

CIENCIA DE DATOS APLICADA



Entrega No. 2 Proyecto CDA

INTEGRANTES

Nombre	Correo Electrónico
Juliana Andrea Galeano Caicedo	ja.galeanoc1@uniandes.edu.co
Juan Nicolás Estepa Guzmán	j.estepa@uniandes.edu.co
Santiago Rodríguez Cruz	s.rodriquez52@uniandes.edu.co
Harvy Benítez Amaya	h.benitez@uniandes.edu.co

INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
MAESTRÍA EN INGENIERÍA DE LA INFORMACIÓN
UNIVERSIDAD DE LOS ANDES
BOGOTÁ D. C.

1. DEFINICIÓN DE LA PROBLEMÁTICA Y ENTENDIMIENTO DEL NEGOCIO

El Ministerio de Minas y Energía (MinEnergía) y la Unidad de Planeación Minero-Energética (UPME) enfrentan dificultades con los modelos actuales de pronóstico de demanda energética a corto plazo, los cuales no son precisos a nivel regional. Esto genera sobrecostos debido a la subestimación y sobreestimación de la demanda, afectando la planificación y operación del sistema energético. Sin embargo, la Ciencia de Datos ofrece una oportunidad para mejorar la precisión de los pronósticos, considerando variables externas y eventos atípicos, lo que optimizaría la gestión energética y reduciría costos.

La estrategia del MinEnergía y UPME incluye el uso de modelos estadísticos y econométricos, la consulta con actores clave del sector energético y el análisis de diferentes escenarios económicos. También incorporan tecnologías emergentes y análisis regionales para una planificación más precisa del sistema. El sector eléctrico colombiano está regido por las Leyes 142 y 143 de 1994, que definen la estructura y funciones del sector, con entidades clave como el MinEnergía, UPME, CREG, XM y las Empresas de Servicios Públicos.

Aunque MinEnergía lidera la dirección del sector y UPME realiza la planeación, los operadores del sistema (XM) son responsables de recolectar datos y predecir la demanda energética mediante un sistema integral que monitorea en tiempo real la generación y consumo de electricidad.

Objetivo del proyecto

Desarrollar un modelo analítico que utilice datos históricos y variables exógenas para predecir la demanda de energía a corto y mediano plazo en Colombia, mejorando la precisión de los pronósticos actuales.

Métricas de negocio

Tolerancia de error de pronóstico $\leq 5\%$: El Consejo Nacional de Operación (CNO), conformado por generadores, transmisores y distribuidores de energía, supervisa las desviaciones de pronóstico de los operadores de red y establece una tolerancia de error de pronóstico $\leq 5\%$, de lo contrario, estas deben ser reportadas a la Superintendencia de Servicios Públicos, puesto que, generan costos adicionales para el sistema.

Reducción de costos operativos: La mejora en la precisión de los pronósticos contribuye a la reducción de costos operativos al minimizar pérdidas por sobreproducción o falta de suministro. Esto optimiza la generación y distribución de energía, permitiendo reinvertir recursos en infraestructura y tecnología.

2. IDEACIÓN

El producto consiste en un modelo analítico predictivo que se ofrece a través de una API. Como corresponde a los operadores del sistema (XM / Expertos en mercado) recolectar y generar datos para predecir la demanda de energía eléctrica en Colombia, estos

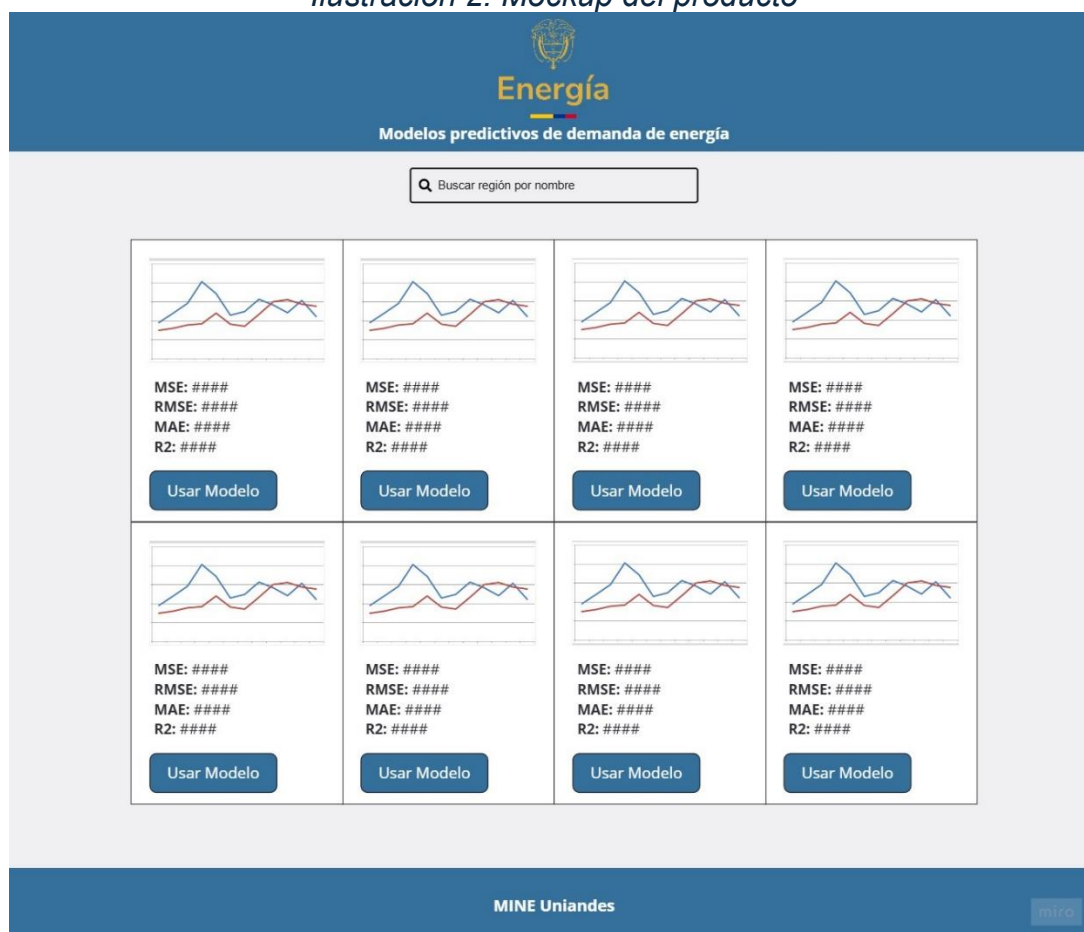
podrían ser considerados como los “usuarios finales” del modelo. Sin embargo, es necesario aclarar que los datos que generará el modelo y que por cuestiones regulatorias deben ser de acceso público (*posiblemente a través del portal web <https://www.integrarme.gov.co>*), pueden ser utilizados por diversos actores del sector energético, incluidos el MInEnergía, empresas de generación, transmisión y distribución, así como las entidades reguladoras CREG y UPME.


Ilustración 1. Diagrama de componentes



Fuente: Elaboración propia.

Ilustración 2. Mockup del producto




Energía
Modelos predictivos de demanda de energía

Seleccionar archivo


datos_a_predecir.csv

Predecir

Volver al inicio

MINE Uniandes

muro


Energía
Modelos predictivos de demanda de energía

Descargar Excel

Descargar CSV - XM

Volver al inicio

Predicciones:

Fecha	Predicción
aaaa-mm-dd	#####
aaaa-mm-dd	#####
aaaa-mm-dd	#####
aaaa-mm-dd	#####
aaaa-mm-dd	#####

MINE Uniandes

muro

Fuente: Elaboración propia.

Actualmente, el modelo se basa en técnicas de series de tiempo, específicamente regresión lineal multivariada, lo que permite realizar pronósticos sobre la demanda energética. Sin embargo, el enfoque presenta varios desafíos:

- Las limitaciones en los pronósticos a corto plazo dificultan la capacidad de respuesta ante cambios repentinos en la demanda
- La baja granularidad de los modelos impide obtener pronósticos precisos a nivel regional o por sectores específicos
- La complejidad en el manejo de variables exógenas, como factores económicos y climáticos, complica la calibración del modelo.

Estos "dolores" resaltan la necesidad de un enfoque más avanzado y flexible que pueda mejorar la precisión y utilidad de las predicciones en el sector energético.

El desarrollo del nuevo modelo analítico predictivo sigue un proceso estructurado que abarca varios requerimientos esenciales. En primer lugar, se inicia con la recopilación de

datos históricos, que son fundamentales para entender patrones y tendencias en la demanda energética. Luego, se procede al procesamiento de estos datos, seguido de su división en conjuntos que permitirán la construcción y entrenamiento del modelo. Este proceso incluye la validación del modelo para asegurar que cumpla con los estándares necesarios antes de pasar a la evaluación final, donde se determinará su efectividad y precisión.

Para la implementación del modelo, se requieren diversos componentes tecnológicos. Esto incluye un servidor dedicado para el entrenamiento del modelo, que garantice el rendimiento adecuado durante el procesamiento de grandes volúmenes de datos. Además, se necesitará un entorno de desarrollo que facilite la creación y ajuste del modelo, así como una interfaz de programación de aplicaciones (API) que permita su integración y acceso por parte de los usuarios. Por último, se proporcionará documentación detallada para facilitar el uso e implementación del modelo, asegurando que todos los actores involucrados puedan beneficiarse de sus capacidades analíticas de manera efectiva.

3. RESPONSIBLE

El modelo analítico predictivo para el sector energético se basa en datos que son de dominio público y que provienen de entidades reguladas con la legislación vigente en Colombia. No existen temas regulatorios adicionales que limiten la privacidad de estas fuentes de datos, ya que, están respaldadas por el marco normativo vigente y que está diseñado para fomentar la transparencia y el acceso a la información. En este contexto, el uso de datos de dominio público para el modelo analítico no solo es legal, sino que también se alinea con los principios de transparencia y acceso a la información que rigen en Colombia. Esto garantiza que el modelo se construya sobre bases sólidas, utilizando información que está disponible para todos y que puede contribuir significativamente a la toma de decisiones informadas en el sector energético, sin vulnerar la privacidad de los datos personales.

4. ENFOQUE ANALÍTICO

Pregunta de negocio

¿Cómo podemos utilizar técnicas avanzadas de analítica y Machine Learning para generar pronósticos más precisos y granulares de la demanda de energía eléctrica en Colombia para horizontes de 1, 3 y 6 meses, aprovechando datos históricos de consumo y variables exógenas relevantes, con el fin de optimizar la toma de decisiones en procesos claves como la planeación de la generación, la operación del sistema interconectado y la gestión comercial de las empresas de energía?

Técnicas propuestas

- LSTM (Long Short-Term Memory): Red neuronal recurrente que captura dependencias a largo plazo en series temporales, ideal para predecir la demanda energética debido a su capacidad para recordar patrones temporales complejos.

- **MLP (Multilayer Perceptron):** Red neuronal con múltiples capas que modela relaciones no lineales entre variables. Útil para predecir la demanda energética con factores adicionales como el clima o eventos especiales.
- **CNN (Convolutional Neural Networks):** Red neuronal que extrae características locales de los datos, eficaz para detectar patrones o picos en la demanda energética mediante la identificación de tendencias a corto plazo.

Métricas de evaluación

Error Absoluto Medio (MAE): Mide el error promedio en las predicciones, indicando cuán cerca están las estimaciones de los valores reales. Un MAE bajo sugiere alta precisión en las predicciones.

Raíz del Error Cuadrático Medio (RMSE): Identifica desviaciones significativas en las predicciones al penalizar errores grandes, lo que ayuda a resaltar fallos críticos en el modelo.

5. RECOLECCIÓN DE DATOS

A continuación, se presentan las fuentes de datos seleccionadas y se describe su utilidad en el modelo analítico predictivo de la demanda de energía.

Tabla 1. Fuentes de datos

VARIABLE OBJETIVO				
NOMBRE	FUENTE	COMPONENTE	TIPO DE DATO	DESCRIPCIÓN
Demanda de energía	XM (Operador del Sistema / Expertos en Mercado)	Energético	Datos estructurados	Un (1) dato por día por región.

VARIABLES INDEPENDIENTES				
NOMBRE	FUENTE	COMPONENTE	TIPO DE DATO	DESCRIPCIÓN
Proyección de la población	DANE	Demográfica	Datos estructurados	Un (1) dato anual por ciudad.
IPC	DANE	Económica	Datos estructurados	Un (1) dato mensual por ciudad.
PIB	DANE	Económica	Datos estructurados	Un (1) dato anual por ciudad.
Temperatura seca	IDEAM	Climática	Datos estructurados	Temperatura media diaria por estación de monitoreo.
Calendario festivos Colombia	Nager.Date	Calendario	Datos estructurados	Días festivos y fechas especiales que alteren la demanda de forma temporal.

Fuente: Elaboración propia.

Demanda de energía: Esta variable proporciona datos históricos de consumo que reflejan patrones y tendencias en el uso de energía. Analizarla permite ajustar el modelo a comportamientos pasados y prever cambios futuros en la demanda.

Proyección de la población: La población es un factor crítico que impacta directamente en la demanda energética. A medida que la población crece, también lo hace el consumo

de energía, especialmente en sectores residenciales y comerciales. Incluir esta variable permite captar cómo los cambios demográficos influyen en la demanda.

Índice de Precios al Consumidor (IPC): El IPC refleja la inflación y el costo de la vida, lo que puede afectar la capacidad de los consumidores para gastar en energía. Un aumento en el IPC podría llevar a una disminución en el consumo energético, especialmente en segmentos vulnerables. Esta variable ayuda a entender la relación entre la economía y la demanda.

Producto Interno Bruto (PIB): El PIB es un indicador clave del crecimiento económico. Un PIB en crecimiento suele correlacionarse con un aumento en la actividad industrial y comercial, lo que generalmente eleva la demanda de energía. Incorporar el PIB permite capturar la dinámica entre el crecimiento económico y el consumo energético.

Temperatura seca: Las variaciones en la temperatura afectan significativamente la demanda energética, especialmente en climas donde la calefacción y la refrigeración son necesarias. Incorporar datos climáticos ayuda a anticipar picos en la demanda durante períodos de temperaturas extremas.

Calendarios festivos Colombia: Los días festivos pueden influir en los patrones de consumo, ya que durante estos períodos la actividad comercial y el comportamiento del consumidor pueden cambiar notablemente. Tener en cuenta estos días permite ajustar las proyecciones de demanda para reflejar estas variaciones estacionales.

6. ENTENDIMIENTO DE LOS DATOS

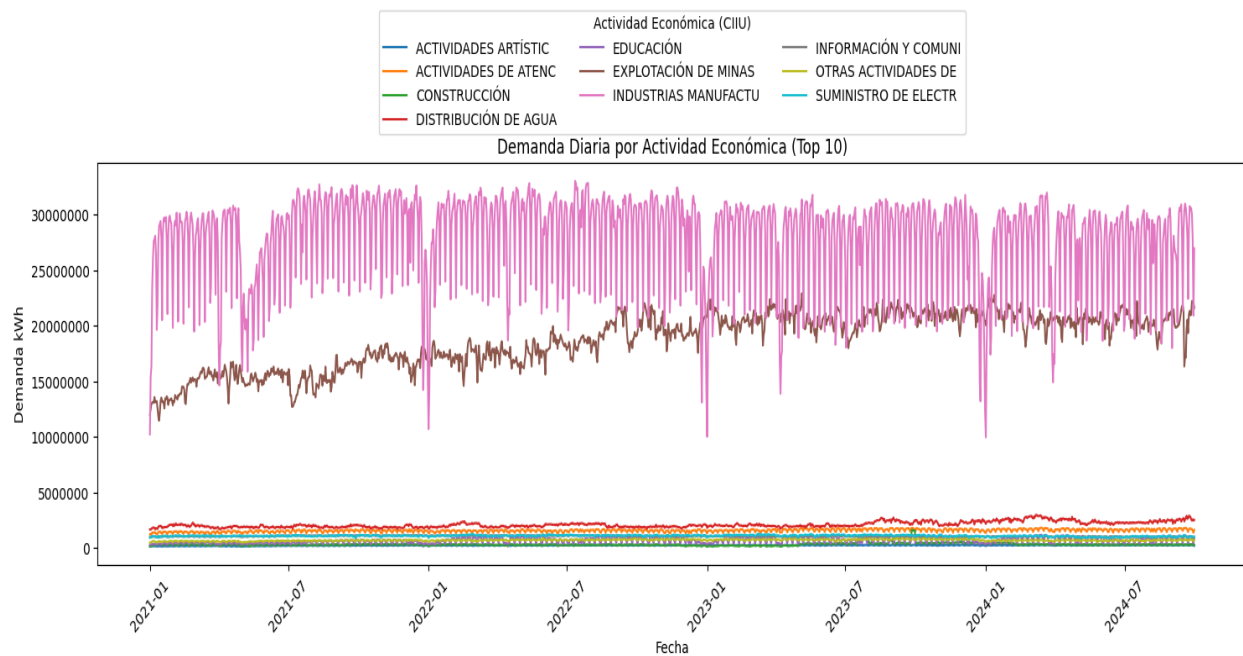
Se realizó análisis no gráfico univariado para cada una de las variables (*variable objeto y variables independientes*) revisando las estadísticas como la desviación estándar, media, máximos y mínimos, ect. Así mismo, se hizo análisis no gráfico bivariado para el entender el comportamiento de la variable objetivo en relación con cada una de las variables independientes, utilizando las técnicas estadísticas mencionadas con anterioridad.

Por otra parte, se realizó análisis gráfico univariado y bivariado para comprender el comportamiento de cada una de las variables y, la relación entre la variable objetivo y las variables independientes, respectivamente. Haciendo uso de diagramas de barra, box plot, gráficas de líneas, etc.

Adicionalmente, se ejecutaron técnicas de análisis de calidad para los datos de cada variable, tales como, validación de valores nulos, transformaciones de cada columna, sumar los valores de demanda horaria para obtener un dato de demanda por día, mapear las regiones del DANE con respecto a las usadas por XM para realizar análisis, etc.

A continuación, se presentan algunos gráficos relacionados con la variable demanda de energía, si se desea conocer un análisis más técnico y detallado de cada una de las variables, el lector puede remitirse al Notebook llamado *Limpieza Y Entendimiento*:

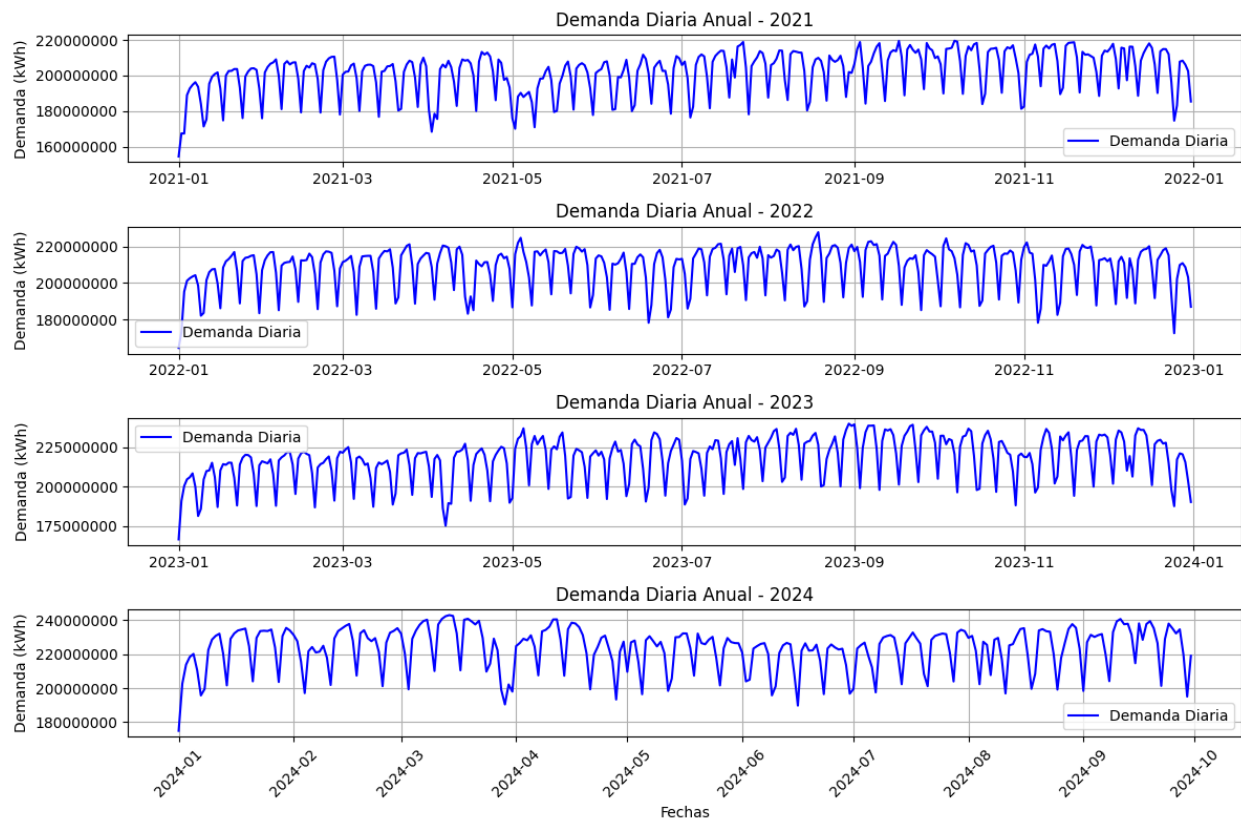
Ilustración 3. Demanda diaria por actividad económica (Top 10)



Fuente: Elaboración propia.

Según la *Ilustración 3*, se puede observar que la actividad económica *Industrias Manufactureras* presenta una alta variabilidad con respecto a *Explotación de minas y canteras*. Adicionalmente, las *Industrias Manufactureras* presentan una leve tendencia de reducción de la demanda de energía, mientras que, para la *Explotación de minas y canteras* se ve un pronunciado crecimiento de la demanda de energía en los últimos años.

Ilustración 4. Demanda diaria anual



Fuente: Elaboración propia.

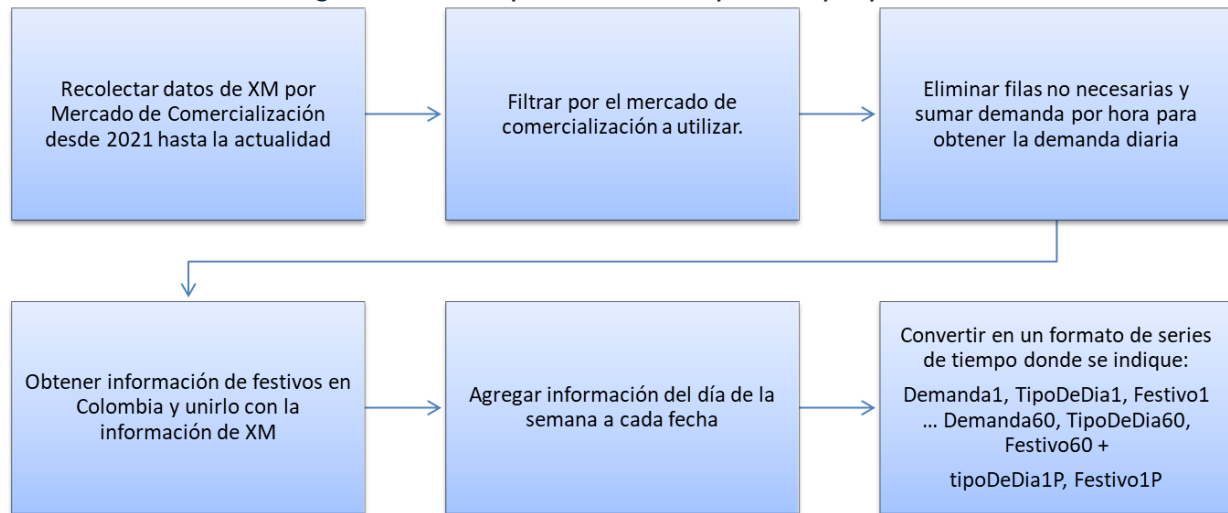
Por último, en la *Ilustración 4* se puede observar que existe un patrón en el comportamiento de la demanda de energía tanto a nivel semanal como mensual que se repite a lo largo de los años, lo cual, sugiere que cierto grado de estacionalidad en la demanda de energía, es decir, su comportamiento es recurrente y predecible para intervalos de tiempos regulares.

7. PREPARACIÓN DE DATOS

Como fuentes de datos para los modelos se seleccionó la demanda de energía por mercado de comercialización en mutuo acuerdo con el cliente, cuya fuente es XM (*operador del sistema*) y que se extraen en formato .csv a través de una API. Adicionalmente, se agregaron los días festivos del año, así como las etiquetas de cada día (*lunes, martes, miércoles, etc*) esto con la finalidad de que los modelos puedan identificar patrones dada la estacionalidad que presenta la variable objetivo en un nivel de granularidad semanal. Se descartaron otras variables consideradas inicialmente como IPC, PIB, Proyección de la población, ya que, no presentaban correlación significativa con la demanda de energía. Así mismo, la variable de Temperatura seca se descartó, puesto que, su comportamiento en el último año ha tenido fluctuaciones significativas con respecto a años pasados debido al cambio climático, lo que podría aumentar la incertidumbre en los modelos. Por último, se omitieron los datos generados durante la pandemia debido al cambio drástico en las tendencias de demanda eléctrica en el país, que afectaban negativamente el desempeño de los modelos.

Por tanto, el proceso de preparación de datos (*ver Ilustración 5*) inicia con recolectar los datos de demanda de energía por mercado de comercialización provistos por XM desde el 01 de enero de 2021 hasta el 30 de septiembre de 2024. Posteriormente, se filtran los datos seleccionando uno (1) de los veinticuatro (24) mercados de comercialización existentes (*ej: Bogotá, Antioquia, Caribe Sol, Caribe Mar, etc.*). Seguidamente, se eliminan las filas no necesarias, es decir, las que pertenecen a un mercado de comercialización diferente al seleccionado. A continuación, se transforman los valores de demanda horaria (*24 valores por día*) sumándolos para obtener el valor total de demanda diaria para un mercado de comercialización en específico, esto de acuerdo con la solicitud del cliente que requiere la predicción de la demanda energética desagregada en días y no en horas. Cabe resaltar que, los datos provistos por XM son de una calidad excelente considerando que no presentan datos faltantes, tampoco valores atípicos a excepción de la naturaleza propia del contexto. Por otra parte, se obtienen la información de los días festivos en Colombia y se integran con los datos de XM (*demanda de energía*). Después, se agrega la información del día de la semana a cada fecha.

Ilustración 5. Diagrama de bloques funcional para la preparación de los datos



Fuente: Elaboración propia.

Por último, se convierte esta información en un formato de series de tiempo equivalente a una matriz, en donde se establece una ventana de tiempo de sesenta (60) días de entrada y treinta (30) días de predicción para cada fila (ver *Ilustración 6*). Adicionalmente, cada día se compone de tres (3) columnas (*Valor Demanda, Tipo de Día, Festivo*), dando un total de noventa (90) columnas. Es importante mencionar que, a medida que se incrementa el número de filas, se está moviendo el arreglo de noventa (90) días en un día calendario, es decir, la primera fila inicia el 01 de enero de 2021 y finaliza el 01 de abril de 2021, por ende, la segunda fila inicia el 02 de enero de 2021 y finaliza el 02 de abril de 2021 y, de esta forma se continua hasta que la fila número 1277 inicia el 02 de julio de 2024 y finaliza el 30 de septiembre de 2024. Todo este proceso se realiza con el fin de tener una matriz de 1278 filas con 90 columnas, generando los datos suficientes para entrenar y testear los modelos.

Ilustración 6. Evidencia preparación datos previo al entrenamiento del modelo

DATASET ORIGINAL																					
ID		Values_code	Values_Market Type	Values_Hour01	Values_Hour02	Values_Hour03	Values_Hour04	Values_Hour05	Values_Hour06	Values_Hour07	...	Values_Hour17	Values_Hour18	Values_Hour19	Values_Hour20	Values_Hour21	Values_Hour22	Values_Hour23	Values_Hour24	Date	Total
0	Mercado Comercializacion	ANTIOQUIA	NO REGULADO	349886.48	349905.59	346123.47	343345.25	348379.97	354034.69	344095.67	...	334811.35	317993.31	337670.89	324738.84	310200.22	294013.40	284725.85	281685.94	12/10/2024	8066362.72
1	Mercado Comercializacion	ANTIOQUIA	REGULADO	707721.11	660481.20	637402.95	629450.09	654538.81	692805.27	774391.54	...	1018159.74	1017339.24	1081872.32	1065654.79	1011911.91	941329.93	859420.89	781332.85	12/10/2024	21674553.45
2	Mercado Comercializacion	ARAUCA	NO REGULADO	91036.91	91590.62	91659.21	91489.46	91490.58	91216.39	90950.01	...	92528.85	92751.55	92631.76	93495.21	94211.71	94155.10	93672.21	93592.85	12/10/2024	2216950.49
3	Mercado Comercializacion	ARAUCA	REGULADO	38565.24	35412.12	34168.77	32972.37	32446.60	30796.52	30920.05	...	38780.21	39839.09	43575.74	43956.48	43686.89	40872.81	37321.19	35699.89	12/10/2024	907668.67
4	Mercado Comercializacion	BAJO PUTUMAYO	NO REGULADO	177.33	190.67	223.32	145.79	193.25	176.79	169.23	...	233.24	209.82	205.00	199.57	200.19	161.84	150.07	148.43	12/10/2024	5009.33
...
78361	Mercado Comercializacion	TULUA	NO REGULADO	6724.49	6768.58	7155.24	7191.53	7188.43	7088.42	7457.83	...	8772.95	8986.19	8707.40	8063.86	8054.62	7513.59	7124.25	7129.52	30/01/2021	185050.97
78362	Mercado Comercializacion	TULUA	REGULADO	18390.38	17151.86	16335.97	15906.30	15975.03	16326.97	14961.11	...	23980.63	23290.82	25425.79	27203.33	26087.47	25008.88	22982.94	20528.09	30/01/2021	511240.84
78363	Mercado Comercializacion	VALLE DEL CAUCA	NO REGULADO	140214.79	139195.75	135068.54	135036.53	133085.87	131396.81	123297.54	...	128185.13	126952.77	133716.06	132256.19	129859.09	123826.92	119728.54	115191.51	30/01/2021	3112450.67
78364	Mercado Comercializacion	VALLE DEL CAUCA	REGULADO	137699.15	128748.65	123197.02	121009.42	120629.68	124766.35	133469.98	...	193627.53	188065.57	196328.77	214746.40	207405.54	196495.13	179530.27	157734.66	30/01/2021	4104754.50
78365	Mercado Comercializacion	VALLE DEL SIBUNDÓY	REGULADO	759.93	704.33	681.81	671.37	696.60	790.98	814.29	...	1124.56	1153.15	1440.42	1700.50	1533.88	1303.77	1056.44	852.55	30/01/2021	25764.79

DATASET MODIFICAD (Ej: Matriz Mercado de Comercialización Antioquia)																					
	fe-1	d-1	t-1	fe-2	d-2	t-2	fe-3	d-3	t-3	fe-4	...	r-27	fe-88	d-88	r-28	fe-89	d-89	r-29	fe-90	d-90	r-30
0	1	4	19727162.18	0	5	21347232.04	0	6	22341657.39	0	...	24327810.66	0	0	28631930.20	0	1	29254290.77	0	2	28365072.68
1	0	5	21347232.04	0	6	22341657.39	0	0	26743404.23	0	...	28631930.20	0	1	29254290.77	0	2	28365072.68	1	3	23777623.01
2	0	6	22341657.39	0	0	26743404.23	0	1	27449375.65	0	...	29254290.77	0	2	28365072.68	1	3	23777623.01	1	4	21672313.28
3	0	0	26743404.23	0	1	27449375.65	0	2	27704951.26	0	...	28365072.68	1	3	23777623.01	1	4	21672313.28	0	5	23196616.81
4	0	1	27449375.65	0	2	27704951.26	0	3	27687397.70	0	...	23777623.01	1	4	21672313.28	0	5	23196616.81	0	6	23173731.79
...
1273	0	3	30626218.53	0	4	30307634.51	0	5	28767680.05	0	...	30855631.61	0	6	26663537.61	0	0	31895413.71	0	1	32893159.84
1274	0	4	30307634.51	0	5	28767680.05	0	6	25238821.82	1	...	26663537.61	0	0	31895413.71	0	1	32893159.84	0	2	32734016.50
1275	0	5	28767680.05	0	6	25238821.82	1	0	25421770.50	0	...	31895413.71	0	1	32893159.84	0	2	32734016.50	0	3	32185419.82
1276	0	6	25238821.82	1	0	25421770.50	0	1	30347331.50	0	...	32893159.84	0	2	32734016.50	0	3	32185419.82	0	4	31752295.56
1277	1	0	25421770.50	0	1	30347331.50	0	2	30960780.79	0	...	32734016.50	0	3	32185419.82	0	4	31752295.56	0	5	29355310.64

Fuente: Elaboración propia.

Nota: Si se desea conocer con más detalle el proceso realizado para preparación de los datos, así como, la evidencia de este proceso para los 24 modelos de predicción de la demanda energética por mercado de comercialización, puede remitirse al capítulo 2 del Notebook Modelo Final.

8. ESTRATEGÍA DE VALIDACIÓN Y SELECCIÓN DEL MODELO

La estrategia de experimentación para entrenar y seleccionar el mejor modelo consta de: 1. Definición de objetivos y métricas de evaluación (*descritas a continuación*), 2. Preprocesamiento de los datos (*ver capítulo 7*), 3. División del conjunto de datos (*descritas a continuación*), 4. Selección de modelos y entrenamiento, 5. Evaluación de modelos y 6. Selección del modelo (*ver capítulo 8*).

En primer lugar, se define el objetivo del desempeño del modelo, el cual, es predecir con el mínimo error posible la demanda diaria energética en el país para cada mercado de comercialización en un horizonte de tiempo de treinta (30) días calendario.

Para esto se utilizarán métricas propias de los modelos, tales como, RMSE y MAE. Es importante mencionar que, para este caso de uso el RMSE (*Error Cuadrático Medio de la Raíz, por sus siglas en inglés*) es un buen indicador de desempeño, en la medida en que se requiere que los modelos sean sensibles a valores atípicos y penalicen fuertemente los errores grandes, buscando evitar predicciones erróneas significativas. Asimismo, el indicador MAE (*Error Absoluto Medio, por sus siglas en inglés*) es otro excelente indicador de desempeño dado el caso de uso, ya que, este mide el error promedio sin importar la dirección, es decir, valores negativos para la sobre-estimación y positivos para la sub-estimación de la demanda diaria de energía, lo cual, resulta bastante útil dado que la exactitud de la predicción es importante pero la dirección de los errores no son tan relevantes como la magnitud de estos.

También, se consideró como métrica de negocio la tolerancia de error de pronóstico $\leq \pm 5\%$ dada por el CNO (*Consejo Nacional de Operación*), puesto que, tener errores en la predicción de la demanda diaria energética por fuera de este umbral genera sobrecostos al sistema.

Por una parte, la sub-estimación ($> +5\%$) de la demanda genera desabastecimiento de energía ocasionando apagones o racionamientos de electricidad que afectan tanto a empresas como consumidores finales (*población en general*). También, producen sobrecostos de emergencia, ya que, se deben encender plantas de generación adicionales para suplir la demanda de energía. De igual manera, la sobre-estimación ($> -5\%$) de la demanda causa remanentes de energía que se traducen en condiciones de operación sub-óptimas, conllevando a ineficiencias y desperdicio de recursos. Por lo tanto, las pérdidas económicas en los dos casos no solo afectan a los generadores de energía sino también al resto del Sistema Interconectado Nacional – SIN (*transmisores, distribuidores, comercializadores, consumidores, etc.*) implicando sobrecostos que asumen el estado, las empresas y la población en general (*tarifas eléctricas más elevadas*).

En segundo lugar, el conjunto de datos “preparados” se dividió en 70% de los datos para entrenamiento de los modelos y 30% restante para prueba de forma secuencial, es decir, en orden cronológico dada la naturaleza del problema. Además, internamente los modelos toman el 20% de los datos de entrenamiento para realizar la validación, por lo cual, finalmente se tiene que el 56% de los datos corresponde a datos de entrenamiento, 14% a validación y 30% para prueba.

Por último, dado que la división cronológica implica que los datos de entrenamiento preceden a los de validación y prueba, esta estrategia asegura que el modelo solo tenga acceso a información pasada para predecir futuros resultados, lo que es una característica clave de las series de tiempo. En este sentido, no es necesario verificar que la distribución de los subconjuntos se conserva respecto al conjunto original, ya que, la separación cronológica respeta las propiedades inherentes de los datos temporales. De hecho, realizar una verificación de distribución no es adecuada en este caso, puesto que, en las series de tiempo el orden temporal es mucho más relevante que la distribución estadística entre los subconjuntos. Por lo tanto, al dividir los datos en un 70% para entrenamiento, 14% para validación y 30% para prueba, la integridad temporal de los datos se preserva y no hay riesgo de que la distribución temporal de los datos se vea alterada. Los subconjuntos reflejan correctamente las secuencias de tiempo y, por ende, se ajustan a los requisitos de los modelos de predicción de series de tiempo.

Nota: Es importante aclarar que se descartaron las métricas de desempeño MSE y R^2 debido a su baja interpretabilidad dada la magnitud de los datos de la variable objetivo y, por no ser útil en modelos complejos donde no existe una relación lineal dando valores bajos a pesar de tener buenas predicciones, respectivamente.

9. CONSTRUCCIÓN Y EVALUACIÓN DEL MODELO

En primer lugar, es necesario mencionar que inicialmente se concebía la realización de un único modelo de predicción de la demanda para todos los mercados de comercialización. Sin embargo, de acuerdo con los resultados obtenidos durante la primera etapa del proyecto se acordó con el cliente realizar un modelo por cada mercado de comercialización.

En segundo lugar, para evaluar el desempeño de los tres (3) modelos propuestos, (*LSTM (Long Short-Term Memory)*, *MLP (Multilayer Perceptron)* y *CNN (Convolutional Neural Network)*) se utilizó el mercado de comercialización de Bogotá. Debido al bajo desempeño obtenido con el modelo LSTM se decidió descartarlo para el entrenamiento y prueba de los 23 modelos restantes.

En tercer lugar, el ajuste de los hiperparámetros se realizó con la librería Keras Tuner para los modelos MLP y CNN porque automatiza la búsqueda de combinaciones óptimas de parámetros, mejorando la precisión de los modelos sin la necesidad de ajustes manuales. Además, permite explorar rápidamente diversas configuraciones y garantiza modelos más robustos y eficientes, lo cual, es crucial para mejorar la predicción de la demanda energética.

A continuación, se detallan las métricas obtenidas para los modelos propuestos.

Ilustración 7. Métricas de desempeño de los modelos propuestos

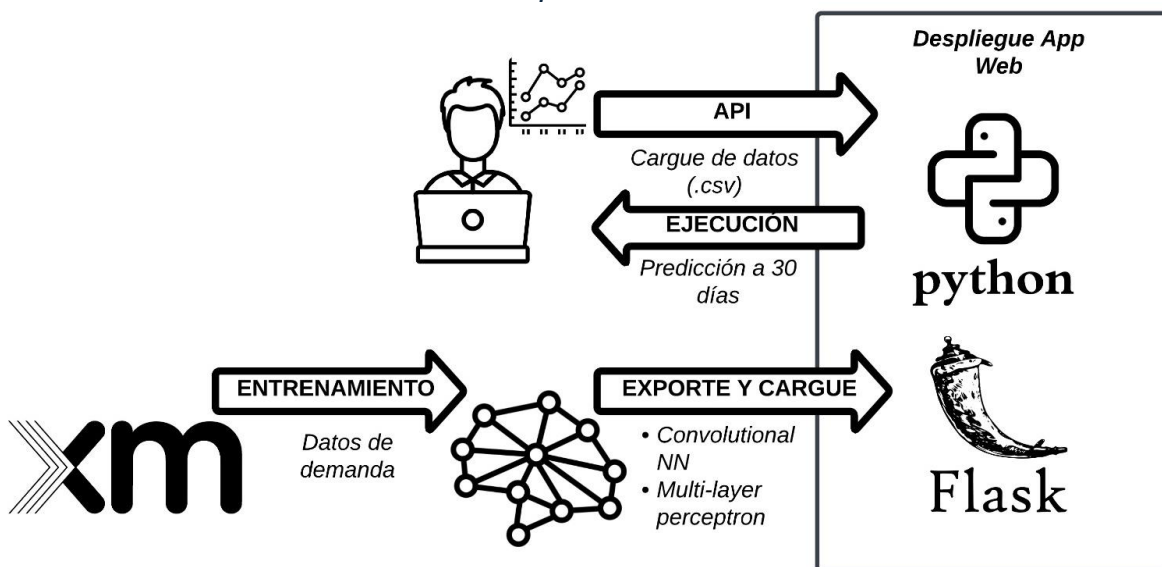
MERCADO DE COMERCIALIZACIÓN BOGOTÁ				
Modelos Series de Tiempo	ENTRENAMIENTO		PRUEBA	
	RMSE	MAE	RMSE	MAE
LSTM	2.205.265	2.108.206	3.618.742	3.553.240
MLP	2.809.426	1.929.105	2.646.202	1.451.526
CNN	2.719.579	1.891.036	2.500.559	1.411.439

Fuente: Elaboración propia.

10. CONSTRUCCIÓN DEL PRODUCTO DE DATOS

Se realizó una aplicación web en donde se pueden consumir predicciones de la demanda energética por mercado de comercialización. Es importante aclarar que, internamente se utiliza el mejor modelo por cada mercado de acuerdo con la estrategia planteada en el capítulo 8. Adicionalmente, se despliega la solución de forma local usando la librería Flask, ya que, es un framework web ligero y flexible para Python, diseñado para desarrollar aplicaciones web de forma sencilla y rápida. La simplicidad en su configuración, junto con un enrutamiento fácil de usar y la integración con el motor de plantillas Jinja2, lo convierten en una excelente opción cuando se busca un control total sobre el diseño de la aplicación sin sacrificar flexibilidad o escalabilidad. (ver Ilustración 8).

Ilustración 8. Arquitectura de la solución



Fuente: Elaboración propia.

Nota: Para conocer a detalle la construcción del producto de datos puede remitirse al siguiente [video](#) (por favor ajustar la calidad de reproducción a 1080p).

11. RETROALIMENTACIÓN POR PARTE DE LA ORGANIZACIÓN

A continuación, se presenta a modo de bitácora un resumen de las diferentes interacciones con el cliente, en donde se detallan los acuerdos más importantes:

Ilustración 9. Bitácora de interacciones y acuerdos con el cliente

RETORALIMENTACIÓN POR PARTE DEL MinEnergía		
Reunión No.	Interacciones	Acuerdos
1	Entendimiento del problema	Se pueden abordar modelos diferentes a una Red Neuronal Recurrente (RNN).
2	Definición de métricas de negocio	Tolerancia de error de pronóstico <5%. <i>Se desconoce el tipo de series de tiempo empleado actualmente y la precisión de estos modelos.</i>
3	Entendimiento de los datos	<u>ETAPA 1:</u> Se utilizará la Demanada SIN por Mercado de Comercialización como variable objetivo.
		<u>ETAPA 2:</u> Se descartan las variables exógenas en la elaboración del modelo, debido a su baja correlación con la variable objetivo. <i>[predicción demanda energética diaria]</i> . Se definen 30 días como el horizonte de tiempo de predicción.
4	Evaluación del modelo	<u>Se acuerda:</u> 1. Realizar un modelo por cada tipo de Mercado de Comercialización para mejorar los pronósticos (24 modelos en total). 2. Entregar una API consumible para XM. 3. Redactar manual de usuario para entendimiento del modelo y posibles ajustes posteriores.
5	Resultados del modelo	Se presentará el desempeño del modelo en función de las métricas del mismo y de la métrica de negocio. <i>Pendiente retroalimentación final por parte de MinEnergía.</i>

Fuente: Elaboración propia.

12. CONCLUSIONES

- ✓ Utilizar un modelo por cada mercado de comercialización mejora notablemente el desempeño de las predicciones en comparación a un único modelo general.
- ✓ Se observó que la demanda energética presenta patrones similares de consumo de manera semanal en la mayoría de los mercados de comercialización.
- ✓ Se presenta alta variabilidad en la demanda energética diaria y baja estacionalidad semanal en los mercados de comercialización cuyas actividades económicas principales corresponden a explotación de petróleo, gas, carbón y minerales (Caribe sol, Caribe mar, Boyacá y Meta).
- ✓ El uso de variables exógenas que presentan variaciones significativas en granularidades mensuales y/o anuales, no son relevantes en modelos predictivos con intervalos de tiempo a corto plazo y que presentan granularidades más bajas.
- ✓ Se propone evaluar el uso de variables exógenas con alta variabilidad y granularidades diarias y/o semanales para mejorar el desempeño de las predicciones de la demanda energética en algunos mercados de comercialización (precio producción energía, precio producción de gas, precio del dólar y precio del petróleo).
- ✓ En un contexto comercial es importante evaluar el modelo desde las métricas de negocio que reflejan el impacto real y la efectividad de este en los objetivos comerciales, y no solo desde las métricas del modelo que miden su precisión y rendimiento técnico, puesto que, se corre el riesgo de que el modelo sea matemáticamente sólido, pero no útil para las decisiones estratégicas del negocio.
- ✓ Recibir retroalimentación constante de los stakeholders en el desarrollo de un modelo de ML es clave para asegurar que el modelo cumpla con los objetivos del negocio, se ajuste a sus necesidades y sea práctico en su implementación, mejorando así su efectividad y adopción.