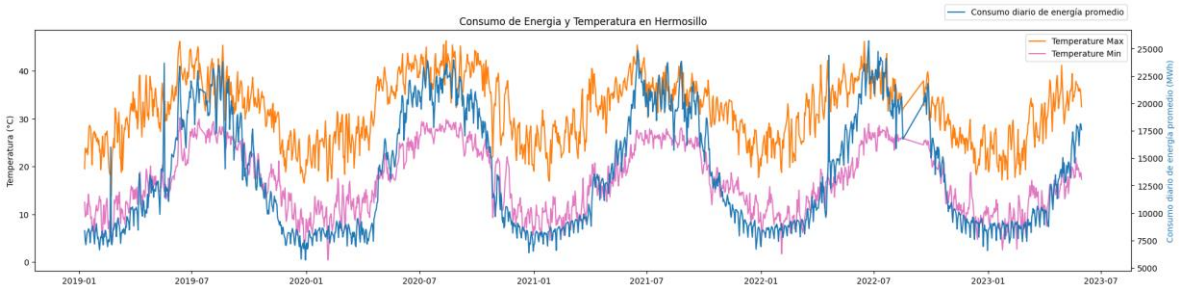

Figura 5 Consumo de energía y temperatura de Hermosillo

Vemos que el comportamiento era algo esperado. Sabemos que cuando la temperatura sube en nuestra ciudad, es rara la casa que no se suministra con un equipo de aire acondicionado, y eso lo podemos ver, mientras más aumente la temperatura, también lo será el consumo de energía. Algo similar pasa con la siguiente gráfica, donde se grafica la humedad relativa contra el consumo.

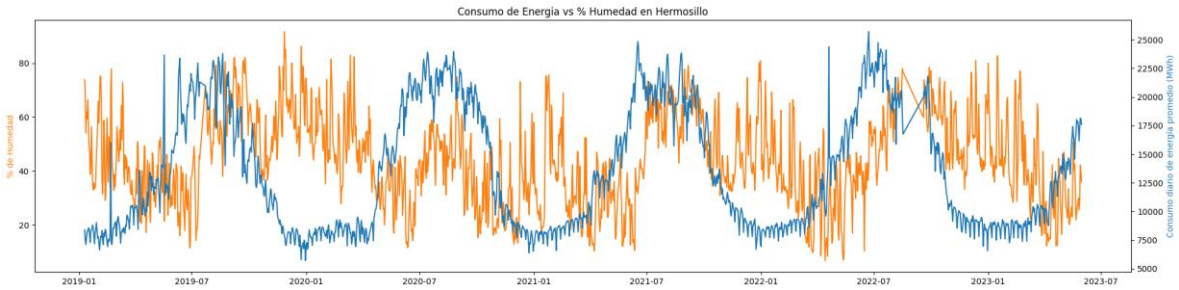

Figura 6 Consumo de energía y humedad relativa de Hermosillo

Esta relación, dada las condiciones de la ciudad, podría decirse que aparenta una relación lineal (negativa). La humedad relativa depende de variables geográficas y climatológicas. Sin embargo, las condiciones de nuestra ciudad (lugar árido, seco), a mayor temperatura, menor humedad relativa, y el consumo de energía nuevamente aumenta.

Graficar hasta encontrar qué variable se ajusta mejor a la demanda de energía no es lo más apropiado, para esto existen técnicas. Una de ellas es la de correlación de variables, y Python tiene sus formas de representar (de forma gráfica) esto.

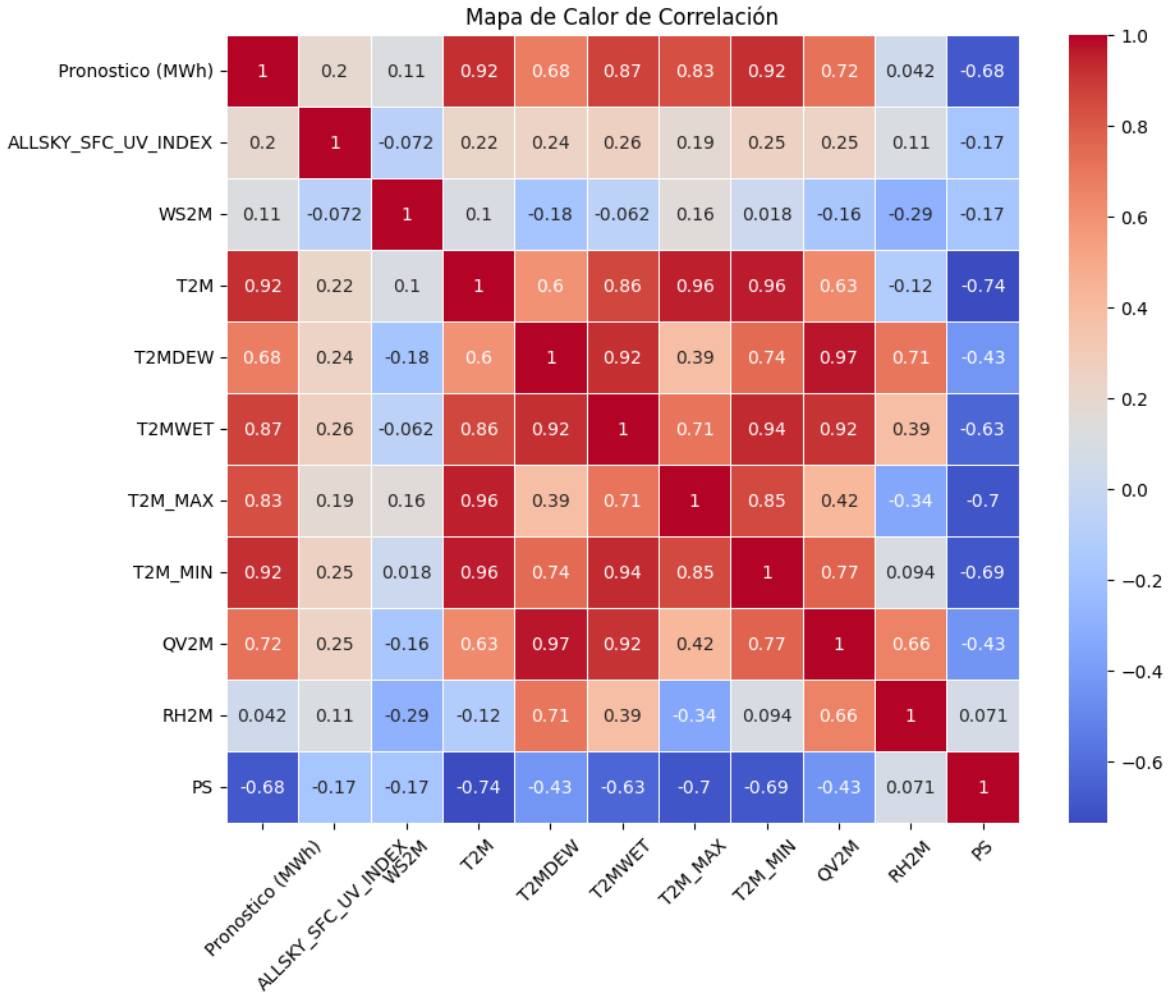

Figura 7 Heatmap de correlaciones

Recordemos que una relación de 0 no significa que dos variables no tengan relación, sino que su relación simplemente no es lineal, puede existir relación entre ellas, pero sería o no lineal, o compleja, dependerá el caso. Una relación negativa indica que, si una variable aumenta, la otra disminuye, pero también hay que recordar que relación no implica causalidad. Esto último aplica para las relaciones positivas, donde si una variable aumenta, la otra también; sin olvidar que relación no implica causalidad.

Viendo este *Heatmap*, queda claro qué variables tienen una fuerte relación con nuestra variable de interés. Este proyecto no es nuevo, está basado de varias referencias, donde una de ellas (Kaggle) trabajaba con 3 variables de clima que creían eran las más relevantes (velocidad del viento, temperatura máxima, y la humedad (en este caso, la relativa). A partir de estas 3 variables, se decidió crear *clusters*; estos sirven para la agrupación de datos, grupos que se forman basados en características propias de los datos, ayudando a encontrar patrones que pueden ayudar en el modelo de predicción.

Para lograr esto, debemos normalizar los datos en un rango de 0 y 1; tenemos entendido que esto dependerá de cada caso, pero en este caso se normalizan dentro de ese rango con tal de formar los *clusters*. Una vez le damos tratamiento a los datos, los aplicamos a nuestro DataFrame para que asigne

los *clusters* correspondientes a nuestros registros. Graficamos para ver más o menos qué podemos visualizar.

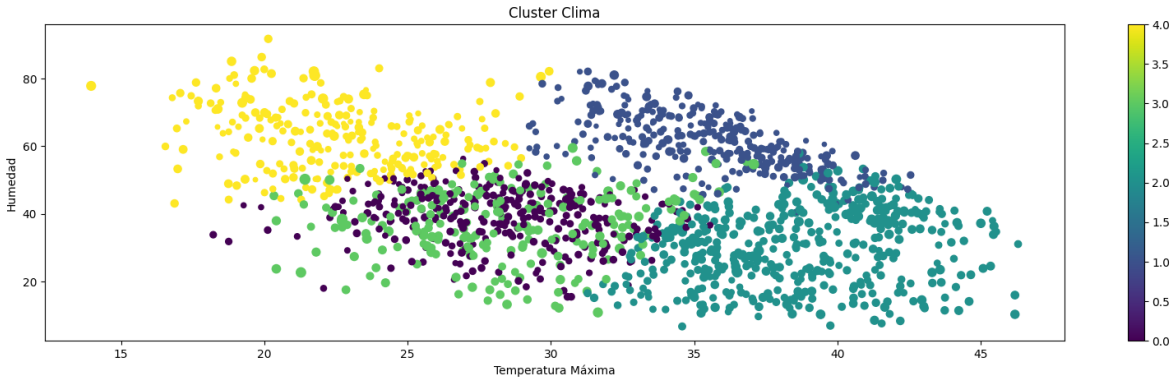

Figura 8 Clusters del Clima

Este tipo de gráficas son muy interesantes, en esta en específico, estamos viendo una representación gráfica de 4 variables, la temperatura máxima, la humedad relativa, la velocidad del viento (no se aprecia mucho pero el tamaño de cada punto indica la magnitud de viento de ese día) y el color (clasificación del *cluster*). La clasificación del *cluster* para cada registro no fue algo que hayamos hecho nosotros de forma directa. Para que se creen los *clusters* hay algo que se llama K-means (K-medias en español), y esto es un algoritmo de aprendizaje automático no supervisado y que nos ayuda a agrupar los datos, haciendo que sean similares entre sí y diferentes entre los demás, basándose en distancias suma de distancias al cuadrado de cada punto al centroide más cercano.

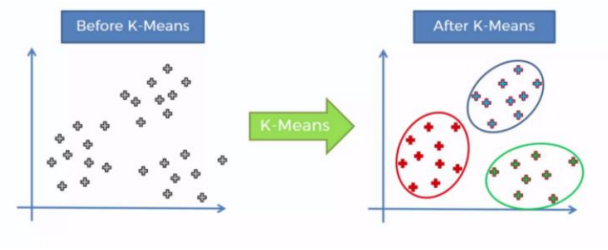

Figura 9 Explicación de lo que hace K-means

Hasta ahorita tenemos un DataFrame que contiene datos climatológicos y la demanda de energía histórica. En este punto es cuando hacemos un *merge* del DataFrame actual con el de los indicadores de días festivos (1 para el día que sí es festivo, 0 para el que no). Tomamos las columnas del DataFrame que incluye la demanda, el *cluster* del clima y el indicador de días festivos. La siguiente gráfica es una representación de los datos a como los tenemos ahorita, sólo para una representación de lo que vamos a hacer.

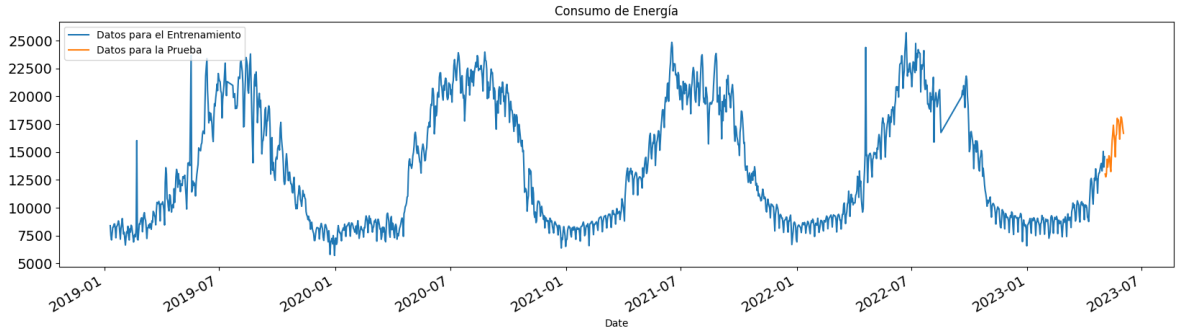

Figura 10 Separación de los datos, los de entrenamiento y los de prueba.

Lo que hicimos aquí fue separar nuestro DataFrame. Lo azul son todos los datos que van a entrar en nuestro modelo, los datos de entrenamiento. Los últimos 30 datos (últimos 30 días, color naranja) son datos que no va a ver el modelo. Estos nos van a servir para medir qué tan buena es la predicción de nuestro modelo; lo vamos a hacer que nos prediga esos 30 días que no va a ver, pero que sabemos cuál es su valor, para después aplicar unos métodos que nos dirán qué tan bueno es este modelo.

Para que nuestro modelo tenga una muy buena predicción, que sea lo más precisa posible, vamos a aplicar una descomposición estacional. Esto es un procedimiento estadístico que se utiliza para descomponer una serie de tiempo, donde sus componentes fundamentales son la tendencia, la estacionalidad y los residuos. El significado de cada uno es el siguiente:

- Estacionalidad (Seasonality): Identifica patrones que se repiten en ciertos intervalos de tiempo, como estaciones del año, meses, semanas o días de la semana.
- Tendencia (Trend): La tendencia representa la dirección general de los datos a medida que se mueven en una dirección específica a lo largo del tiempo.
- Residuo (Residual): Los residuos son las partes de los datos que no pueden explicarse por la estacionalidad ni por la tendencia. Representan las variaciones impredecibles o aleatorias en los datos.

Para lograr visualizar y descomponer la serie de tiempo, existen funciones que se importan en Python, y que se pueden graficar para ver qué nos indican.

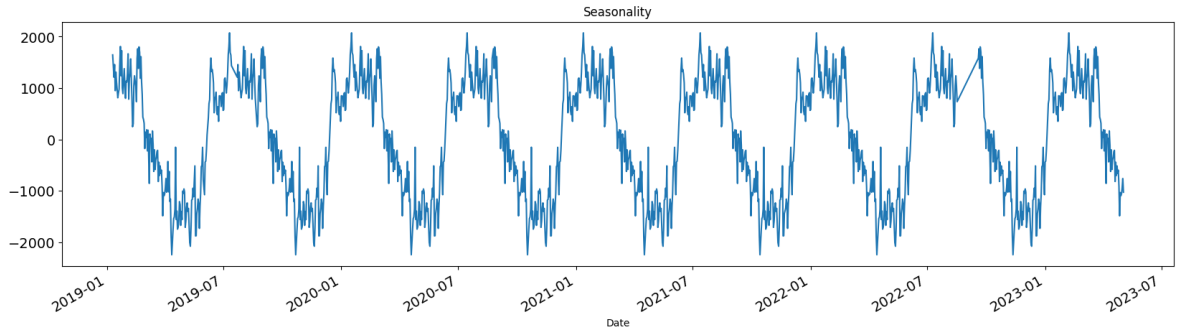

Figura 11 Estacionalidad

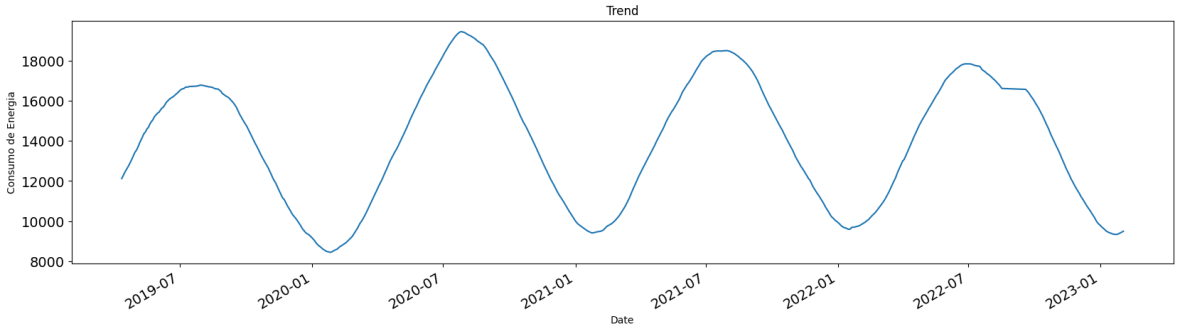

Figura 12 Tendencia

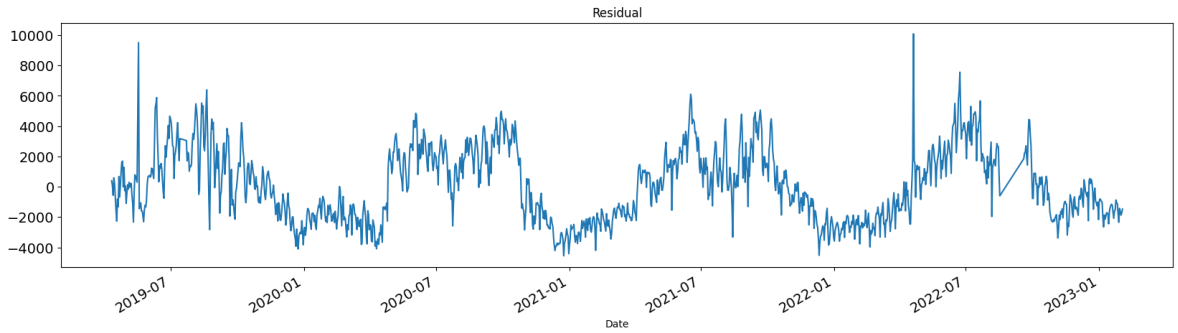

Figura 13 Residuo

Después de ver esta descomposición, es en este momento en el que vamos a trabajar con un modelo estadístico bastante utilizado en problemas de series de tiempo, del cual también nos basamos de la bibliografía dado que no es el único, SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous variables). Al utilizar SARIMAX, se pueden estimar los parámetros del modelo a partir de datos históricos y luego utilizar el modelo ajustado para realizar pronósticos futuros. En términos sencillos, SARIMAX permite modelar y predecir una serie de tiempo teniendo en cuenta tanto los patrones estacionales y de tendencia inherentes a la serie como las influencias externas o factores exógenos que pueden afectarla.

Una vez que aplicamos los modelos estadísticos necesarios para afinar, o ajustar, el modelo, observamos el comportamiento que se describe por parte del modelo.

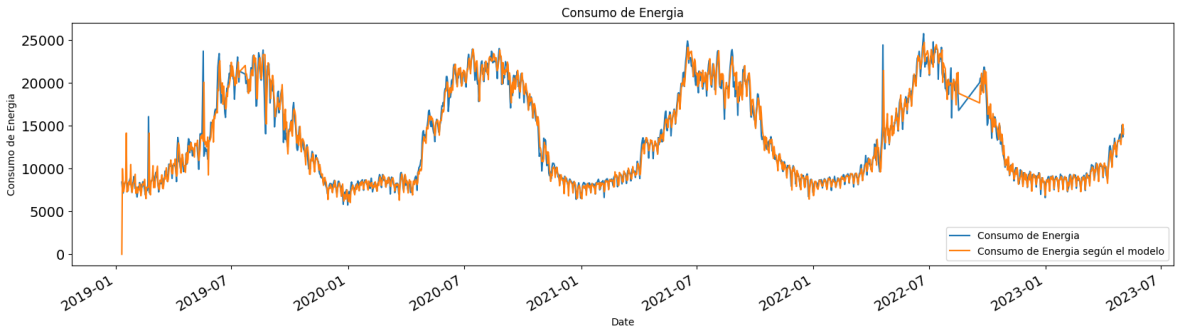

Figura 14 Consumo de energía

Esta gráfica es una representación del comportamiento del modelo en comparación a los datos históricos utilizados para su construcción. Lo azul son los datos de entrenamiento. Lo naranja sería

la forma en la que el modelo ajustó los datos. Entonces, ya tenemos un modelo y aparénteme su comportamiento es aceptable. Sin embargo, si queremos saber si es un buen modelo, o no, hay que ponerlo a prueba, y para esto es que utilizaremos los 30 últimos datos que no vio el modelo. Nosotros sabemos los valores de esos 30 días, pero veamos qué valores "predice" (porque ya los conocemos), y a partir de esto, haremos uso de dos conceptos, MAE y MAPE. Pero primero veamos las gráficas.

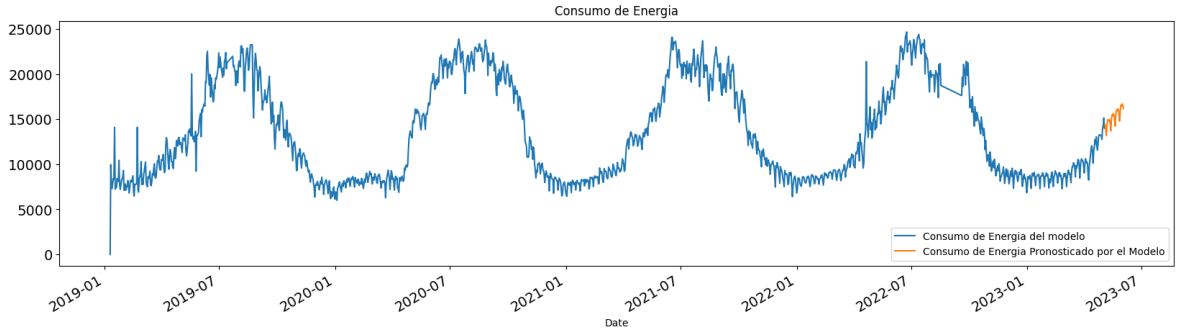

Figura 15 Consumo de energía de los últimos 30 días según el modelo

Vemos que, con este modelo de series temporales, obtenemos una predicción. ¿Qué tan buena es la predicción? A simple vista no lo podemos decir, pero para evaluarlo existen métodos, métodos ya antes mencionados.

- MAE (Mean Absolute Error): Medida de error absoluto promedio. Calcula la diferencia absoluta promedio entre las predicciones del modelo y los valores reales.
  - $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$
- MAPE (Mean Absolute Percentage Error): Medida de error porcentual promedio. Calcula el porcentaje promedio de diferencia absoluta entre las predicciones del modelo y los valores reales en relación con los valores reales.
  - $\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| * 100$

Retomando, con estos dos métodos obtenemos, para este caso y modelo, los siguientes valores.

- MAE: 956.12 MWh
- MAPE: 5.9669 %

¿Estos valores qué nos dicen? Bueno, Con estos valores de MAE y MAPE podemos ver que, aparentemente, el MAPE no es muy grande, comparando con los valores que trabajamos que son de magnitudes más grandes, entonces podríamos pensar que no es mucho. Sin embargo, hay que poner atención a las unidades, hablamos de casi 1000 MWh, o sea, casi 1,000,000 de KWh. Si el consumo por habitante es de unos 2,110 KWh (aproximadamente) en México (Gestión de Energía En México, 2020), entonces estaríamos hablando de 474 (redondeado) personas, menos que más (redondeamos el MAE). Si comparamos con el tamaño de la población en México, es muy poco, pero el impacto sigue estando ahí.

Si nos vamos con el MAPE, vemos que, dentro de lo que cabe, es un porcentaje aceptable, un porcentaje de error muy acertado para el modelo que creamos con las variables que utilizamos. Estos dos conceptos nos ayudan a evaluar nuestro modelo de regresión, para un mejor análisis de este.

Conclusión

Este trabajo se hizo con fines prácticos para emplear métodos estadísticos que nos ayudaran a analizar, evaluar, entender y resolver el objetivo planteado. Hicimos uso de varios recursos para poder llevarlo a cabo y entender, paso a paso, lo que estaba pasando. Si bien hay uso de conceptos nuevos (como el SARIMAX), se hizo un esfuerzo por entender, a grandes rasgos, lo que había de fondo. Los problemas de series de tiempo son interesantes, y en parte, creemos que son algo complejos de entender. Sin embargo, gracias a las gráficas, análisis de los datos dentro de los DataFrames, y métodos, o conceptos, utilizados, se pudo lograr un alcance bastante aceptable, permitiéndonos poder resolver el problema.

El modelo que se generó lo calificaríamos como bueno, o aceptable. Aunque hay secciones que mejorar, como la cantidad de datos, variables, etc., la finalidad de esto era meramente académica, para un proyecto profesional, se tendría en cuenta más conceptos, más técnicas y un equipo más grande de trabajo. Esto ya que no podemos tener un conocimiento sobre un universo tan amplio como lo es la energía, desde un inicio vimos su área de impacto y es prácticamente en todo, por lo tanto, hay variables que están en zonas en las cuales no tenemos conocimiento a priori.

Bibliografía

Real Python. (2021, March 17). *Python AI: How to Build a Neural Network & Make Predictions*. Realpython.com; Real Python. https://realpython.com/python-ai-neural-network/

rheajgurung. (2018, December 11). *Energy Consumption Forecast*. Kaggle.com; Kaggle. https://www.kaggle.com/code/rheajgurung/energy-consumption-forecast

OpenAI. (2021). ChatGPT (Version 3.5) [Software]. Recuperado de https://openai.com

CIMAT. (2023). Women in Data Science (WiDS) 2023 [YouTube Video]. In *YouTube*. https://www.youtube.com/watch?v=QcP3tU3uMYo&t=5595s

Nguyen, B. (2021, February 6). *End-to-End Time Series Analysis and Forecasting: a Trio of SARIMAX, LSTM and Prophet (Part 1)*. Medium; Towards Data Science. https://towardsdatascience.com/end-to-end-time-series-analysis-and-forecasting-a-trio-of-sarimax-lstm-and-prophet-part-1-306367e57db8#:~:text=What%20is%20SARIMAX%3F,%2Daverage%20term%20(MA)

Gestión de energía en México. (2020). DatosMundial.com. https://www.datosmundial.com/america/mexico/balance-energetico.php