PREDICIÓN DE LA DEMANDA DE ENERGÍA ELÉCTRICA EN ANTIOQUIA

Estudiante: Anderson Sebastián Torres Sánchez
(https://github.com/Astorress/especializacion-analitica-monografia),
Asesor: Walter Mauricio Villa Acevedo.
Especialización en Analítica y Ciencia de Datos
Departamento de Ingeniería de Sistemas
Universidad de Antioquia, Colombia

Resumen – En este documento se presenta las predicciones iniciales de la demanda de energía eléctrica en Antioquia, utilizando métodos de machine learning tales como regresiones lineales, decision tree o arboles de decisión empleados en auto regresiones. Además se presenta el proceso de tratamiento de datos utilizado para la limpieza y preparaciones de los datos utilizados en los modelos de predicción, por ultimo las métricas de desempeño de los modelos y conclusiones y hallazgos encontrados.

Índice de Términos – API, Auto regresiones, demanda, IDEAM, energía, machine learning, MAPE, métricas, predicción, series de tiempo, XM.

I. INTRODUCCIÓN

El pronóstico de la demanda de energía eléctrica es un proceso fundamental para la toma de decisiones operativas y estratégicas en el mercado energético del país, cuya falta de precisión puede generar altos costos económicos [1]. Dichas predicciones pueden determinar de forma previa la carencia de energía eléctrica en el sistema y evitar de esta forma razonamientos de energía. Antioquia, al ser el segundo departamento con más población del país, que según cifras el DANE [2] representa el 13.5% de la población total del país, además de representar el 5.44% del territorio nacional [3] y poseer el segundo PIB más alto del país [4], representa, es uno de los focos más importantes del sector eléctrico del país.

Antioquia es el departamento con mayor capacidad de generación instalada en Colombia, teniendo según datos analizados del API de XM, una capacidad instalada de 5789MW, que constituye alrededor del 30% de la capacidad total instalada del país.

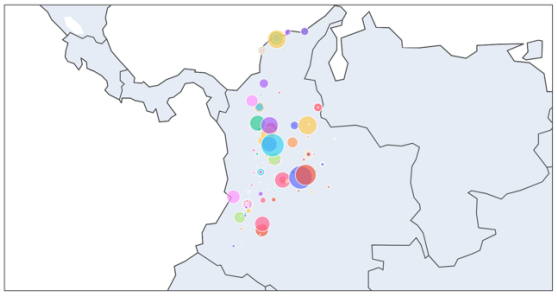

Figura 1: Mapa de la distribución de recursos de generación del país. Elaboración propia a partir de los datos del API de XM.

A su vez, Antioquia es la cuarta regiones del país con mayor demanda de energía eléctrica del país según los datos analizados del API de XM, superada por la región caribe, la región centro y la región oriente (se debe tener en cuenta que estas regiones son según la categorización dada por XM, donde la región caribe está conformada por los departamentos La Guajira, Cesar, Magdalena, Atlántico, Bolívar y Sucre, la región centro está formada por Bogotá D.C Cundinamarca y Meta, y la región oriente está formada por Norte de Santander, Santander y Boyacá).

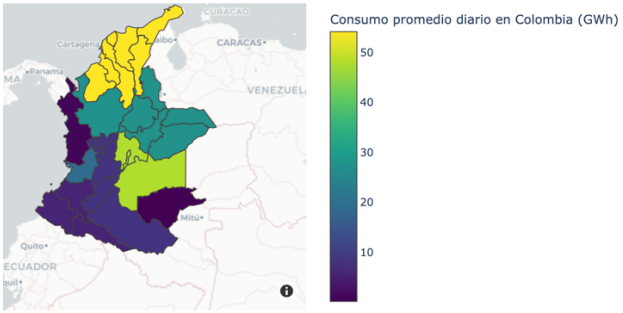

Figura 2: Mapa de calor según el consumo promedio diario de energía eléctrica en las regiones del país. Elaboración propia a partir de los datos del API de XM.

Teniendo así, Antioquia un consumo de energía similar al consumo combinado de departamentos como Norte de Santander, Santander y Boyacá.

Debido a lo expuesto anteriormente, se puede notar la gran importancia del departamento de Antioquia en el sector eléctrico del país y su gran relevancia en la elaboración de modelos de predicción de la demandada de energía eléctrica. Para lograr este objetivo, en la sección 2 de este documento se explica el proceso de extracción de diversas fuentes de datos, así como la preparación y limpieza de datos utilizados, en la sección 3 se realiza un análisis exploratorio de los datos, en la sección 4 se realizan diversos modelos de predicción de la demanda, utilizando métodos paramétricos y no paramétricos de machine learning, además de método propios de las series de tiempo, todos con sus respectivas métricas de validación, en la sección 5 se presenta las conclusión y por último en la sección 6 se presentan las referencias.

II. EXTRACIÓN, LIMPIEZA Y PREPARACIÓN DE DATOS

Para la extracción de los datos de demanda de energía eléctrica del país, en un inicio se planteó utilizar los datos existentes en el API de XM [5], dichos datos fueron planteados en un inicio como datos horarios de consumo, pero luego de una larga búsqueda, esto no fue posible, pudiendo solo encontrar datos de demanda diarios, desde el 1 de enero del 2018 hasta el 20 de mayo del 2023. Dichos datos constan de tres columnas, fecha en formato yyyy-mm-dd, región y consumo, teniendo en total 19660 registros para todas las

regiones del país y 1966 registros para Antioquia. Debido a la imposibilidad actual de obtener los datos horarios de demanda, se decide que la primera iteración del modelo sea realizada con datos diarios de demanda.

	fecha	region	consumo
0	2018-01-01	Antioquia	17.978862
1	2018-01-01	Caribe	41.495939
2	2018-01-01	Centro	30.340973
3	2018-01-01	Chocó	0.601889
4	2018-01-01	CQR	5.528218
...	...	...	...
19655	2023-05-20	Guaviare	0.159026
19656	2023-05-20	Oriente	36.335262
19657	2023-05-20	Sur	6.069049
19658	2023-05-20	THC	9.132351
19659	2023-05-20	Valle	19.560268

Figura 3: Tabla con los datos de la demanda de energía eléctrica en el país para todas sus regiones desde el 1 de enero del 2018 hasta el 20 de mayo del 2023.

Como se puede observar, se presenta una baja dimensionalidad de los datos, teniendo solo como variable predictora de la demanda de energía, la fecha, por tal motivo, se decidió agregar variables adicionales, que aumentan la dimensionalidad de los datos y aumentando las variables predictoras de la demanda. Para esto se utilizó la variable de fecha, extrayendo de esta información como el día de la semana, mes y una variable de fin de semana, donde se le asigna un 1 si es domingo o festivo y cero si es el resto de los días. Estas variables fueron adicionadas no solo con el objetivo de aumentar la dimensionalidad de los datos, también cumplen una función importante en la predicción, ya que por lo general los días de la semana son similares en demanda entre sí, es decir, los lunes tiene una demanda de energía similar entre ellos y en los fines de semana se tiene por lo general un consumo menor que en los días de semana.

Además de las variables antes mencionadas, se agregó la variable de temperatura promedio diario del departamento, está información fue extraída del API de Socrata, que se conecta con la plataforma de datos abiertos [6], extrayendo información de los medidores de temperatura administrados por el IDEAM. Esta variable se agregó con el objetivo se validar el supuesto de que la temperatura es un variable relevante para demanda de energía eléctrica, esto se puede notar en la Figura 2 de este documento, donde la región con una mayor de demanda de energía fue la zona caribe del país, zona donde la temperatura promedio es más alta que en otras regiones del país. Obteniendo así un total de siete variables en el dataset, cinco de ellas variables predictoras (fecha, mes, dia_semana, fin_de_semana y temperatura), una variable objetivo (consumo) y una variable irrelevante (region).

	fecha	region	consumo	mes	dia_semana	fin_de_semana	temperatura
0	2018-01-01	Antioquia	17.978862	1	1	0	18.610417
1	2018-01-02	Antioquia	22.849402	1	2	0	18.683333
2	2018-01-03	Antioquia	24.382203	1	3	0	19.962500
3	2018-01-04	Antioquia	24.470359	1	4	0	19.468750
4	2018-01-05	Antioquia	24.548823	1	5	0	18.760417
...	...	...	...	...	...	...	...
1961	2023-05-16	Antioquia	31.785307	5	2	0	NaN
1962	2023-05-17	Antioquia	31.150905	5	3	0	22.045833
1963	2023-05-18	Antioquia	32.391560	5	4	0	23.995833
1964	2023-05-19	Antioquia	32.464375	5	5	0	24.304167
1965	2023-05-20	Antioquia	29.338370	5	6	1	22.150000

Figura 4: Tabla con el dataset final.

Al dataset final obtenido, se le realiza un proceso de limpieza y preparación de datos, basado en detección de datos atípicos e imputación de datos nulos. En la detección de datos atípicos se encontraron datos atípicos naturales, pero no datos atípicos no naturales, es decir, datos que son posibles en las mediciones de demanda o temperatura, pero que se salen de la tendencia normal de los datos, por ejemplo, temperatura promedio de 29°C, no son temperaturas comunes en Antioquia (esto si se tienen en cuenta todos los municipios del departamento, se sabe que hay municipio con temperatura promedio mayores), pero no son valores imposibles de que sucedan. Por este motivo, no se eliminan datos atípicos. Por último, en la imputación de datos, se encontraron nulos en la variable de temperatura, en total el 4.6% de los datos de esta variable son datos nulos, por lo cual, se realiza una imputación de datos con el método KNNImputer.

Todo el proceso de extracción de datos de las API de XM y Socrata, limpieza y preparación de datos, así como los gráficos mostrados en la sección 1 y 2 de este documento, se encuentran en el archivo *"extraccion_procesamiento_y_limpieza_datos.ipynb"* del repositorio de Github. Además, toda la información utilizada y procesada es guarda en la carpeta data del mismo repositorio.

III. ANALISIS EXPLORATORIO

Como se ha venido mencionado, este es un problema de series de tiempo, las cuales son sucesiones de datos ordenados cronológicamente, espaciados a intervalos iguales o desiguales, la predicción en series de tiempo o forescasting consiste en predecir el valor futuro de una variable a través de valores pasados o empleando variables externas [7]. Por este motivo, las series de tiempo deben ser analizadas pensando en cómo se relacionan las variables externan con la variable objetivo y como se relaciona la variable objetivo con ella misma en intervalos de tiempo diferentes. Esto último es muy importante, ya que hay existir una relación entre la misma variable en instantes de tiempo diferentes, se tiene que validar la autocorrelación de la variable, ya que el valor de t_{n+1} dependerá de los t_n anteriores. En el caso de las variables externas, existe un tipo muy importantes para realizar las predicciones de series de tiempo y son las denominadas variables exógenas, estás son variables externas, cuyo valor se conoce a futuro [8], algunos ejemplos tipos puede ser festivos, mes del año, día de la semana, hora del día.

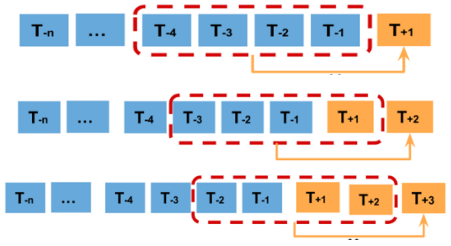

Figura 5: Grafico explicativo de la relación de una variable con ella misma en el tiempo [7].

Para análisis exploratorio se plantea validar los supuestos que fueron utilizados en la etapa de limpieza y preparación de datos, tales como correlaciones entre demanda de energía y temperatura, diferencias entre los comportamientos de la demanda de los fines de semana y días en semana y, gráficos de autocorrelación de datos.

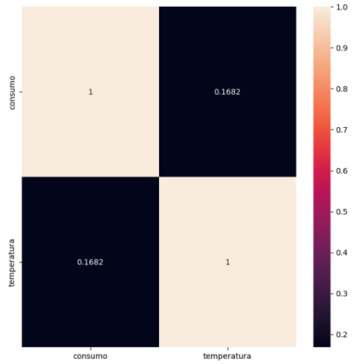

Figura 6: Matriz de correlación de Spearman.

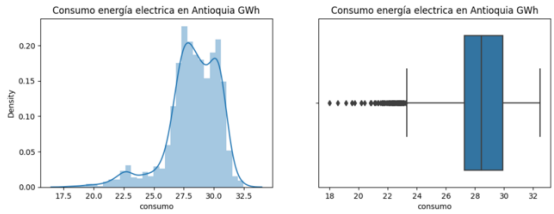

Figura 7: Distribución de la demanda de energía para los días en semana.

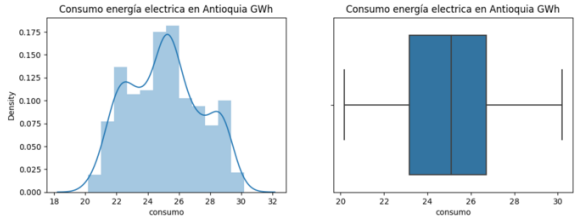

Figura 8: Distribución de la demanda de energía para los fines de semana.

Como se puede observar en la Figura 6 no existe correlación entre la variable de temperatura y demanda de energía. Además, al comparar las Figura 7 y 8, se puede notar que la distribución de los datos es diferente para los días en semana y los fines de semana, lo cual indica que esta variable es una buena cantidad como variable exógena del modelo.

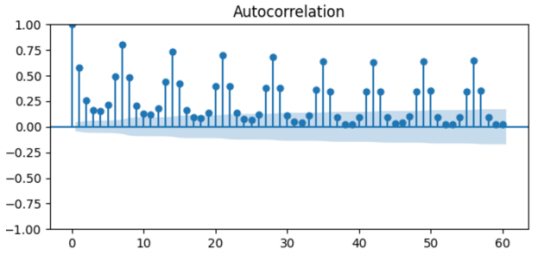

Figura 9: Grafico de autocorrelación de la demanda de energía teniendo en cuenta todos los días.

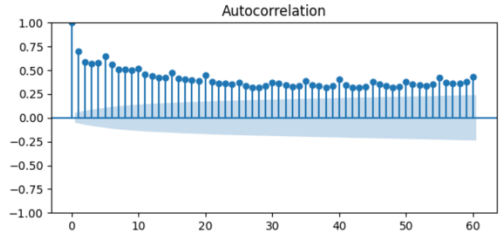

Figura 10: Grafico de autocorrelación de la demanda de energía para los días en semana.

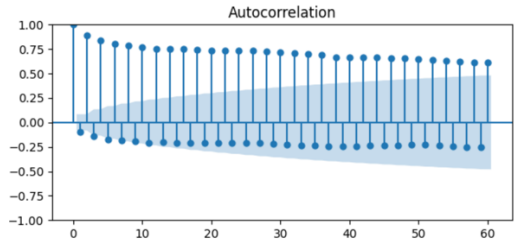

Figura 11: Grafico de autocorrelación de la demanda de energía para los fines de semana

En las Figura 9, 10 y 11, se observa que la autocorrelación de la demanda de energía se correlación mejor cuando se tiene ventanas de tiempo de múltiplos de 7, 5, y 2, respectivamente, es decir, los lunes entre ellos mismo tienen una alta correlación y a su vez indica que la mejor forma de predecir un lunes es con un lunes de una semana anterior, lo mismo para los días domingos, etc.

IV.MODELOS DE PREDICCIÓN

Cuando se realizan modelos de predicción de series de tiempo, es necesario realizar predicción por medio de modelos o métodos de auto regresión, que consisten en métodos que toman en cuenta la temporalidad de los datos y realizan las predicciones a partir de estos datos temporales. En este caso, se usará las librerías skforecast, que es una librería especializada de Python para este tipo de problemas.

Esta librería contiene una gran cantidad de métodos y funciones similares a sklearn, pero en series temporales, tales como lo son predicciones, optimización de hiperparámetros con grid search, etc. Para obtener un modelo de predicción es necesario utilizar modelos de regresión tales como regresión lineal, regresión lineal L2, decision tree regressor, random forest regressor, etc. y para luego ser pasado como parámetro de entrada de una función de auto regresión, es decir, al final se usan los modelos de machine learning comunes en auto regresiones.

Para esta predicción inicial de la demanda de energía eléctrica, se usarán 10 modelos de auto regresión, en los cuales se usarán modelos de entrada la regresión lineal y random forest regressor, estos modelos serán puestos a prueba con datos normalizados, datos sin normalizar, optimizando hiperparámetros en el caso del random forest y teniendo en cuenta variables exógenas como la temperatura o el fin de semana. Con el objetivo de evaluar el mejor modelo, se utilizan las métricas MSE, RMSE, MAE, MedAE y MAPE, siendo esta última a la que más relevancia se le dará.

Otra particularidad de las series de tiempo y es que la división entre train y test no se puede darse de forma aleatoria, debe darse de forma ordenada, es decir, los primero 80% de registros son para el train y el 20% restante son para el test. En este caso todos los modelos fueron entrenado con el 80% de los registros y el 20% restante son para el test.

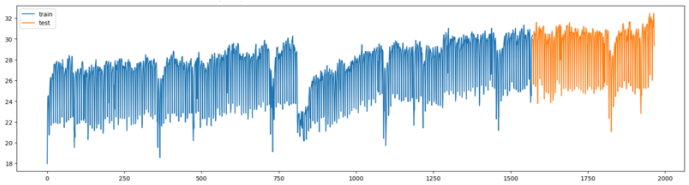

Figura 12: Grafico de la partición de los datos de train y test de la demanda de energía eléctrica.

En total los datos de train cuentan con 1572 registros, que van desde el 1 de enero del 2018 hasta el 21 de abril del 2022 y los datos de test cuentan con 394 registros que van desde el 22 de abril del 2022 hasta el 20 de mayo del 2023.

Ahora, si bien los modelos de auto regresión son muy útiles en series de tiempo, cuando con una gran desventaja y es que debidos a su propia naturaleza, dichos modelos van perdiendo variabilidad conforme se avanza con la predicción en el tiempo, debido a que las predicciones van dependiendo predicciones realizadas antes, haciendo que estos modelos sean utilices a corto plazo y poco eficientes a largo plazo, tal como se muestra en la Figura 5. Por este motivo, solo le prestará atención a las primeras 2 semanas de la predicción y sobre estos datos se aplicarán las métricas de validación de los modelos.

A continuación, se muestra una tabla de resumen los modelos realizados y sus respectivas métricas de validación.

	modelo	mse	rmse	mae	medAE	mape
0	Auto regresión con random forest con variable exógena de fin de semana	1.7478	3.0550	1.3921	1.2667	0.0472
1	Auto regresión con random forest con todos los dias con hiperparámetros con datos norm...	1.7446	3.0435	1.5836	1.5306	0.0527
2	Auto regresión con regresión lineal con variable exógena de temperatura	2.2170	4.9153	1.7537	1.3621	0.0592
3	Auto regresión con regresión lineal con variable exógena de fin de semana	2.0188	4.0755	1.8326	1.6781	0.0626
4	Auto regresión con random forest con todos los dias	2.3623	5.5803	1.8751	1.5756	0.0661
5	Auto regresión con regresión lineal con todos los dias	2.3733	5.6326	2.0152	1.9980	0.0674
6	Auto regresión con regresión lineal con todos los dias con datos normalizados	2.3733	5.6326	2.0152	1.9980	0.0674
7	Auto regresión con random forest con todos los dias con hiperparámetros	2.5618	6.5629	2.1408	1.6289	0.0713
8	Auto regresión con random forest con todos los dias con datos normalizados	2.5031	6.2657	2.2520	2.1652	0.0780
9	Auto regresión con random forest con variable exógena de temperatura	2.5495	6.5001	2.3513	2.3030	0.0805

Figura 14: Resumen de las métricas para los modelos realizados.

Como se puede observar el mejor modelo fue el realizado con random forest con variable exógena el fin de semana. A continuación, se presentan los gráficos de la predicción.

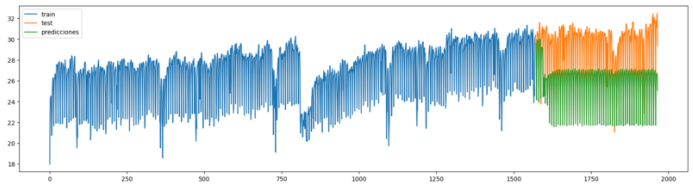

Figura 15: Grafico de la predicción del mejor modelo en todo el intervalo de test.

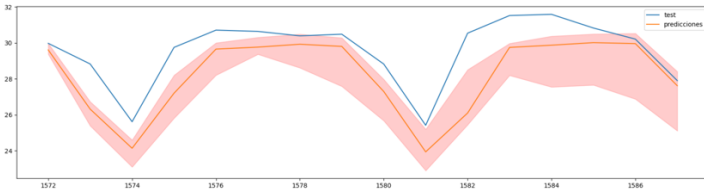

Figura 16: Grafico de la predicción del mejor modelo en las primeras dos semana de la predicción.

Todo el proceso de análisis exploratorio inicial de los datos, el detalle de los modelos de predicción de series de tiempo, así todo los gráficos utilizados en la sección 3 y 4 de este documento, se encuentran en el archivo "*analisis_datos_y_generacion_modelos.ipynb"* del repositorio de Github. Además, toda la información utilizada y procesada es guarda en la carpeta data del mismo repositorio.

V. CONCLUSIONES

- Si bien los datos que pudieron obtener no fueron los esperados en un inicio los modelos de predicción obtenidos fueron muy buenos según la métrica del MAPE ya que se encuentran por debajo del 0.2 o del 20%.
- La variable de temperatura a diferencia de como se pensaba en un inicio no tuvo relevancia en la predicción. Esto puede darse a que los datos del API de Socrata del IDEAM no tiene datos todos los municipios de Antioquia y se sacan de dicho promedio de temperatura muchos municipios.
- La variable exógena del fin de semana es quizás uno de los mejores predictores para este tipo de problemas, para la próxima iteración de este problema, se realizará un calendario de días con días festivos en todo el intervalo de los datos para tener en cuenta también estos días como variable exógena.
- Dado que el problema principal, es la predicción de la demanda de energía con datos horarios, se está buscando la forma de obtener estos datos con el ente regulador del mercado XM, para así mejorar y obtener mejores modelos que puedan ser usados en caso reales y dato que se pueden tener datos de temperatura también horarios con el API de Socrata, se puede mejorar la relación entre la variable de la demanda de energía eléctrica y la temperatura.
- En la especialización se debe incluir algunos temas relacionados con series de tiempo, ya que son un tipo de problemas comunes en el mundo real y la información que se brinda es nula, haciendo que en este tipo de problemas se deba gastar una gran cantidad de tiempo adicional en la búsqueda y aprendizaje.

VI.REFERENCIA

[1] *D. O. Garzón Medina, G.A. Marulanda García, "Estimación del consumo eléctrico colombiano en el corto plazo y largo plazo empleando regresiones multivariables y series temporales", 2017, pág. 1*

[2] *DANE, "RESULTADOS CENSO NACIONAL DE POBLACIÓN Y VIVIENDA 2018", 2018, pág. 52.*

[3] *TodaColombia, "Departamento de Antioquia", 2019.* https://www.todacolombia.com/departamentos-de-colombia/antioquia/index.html

[4] *DANE, "PIB por departamento", 2023.* *http://www.dane.gov.co/index.php/estadisticas-por-tema/cuentas-nacionales/cuentas-nacionales-departamentales*

[5] *XM. Equipo de Analitica, "Documentación API".* https://github.com/EquipoAnaliticaXM/API_XM

https://sinergox.xm.com.co/Paginas/Home.aspx

[6] *Datos Abierto.* https://www.datos.gov.co/

[7] J. *Amat Rodrigo, J. Escobar Ortiz, "Forecasting series temporales con Python y Scikit-learn" , 2021.*

https://www.cienciadedatos.net/documentos/py27-forecasting-series-temporales-python-scikitlearn.html

[8] J. *Amat Rodrigo, J. Escobar Ortiz, "Predicción (forecasting) de la demanda eléctrica con Python" , 2021.*
https://www.cienciadedatos.net/documentos/py29-forecasting-demanda-energia-electrica-python