Pronóstico de demanda de energía eléctrica en un mercado de comercialización en Colombia.

Juan Esteban Rojas Serna.

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor
Walter Mauricio Villa Acevedo, PhD Electrical Engineering
Álvaro Jaramillo Duque, PhD Electrical Engineering

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2024

Cita	(Rojas Serna, 2024)
Referencia **Estilo APA 7 (2020)**	Rojas Serna, J.E. (2024). Pronóstico de demanda de energía eléctrica en un mercado de comercialización en Colombia. Trabajo de grado especialización, Especialización en Analítica y Ciencia de Datos. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte VI.

Centro de Investigación Ambientales y de Ingeniería (CIA).

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Centro de Documentación Ingeniería (CENDOI)

Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.
Decano: Julio Cesar Saldarriaga Molina
Jefe departamento: Diego José Luis Botia Valderrama

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Siglas, acrónimos y abreviaturas

ARIMA	AutoRegressive Integrated Moving Average
CND	Centro Nacional de Despacho
CNO	Consejo Nacional de Operación
CREG	Comisión de Regulación de Energía y Gas
DANE	Departamento Administrativo Nacional de Estadística
IDEAM	Instituto de Hidrología, Meteorología y Estudios Ambientales
IPC	Índice de Precios al Consumidor
IPP	Índice de Precios al Productor
GRU	Gated Recurrent Unit
LSTM	Long Short Term Memory
MC	Mercado de comercialización
NREL	National Renewable Energy Laboratory
RNN	Recurrent Neural Network
SARIMAX	Seasonal Autoregressive Integrated Moving Average with Exogenous regressors
SIN	Sistema Interconectado Nacional
SUI	Sistema Único de Información de la Superintendencia de Servicios Públicos Domiciliarios
TRM	Tasa Representativa del Mercado
UCP	Unidad de Control de Pronóstico de demanda

Resumen

El presente trabajo ha sido desarrollado con la finalidad de aplicar modelos de Machine Learning para ser aplicados al pronóstico horario de la demanda de energía eléctrica para un mercado de comercialización en Colombia, específicamente el correspondiente al departamento de Antioquia, que permitan disminuir el porcentaje de desviación de los pronósticos con respecto a la demanda de energía eléctrica real vs los pronósticos realizados actualmente y/o que dicho nivel de desviación permanezca por debajo del +/-4%. Para esto se ha empleado diferentes fuentes de información abiertas para la obtención de los datos históricos relacionados con la demanda de energía eléctrica, cantidad de usuarios conectados al sistema en cada uno de los municipios del departamento, temperatura ambiente, precipitaciones, entre otras variables climatológicas de cada uno de los municipios, así como variables macroeconómicas a nivel nacional.

Dado que los datos provienen de diferentes fuentes y presentan una granularidad temporal diferente, se hace necesario la aplicación de diferentes metodologías para la preparación y homogenización de los datos para ser anexados en un único dataset que brinde información sobre todas las variables mencionadas para cada hora dentro del período de tiempo seleccionado, el cual, para efectos del presente trabajo, se ha tomado como del 2018-01-01 hasta el 2024-03-31. Al conjunto de datos resultante se le aplicaron diferentes análisis descriptivos para profundizar en el entendimiento y comportamiento de los datos y posteriormente se construyeron los respectivos conjuntos de datos de entrenamiento y validación, los cuales son aplicados a los diferentes modelos empleados como lo son las regresiones lineales, modelo SARIMAX y redes neuronales recurrentes. Al final del informe se presentan los resultados obtenidos concluyendo que en efecto es posible aplicar modelos que brinden niveles de desviación por debajo del valor objetivo.

Repositorio Github: https://github.com/jestebanrojas/analitica-y-ciencia-datos-pronostico-demanda-energia.git

Palabras clave: Demanda de Energía Eléctrica, Pronóstico, Machine Learning, series de tiempo, Redes Neuronales, Regresión Lineal.

Abstract

The present work has been developed with the purpose of applying Machine Learning models to forecast hourly electricity demand for a specific market in Colombia, specifically in the department of Antioquia. The goal is to reduce the percentage deviation of the forecasts compared to the actual electricity demand, either by improving the current forecasts or by keeping the deviation level below +/-4%. To achieve this, various open sources of information have been used to obtain historical data related to electricity demand, the number of users connected to the system in each municipality of the department, ambient temperature, precipitation, and other climatological variables for each municipality. Additionally, macroeconomic variables at the national level have been considered.

Given that the data comes from different sources and exhibits different temporal granularity, it becomes necessary to apply various methodologies for the preparation and standardization of the data to be included in a single dataset that provides information on all the mentioned variables for each hour within the selected time period. For the purposes of this work, the time period has been chosen from January 1, 2018, to September 30, 2023. Different descriptive analyses were applied to the resulting dataset to deepen the understanding of the data and its behavior. Subsequently, the respective training and validation datasets were constructed, which are applied to the different models used such as linear regressions, SARIMAX model, and recurrent neural networks. At the end of the report, the obtained results are presented, concluding that indeed it is possible to apply models that provide deviation levels below the target value.

Github Repository: https://github.com/jestebanrojas/analitica-y-ciencia-datos-pronostico-demanda-energia.git

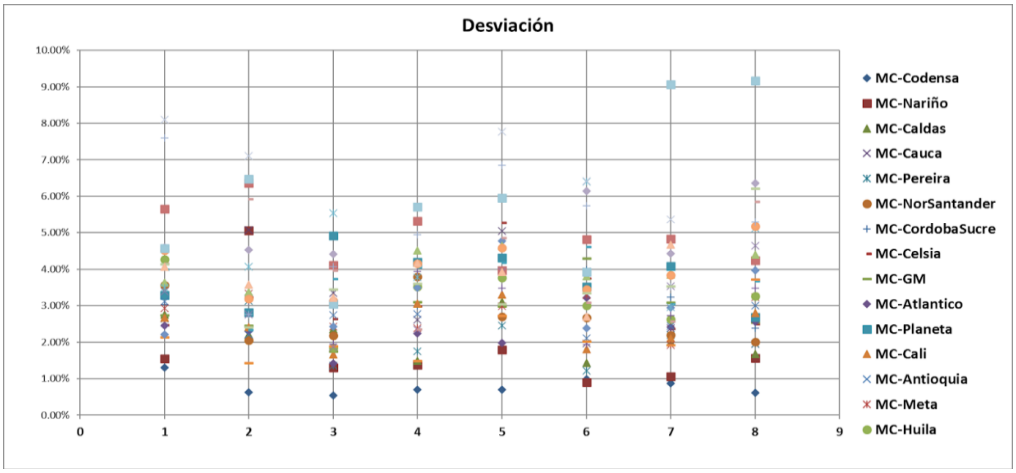
Keywords: Electricity Demand, Forecast, Machine Learning, Time Series, Neural Networks, Linear Regression.

1. Descripción del problema

La Comisión de Regulación de Energía y Gas, CREG, en cumplimiento de sus atribuciones, establece el código de redes en la resolución 025 de 1995 y aquí establece las responsabilidades en el pronóstico horario de la demanda y por su parte el CNO - Consejo Nacional de Operación, en el acuerdo CNO 1303 de 2020 determina la mecánica para la presentación al CND - Centro Nacional de Despacho-, por parte de las empresas responsables, el respectivo pronóstico para su mercado de comercialización. Los operadores de red, como agentes responsables del pronóstico para su respectivo mercado de comercialización, deben presentar al CND el pronóstico de demanda horaria según sus criterios e información disponible, o en su defecto pueden acogerse a los pronósticos de demanda estimados por el CND. Dicho pronóstico deberá estar lo más ajustado posible a la demanda real a fin de garantizar el adecuado despacho de la generación. Para tal fin, los operadores de red han de contar con herramientas técnicas y metodológicas para la estimación de la demanda y estar en la capacidad de suministrar al CND el pronóstico para los siguientes 7 días calendario con un detalle horario por cada día.

De acuerdo con los indicadores publicados por XM en su portal web, se tienen los siguientes porcentajes de desviación entre el pronóstico y la demanda real por cada uno de los mercados de comercialización en el país para los diferentes meses del año 2023.

Figura 1.
Desviación mensual por mercado de comercialización de la demanda real vs la demanda pronosticada



Nota: Elaboración propia a partir de información pública de XM.

El Acuerdo CNO 1303 (2020) considera que valores de desviación que estén por encima del 5%, pueden poner en riesgo la atención de la demanda y en caso de que para un mercado de comercialización se supere este límite por dos días consecutivos, el operador de red deberá realizar el respectivo análisis y presentar al CND las acciones de mejora a aplicar para evitar la futura ocurrencia de las desviaciones y hacer seguimiento a fin de garantizar la efectividad de estas.

Por otro lado, el proyecto de resolución CREG 100 de 2019 “por medio de la cual se proponen modificaciones a las Resoluciones CREG 025 de 1995 y CREG 063 de 2000 y se establecen otras disposiciones”, da señales de las intenciones del regulador en incorporar ajustes a la dinámica de pronóstico de la demanda, en la cual, entre otros ajustes, plantea la asignación a los comercializadores un porcentaje de los costos de las reconciliaciones positivas por desviaciones de la demanda que superen el umbral del +/- 4%, dando cuenta de la necesidad de avanzar en los modelos de predicción de la demanda a fin de alejarse del umbral del 4% de la desviación entre la demanda pronosticada con la demanda real.

Como se puede observar en la Figura 1, es evidente que es frecuente que se presenten desviaciones que superen el umbral del 5% y también un mayor número de ocasiones en las que se supera el 4% estando cerca del umbral definido, dando cuenta de la necesidad de establecer mecanismos que faciliten la estimación de la demanda y disminuyan los niveles de desviación con respecto a la demanda real.

1.1. Problema de negocio.

Se requiere realizar el pronóstico de demandas operativas del Sistema Interconectado Nacional - SIN - por parte de un operador de red para su mercado de comercialización de forma horaria para un período de 7 días con un nivel de desviación que mejore las predicciones actuales y/o que en todo caso no supere el 4% con respecto a la demanda real, por lo que en el presente trabajo se abordará el mercado de comercialización correspondiente a Antioquia.

1.2. Aproximación desde la analítica de datos.

Se aplicarán varias técnicas propias de modelos de regresión lineal, redes neuronales y una variación del método ARIMA, el cual incorpora un componente temporal y variables

exógenas, este es conocido como método SARIMAX. Luego de aplicar estos modelos y realizar iteraciones con diferentes parámetros, se tomará aquel que mejor nivel de precisión presente en el pronóstico. El método ARIMA (AutoRegressive Integrated Moving Average) o sus variantes se perfilan como una posible alternativa a trabajar dado que puede ser empleado para encontrar patrones para una predicción a partir de datos del pasado y no por variables independientes (Ortuño, Ramos, & Senent, 2018). Las redes neuronales también podrán ser exploradas dado que se han empleado con éxito en ejercicios de predicción en series de tiempo con alto grado de precisión, con antecedentes en su aplicación en casos tales como generación de electricidad y consumo de gas natural (Escobar, Luis; Valdés, Julio; Zapata, Santiago, 2010).

Ahora bien, teniendo en cuenta que se cuenta con información específica de temperatura ambiente, precipitaciones, usuarios y variables macroeconómicas, durante el desarrollo se evaluará también la posibilidad de emplear regresiones lineales para cada hora de cada tipo de día, lo que podría resultar en un modelo específico para cada combinación posible entre cada hora (desde las 0:00h hasta las 24:00h) para cada tipo de día, eliminando, bajo este enfoque, la concepción de un modelo dependiente meramente del tiempo sino que incluya posibles variables independientes que definan la demanda de energía. Esto a fin de evaluar si con este método se obtienen mejores precisiones en el pronóstico y se evidencia que bajo este enfoque se pueda inferir que no exista una dependencia exclusivamente del tiempo sino una mayor dependencia de otras variables predictoras.

Para la ejecución de los modelos mencionado se emplearán herramientas tales como Python y librerías disponibles tales como Scikit-learn, stats-models, statsForecast, entre otras.

1.3. Origen de los datos.

Teniendo en cuenta lo mencionado en el numeral 1.2, se requerirán datos históricos desde el 1 de enero de 2018 tales como la demanda de energía eléctrica histórica para el mercado de comercialización específico, así como información histórica de temperatura ambiente, precipitaciones, usuarios y variables macroeconómicas.

Demanda histórica por mercado de comercialización con resolución horaria desde 2018. Información disponible directamente de la fuente oficial (XM) con la información de la demanda real del Sistema Interconectado Nacional.

Históricos de temperatura ambiente y precipitaciones con resolución horaria por municipio perteneciente al mercado de comercializador del Operador de Red específico. Información disponible en bases de datos de la NASA, el National Renewable Energy Laboratory (NREL), el IDEAM, entre otros.

Pronóstico de temperatura y precipitación para un período de 7 días de la semana siguiente con resolución horaria. Información tomada de las publicaciones oficiales del IDEAM.

Indicadores macroeconómicos históricos (mensuales / anuales) tales como TRM, IPP e IPC en Colombia desde el 2018. Información disponible en fuentes tales como el DANE y Banco de la República.

Usuarios totales conectados al Operador de Red por Municipio a final de cada mes. Información disponible en el SUI - Sistema Único de Información de Servicios Públicos Domiciliarios.

Al tratarse de data oficial de entidades como XM y la superintendencia de servicios públicos domiciliarios (Demandas operativas de energía y cantidad de usuarios) que corresponde a reportes de carácter regulatorio y periódicos, se cuenta con información que cumple con criterios de completitud y sometida a diferentes procesos de calidad de datos por parte de las entidades encargadas de su reporte; con respecto a la información climática y macroeconómica se emplearán también fuentes oficiales o reconocidas en su materia, quienes publican de forma recurrente dicha información.

La información a emplear corresponde a información de dominio público dispuesta en los portales web específicos de cada entidad mencionada, la cual puede ser consultada y exportada en archivos tipo CSV y/o Excel para diferentes rangos de tiempo según la entidad. Para el caso de las demandas operativas históricas se puede acceder a la información por cada mercado de comercialización de un operador de red específico, con archivos mensuales desde el

año 2018. Los pronósticos de temperatura pueden hallarse en el portal web del IDEAM para cada municipio para la siguiente semana; por su parte los valores de temperatura histórica por horas se pueden obtener de repositorios como el de NREL o la NASA, para lo cual se tomará como referencia la temperatura en las coordenadas del casco urbano de cada municipio. La información de los usuarios conectados en cada municipio se encuentra disponible en el portal del SUI, Aquí se puede descargar la información de cada mes con la cantidad de usuarios según su tipo discriminados por municipio para un departamento seleccionado.

Los indicadores macroeconómicos serán obtenidos a través de información disponible en el DANE y de ser necesario en el Banco de la República.

1.4. Métricas de desempeño.

Se describen a continuación las métricas que serán empleadas para la estimación del desempeño de los modelos usados. Las métricas a emplear serán el R^2, MSE, RMSE, MAE y MAPE. Dado que el objetivo del presente trabajo pretende identificar un modelo que permita una desviación en el pronóstico inferior al +/- 4%, será de especial interés monitorear el resultado que arroje la métrica MAPE.

Coeficiente de determinación (R^2): Medida estadística que indica la proporción de la variabilidad de la variable dependiente que es explicada por el modelo de regresión. Se calcula como la proporción de la suma de los cuadrados de la regresión (SSE) respecto a la suma total de los cuadrados (SST) (Montgomery, Peck, & Vining, 2012).

$$R^2 = 1 - \frac{SSE}{SST}$$

Donde, SSE es la suma de los cuadrados del error y SST es la suma total de los cuadrados.

$$SSE = \sum_{t=1}^{T} (y_t - \hat{y_t})^2$$

$$SST = \sum_{t=1}^{T} \left(y_t - \underline{y}\right)^2$$

Siendo T el número de observaciones, y_t los valores observados, $\hat{y}_t$ los valores predichos por el modelo y $\underline{y}$ el valor medio de todos los valores observados.

Si bien el R^2 no es comúnmente empleado en el contexto de redes neuronales, dado que el presente trabajo abordará un enfoque tanto con regresiones lineales como en Redes Neuronales, se empleará en conjugación con otras métricas comúnmente empleadas.

Ahora bien, para medir la exactitud de modelos que trabajan pronósticos, Hyndman (2018), resalta como métricas adecuadas el MSE, RMSE, MAE y MAPE.

Error cuadrático medio (MSE): Es una medida de que tan bien las predicciones coinciden con las observaciones, el MSE es la suma de los cuadrados de las diferencias entre los valores predichos por el modelo y los valores observados. Cuanto menor sea el MSE, mejor será el ajuste del modelo.

$$MSE = \frac{1}{T}\sum_{t=1}^{T} (y_t - \hat{y_t})^2$$

Raíz del error cuadrático medio (RMSE): Corresponde a la raíz cuadrada del MSE. Este suministra una medida en las mismas unidades de la variable que se está prediciendo, dando cuenta del error promedio de las predicciones del modelo. Un valor más bajo de RMSE indica un mejor ajuste del modelo.

$$RMSE = \sqrt{\frac{1}{T}\sum_{t=1}^{T} (y_t - \hat{y_t})^2} = \sqrt{MSE}$$

Error absoluto medio (MAE): Es la media aritmética de los valores absolutos de los errores entre las predicciones del modelo y los valores observados. Proporciona una medida de la magnitud promedio del error de predicción. Al igual que el MSE y el RMSE, un valor más bajo de MAE indica un mejor ajuste del modelo.

$$MAE = \frac{1}{T}\sum_{t=1}^{T} |(y_t - \hat{y_t})|$$

Error porcentual absoluto medio (MAPE): Es una medida del error porcentual promedio entre las predicciones del modelo y los valores observados. Se calcula como el promedio de los errores porcentuales absolutos de cada predicción. Es útil para evaluar la precisión de las predicciones en relación con la escala de las variables. Un valor más bajo de MAPE indica una mejor precisión del modelo.

$$MAPE = \frac{100}{T}\sum_{t=1}^{T} \frac{|(y_t - \hat{y}_t)|}{y_t}$$

2. Objetivos

2.1. Objetivo general.

Desarrollar un modelo para el pronóstico de la demanda de energía eléctrica en el mercado de comercialización de Antioquia - Colombia, cumpliendo las disposiciones emitidas por la CREG y el CNO para el pronóstico de la demanda, usando modelos de Machine Learning que minimicen el error existente entre el pronóstico y la demanda real.

2.2. Objetivos específicos.

- Construir una base de datos con la información histórica de las variables relevantes que influyen en el comportamiento de la demanda de energía eléctrica que provienen de diferentes fuentes disponibles.
- Realizar el procesamiento y análisis de las variables de la base de datos para la identificación de las variables con mayor influencia sobre el comportamiento de la demanda de energía eléctrica.
- Desarrollar y probar varios modelos de machine learning para la predicción de la demanda de energía eléctrica para los siguientes siete días a partir de la información procesada con las variables predictoras elegidas.
- Validar los modelos propuestos para el pronóstico de la demanda de energía eléctrica usando métricas de desempeño y los datos de prueba

3. Datos

3.1. Datos originales

Como se ha mencionado en el numeral 1, la información a ser considerada corresponde a: demanda de energía eléctrica histórica para el mercado de comercialización específico, información histórica de temperatura ambiente, precipitaciones, usuarios y variables macroeconómicas; a continuación, se hace una descripción de los datos:

3.1.1. Histórico de demanda de Energía Eléctrica.

La información correspondiente a la demanda de energía eléctrica histórica para el mercado de comercialización - MC - seleccionado, el cual para el caso del presente trabajo se define que será Antioquia, se recurre al portal de XM en el cual se realiza el respectivo cargue de manera mensual para todos los MC.

En este caso, la información se encuentra en ficheros que responden a la siguiente estructura: año/mes/informe_mc.xls o año/mes/informe_mc.xlsx donde el archivo año/mes/informe_mc.xls o año/mes/informe_mc.xlsx corresponde al informe de los indicadores de pronósticos oficiales de demanda para cada mercado de comercialización, por lo que se procede con la descarga de los datos para el MC seleccionado (Antioquia). La extensión y nombre del archivo puede variar según la época y cambios en el uso de las herramientas de excel empleadas por el Centro nacional de Despacho - CND - para el respectivo reporte.

Una vez se han descargado los datos históricos entre el 2018-01-01 y el 30-09-2023, se deberá proceder con la unión de cada uno de los archivos para la conformación de un dataset único con todo el histórico de la demanda.

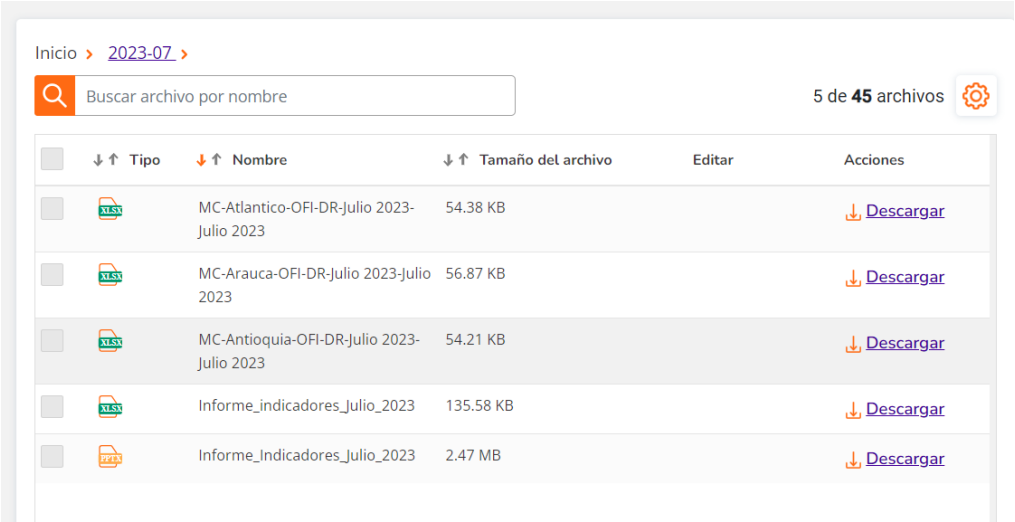
La estructura de los datos es tal y como se presentan en la Tabla 1.

Tabla 1. Estructura del dataset Histórico de demanda de Energía Eléctrica (demanda.csv)

Nombre	Descripción	Tipo
UCP	Unidad de Control de Pronóstico. Para efectos del presente trabajo corresponde a Antioquia.	Categórica
Variable	Tipo de variable del reporte, para efectos del presente trabajo siempre corresponderá a “Demanda Real”.	Categórica
FECHA	Fecha en formato Año-Mes-Día al cual corresponde la demanda real.	Datetime
TIPO_DIA	Tipo de día al cual corresponde la demanda real, indicando si corresponde a un día entre semana, fin de semana, día festivo o alguna fecha especial.	Categórica
P1…P24	Energía demandada para la UCP en cada una de las 24 horas del día indicado en el campo FECHA.	Numérica
Total	Energía total demandada para la fecha indicada en el campo FECHA.	Numérica
PO19, PO20, PO21	Demanda de potencia máxima para las horas 19, 20 y 21 de la fecha indicada en el campo FECHA.	Numérica

En la Figura 2 se ilustra la disposición de los archivos en el repositorio web dispuesto por XM.

Figura 2.
Estructura de archivos con información histórica de la demanda en repositorio de XM.



En la Tabla 2 se muestra la descripción del dataset resultante.

Tabla 2. Descripción del archivo demanda.csv

Descripción	Características
Nombre del dataset	demanda.csv
Tamaño	633 kb

Cantidad Registros	2099

3.1.2. **Histórico de variables climatológicas.**

Para este caso, la información climatológica es la correspondiente a la información histórica con desagregación horaria de temperatura, precipitaciones, humedad relativa, radiación solar, claridad del cielo, entre otros, de los municipios del departamento del mercado de comercialización elegido. La información se toma del portal de la NASA dispuesto para tal fin y se emplea la API disponible en: // https://power.larc.nasa.gov/data-access-viewer/. Para descargar esta información se hace necesario contar con las coordenadas de los municipios respectivos, por lo que se toma la información disponible en el DANE y se dispone previamente para ser empleada por un script de Python que se encargará de correr la API para cada uno de los municipios del departamento.

La estructura de los datos entregada por la API es como se presenta en la Tabla 3.

Tabla 3. Estructura del dataset con el histórico de variables climatológicas (clima.csv)

Nombre	Descripción	Tipo
YEAR	Año	Numérica
MO	Mes	Numérica
DY	Día	Numérica
HR	Hora	Numérica
ALLSKY_SFC_SW_DWN	Irradiancia de onda corta que llega a la superficie terrestre desde todas las direcciones del cielo, independientemente de las condiciones meteorológicas o la presencia de nubes medido en Wh/m^2.	Numérica
ALLSKY_KT	Indicador de la claridad del cielo en relación con la radiación solar incidente.	Numérica
T2M	Temperatura a 2 m en °C	Numérica
RH2M	Humedad relativa a 2 metros en %	Numérica
PRECTOTCORR	Precipitación Corregida en mm/hora	Numérica
CLRSKY_SFC_SW_DWN	Cantidad de radiación solar de onda corta que llega a la superficie terrestre bajo condiciones de cielo despejado, es decir, sin nubes ni obstrucciones atmosféricas medido en Wh/m^2.	Numérica
T2MWET	Temperatura de bulbo húmedo a 2 m en °C	Numérica

Es de anotar que adicional a la información contenida en la estructura anterior, la información para cada coordenada viene acompañada de un "Header", el cual debe ser eliminado para ser anexado a un dataset con la información para todas las coordenadas, esto último también se realiza en el script de Python mencionado previo a almacenar el dataset en un archivo csv.

La Tabla 4 muestra la descripción del dataset resultante para variables climatológicas.

Tabla 4. Descripción del archivo clima.csv

Descripción	Características
Nombre del dataset	clima.csv
Tamaño	376.2 Mb
Cantidad Registros	2099

3.1.3. **Histórico de indicadores macroeconómicos.**

3.1.3.1. Histórico TRM

El dataset TRM se toma de la página web del Banco de República de Colombia, en el siguiente enlace: https://www.banrep.gov.co/es/estadisticas/trm. Allí nos dirigimos al botón Descargar y se descarga toda la información histórica del precio de cierre de la TRM en Colombia desde el año 1991 hasta el día anterior que se descargue el archivo. Este archivo de la TRM consta de 2 columnas como se presenta en la Tabla 5.

Tabla 5. *Descripción del dataset TRM descargado del Banco de la República*

Nombre	Descripción	Tipo
Fecha	Fecha al cierre de ese día de la TRM	Datetime
TRM	Valor de la TRM al cierre del mercado de divisas	Numérica

Al seguir los pasos anteriores, se descarga un archivo con las siguientes características:

Tabla 6. Características del dataset TRM descargado del Banco de la República.

Descripción	Características
Nombre del dataset	TRM.xslx
Tamaño	175 kb
Cantidad Registros	11687

En la Figura 3 se puede evidenciar el lugar en la página web donde se descarga la información para la TRM, adicionalmente se encuentra un gráfico del comportamiento de la TRM en Colombia:

Figura 3.
Lugar en la página del Banco de la República en donde se puede descargar la información histórica de la TRM. Adicional el gráfico del comportamiento de la TRM ofrecida por la página.

3.1.3.2. Histórico IPP e IPC

El dataset IPC e IPC corresponde a la información histórica con resolución mensual desde el año 2018 a 2024 del comportamiento de estas dos variables macroeconómicas. Estos se pueden descargar de la página del DANE y la estructura del dataset en la Tabla 7.

Tabla 7. Descripción del Dataset descargado del DANE sobre el IPC e IPP

Nombre	Descripción	Tipo
Fecha	Mes y año de la información del IPP e IPC	Datetime
IPP	Unidades del IPP	Numérica

IPC	Unidades del IPC	Numérica

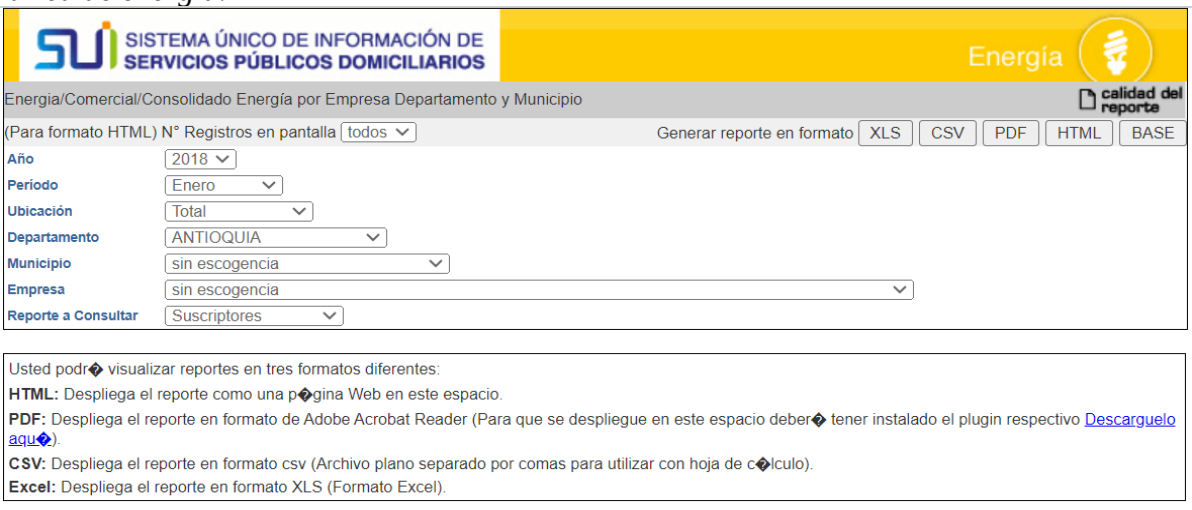
Complementando la información de la Tabla 7, el dataset Evolución IPC E IPP base 2008 - Resolución_ Mensual.xlsx, tiene un tamaño de 8 kb y 71 filas.

3.1.4. **Histórico de usuarios conectados al sistema.**

El dataset usuarios.csv, corresponde a la información histórica de la cantidad de usuarios conectados a la red de energía a nivel nacional desde enero del 2018 a marzo del 2024.

Para conocer esta información se descargó de la página web del SUI (Sistema único de información de servicios públicos domiciliarios, ver Figura 4). En dicho aplicativo solo se permite descargar 1 mes a la vez por cada año de todos los municipios de Antioquia, por ello se descargaron 69 archivos .csv, los cuales posteriormente se unieron en un solo archivo usuarios.csv de forma ordenada usando el lenguaje de programación Python. El enlace para descargar los archivos es el siguiente: http://reportes.sui.gov.co/fabricaReportes/frameSet.jsp?idreporte=ele_com_096.

Figura 4.
Estructura de la página del SUI donde se descarga la información de los usuarios conectados a la red de energía.



Los parámetros tomados fueron los siguientes:

- Año: 2018 hasta 2024
- Periodo: Enero a Diciembre, para el año 2024 hasta marzo
- Ubicación: Total (incluye Urbano, Rural, Centro Poblado)

- Departamento: ANTIOQUIA
- Municipio: Sin escogencia (Todos los municipios del departamento)
- Empresa: Sin escogencia (Todas las empresas)
- Reporte a consultar: Suscriptores (usuarios)

Luego de descargar y unir los archivos de cada mes, el resultado es el dataset usuarios.csv con un peso de 609 kb y una cantidad de registros de 8488. Cuenta con la estructura presentada en la Tabla 8.

Tabla 8. Dataset resultante de usuarios después de unir todos los archivos del SUI

Nombre	Descripción	Tipo
MUNICIPIO	Nombre del municipio de Antioquia.	Categórica
AÑO	Año correspondiente al reporte de usuarios por municipio.	Numérica
ESTRATO	Cantidad de usuarios del estrato 1 al estrato 6 en Antioquia por municipio en determinado día.	Numérica
TOTAL RESIDENCIAL	Cantidad de usuarios totales (Suma del estrato 1 al estrato 6) en Antioquia por municipio en determinado día.	Numérica
INDUSTRIAL	Cantidad de usuarios industriales en Antioquia por municipio en determinado día.	Numérica
COMERCIAL	Cantidad de usuarios comerciales en Antioquia por municipio en determinado día.	Numérica
OFICIAL	Cantidad de usuarios oficiales en Antioquia por municipio en determinado día.	Numérica
OTROS	Cantidad de usuarios otros en Antioquia por municipio en determinado día.	Numérica
TOTAL NO RESIDENCIAL	Cantidad de usuarios totales no residenciales (Suma de industrial, comercial, oficial, otros) en Antioquia por municipio en determinado día	Numérica

3.2. Datsets

Dadas las características propias de cada fuente de datos, se deberá realizar un proceso de preparación de estos, para garantizar que, para el rango de tiempo histórico definido, se cuente con información de las diferentes variables, esto es que para cada hora de cada día se cuente con:

- Demanda de energía real.
- Temperatura promedio de cada municipio, asumida como la temperatura percibida en las coordenadas del casco urbano del municipio.

- Cantidad de clientes conectados a la red en cada municipio, para lo cual se asumirá como un valor constante para cada día y hora de cada mes según los registros reportados en el SUI.
- Precipitaciones registradas en promedio en cada municipio.

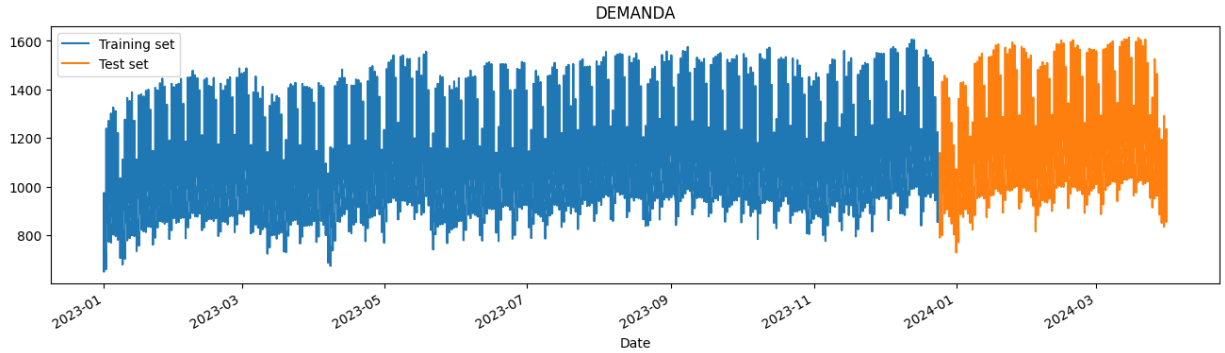
Teniendo en cuenta lo anterior y una vez se ha consolidado cada uno de los datasets resultantes de los numerales 5.1.1 a 5.1.4, se procede a realizar las transformaciones necesarias para la construcción de un único dataset, de modo tal que todas las variables queden a nivel horario y que de este modo puedan ser empleadas como variables predictoras en la aplicación de los diferentes modelos de Machine Learning que así lo requieran. Dichas transformaciones se describen a continuación:

- Consolidar la información de usuarios, clima y datos macroeconómicos, en un dataset que contendrá información para el departamento de Antioquia en cada día/hora del periodo 01/01/2018 al 31/10/2023. Para esto se carga el dataset con las variables climatológicas y se une con el dataset de usuarios, teniendo como llaves las columnas municipio, año y mes. En este caso para cada día/hora en cada municipio se toma como cantidad de usuarios la correspondiente a la cantidad de usuarios del mes para dicho municipio reportado, dado que esta información no se tiene con la resolución día/hora.
- Se procede a realizar una ponderación de cada una de las variables climatológicas para cada municipio, con la cantidad de usuarios para cada período de tiempo y obtener posteriormente un valor promedio ponderado para Antioquia para cada variable climatológica.
- Posteriormente se une el dataset anterior con los dataset que contienen la información histórica de la TRM así como del IPP e IPC.
- Por otra parte, se debe adecuar la información histórica de la demanda, con la finalidad de transformar las columnas con la demanda de cada hora en registros individuales para luego unir el dataset con el resultado del paso anterior, obteniendo así el dataset que servirá para la aplicación de los diferentes modelos.

3.2.1. **Conjunto de datos de entrenamiento y de validación.**

Con miras a establecer un escenario de entrenamiento y pruebas similar para todos los modelos, se establece como conjunto de datos de entrenamiento, la información disponible desde el primero de enero de 2021 hasta el 31 de diciembre de 2023 y como conjunto de datos de pruebas a partir del 1 de enero de 2024 hasta el 31 de marzo de 2024 como se puede observar en la Figura 5.

Figura 5.
Conjunto de datos de entrenamiento y prueba

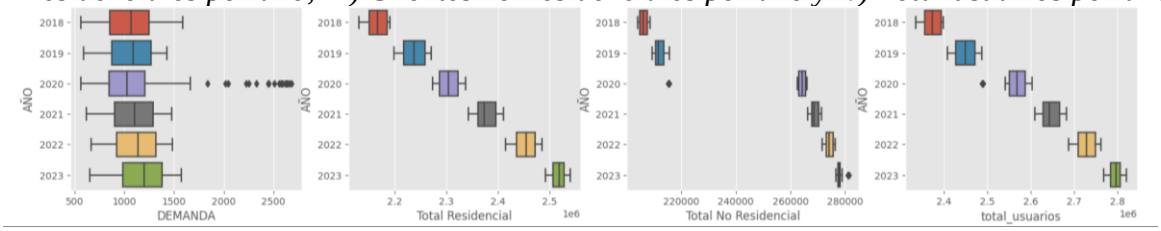


Nota: para efectos visuales se ilustra el conjunto de datos de entrenamiento desde el 01-01-2023, sin embargo, el conjunto de datos empleado inicia desde el 01-01-2021.Elaboración propia.

3.3. Analítica descriptiva

La Figura 6 muestra una tendencia general por cada año en el crecimiento tanto en la demanda como en los usuarios residenciales, no residenciales y totales. Se aprecia que la demanda en el año 2020, derivado de la pandemia por COVID19 tuvo una demanda horaria inferior a 2019, pero del 2021 en adelante se observa nuevamente una tendencia al alza.

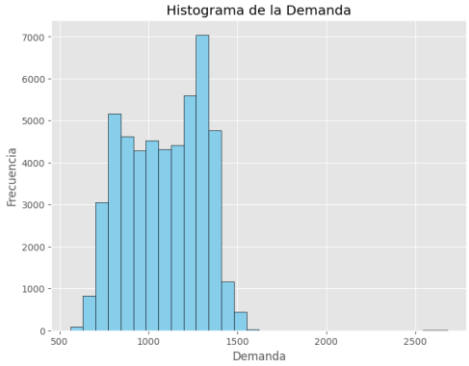
Figura 6.
Diagrama de cajas y bigotes: i) Demanda de energía horaria por año, ii) Clientes Residenciales por año, iii) Clientes no Residenciales por año y iv) Total usuarios por año.



Elaboración propia.

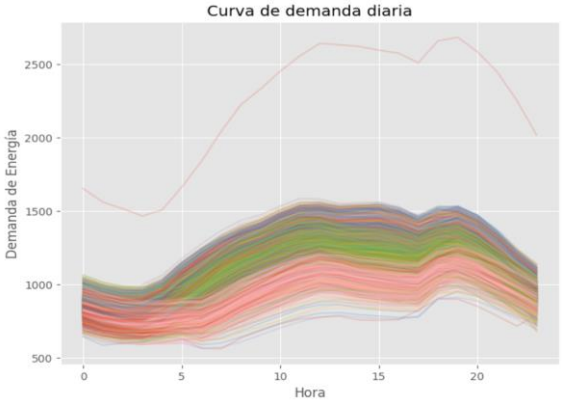
Ahora bien, tanto en la Figura 6 como en la Figura 7 se puede apreciar que existen algunos valores atípicos para las demandas horarias, los cuales se encuentran en magnitudes superiores a los 2.000 MWh, aspecto que se puede apreciar con mejor detalle en la Figura 8, en la cual se evidencia que los valores atípicos corresponden a un día en específico, el cual se encuentra significativamente por encima de los valores típicos para cada una de las horas del día.

Figura 7.
Histograma de frecuencia demanda horaria de energía.



Elaboración propia.

Figura 8.
Curva horaria de la demanda de energía real.

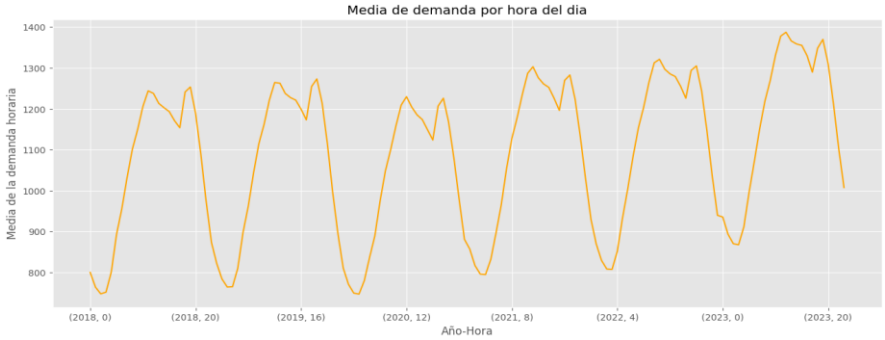


Elaboración propia.

De la Figura 8, se puede apreciar que la demanda horaria presenta un patrón característico para cada hora del día, según su tipo, teniendo que, para las franjas comprendidas entre las 11:00 y las 12:00, así como a las 19:00, se presentan típicamente los momentos de máxima demanda del día. Como se puede apreciar en la Figura 9, la demanda va creciendo en el tiempo conservando dicho patrón diario, salvo en 2020 en donde la demanda decreció con respecto al año inmediatamente anterior, pero en este caso también se preserva el patrón.

Figura 9.

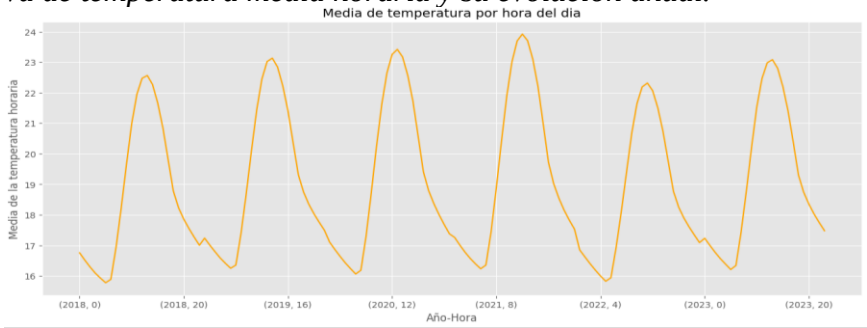
Curva de demanda media horaria y su evolución anual.



Elaboración propia.

Figura 10.

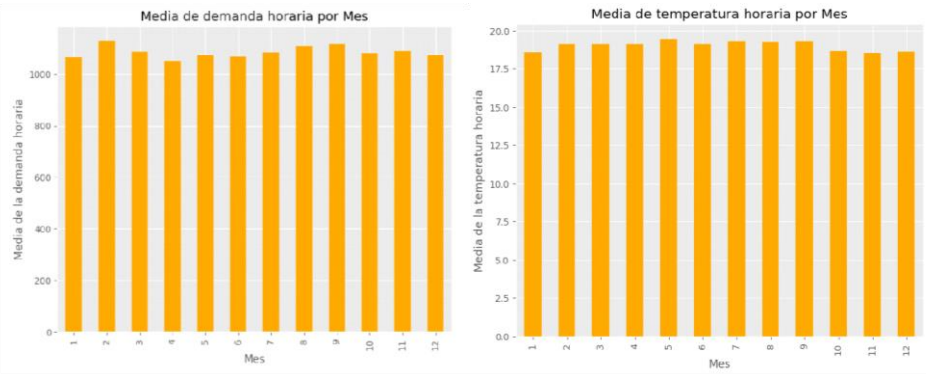
Curva de temperatura media horaria y su evolución anual.



Elaboración propia.

Con miras a identificar si existe una relación clara entre la temperatura ambiente, ver Figura 10, y la demanda de energía, se procede en primera instancia a evaluar si existe algún patrón específico en el comportamiento mensual. La Figura 11, muestra la demanda de energía horaria promedio durante cada mes, así como también la temperatura promedio mensual, al realizar el análisis comparativo entre ambos se identifica que no necesariamente en los meses donde se presentan mayores temperaturas, la demanda horaria promedio sea superior, esto solo parece cumplirse para el mes de septiembre, pero no así para el mes de mayo, el cual presenta las mayores temperaturas, pero la demanda no se encuentra entre las máximas del mes.

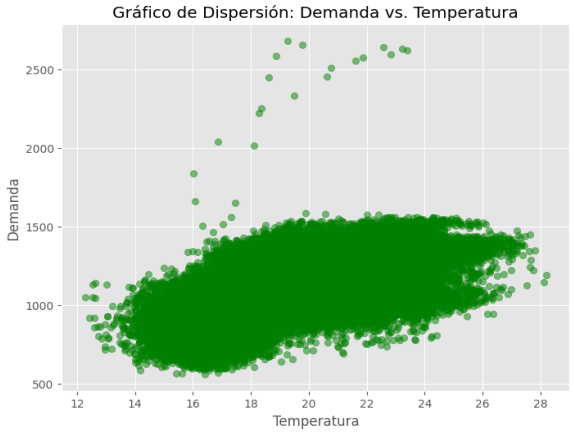
Figura 11.
i) Media de la demanda de energía horaria por mes del año, ii) Temperatura media por mes del año.



Elaboración propia.

Considerando el análisis anterior, se procede a realizar un análisis adicional, aplicando en este caso un diagrama de dispersión entre la temperatura horaria y demanda de energía horaria, cuyo resultado se aprecia en la Figura 12. En este caso se puede apreciar una ligera tendencia a que en la medida que aumenta la temperatura, la demanda de energía es más elevada, lo que da cuenta de que el análisis a aplicar deberá contemplar aspectos como la hora del día a la que se presenta cierta temperatura, así como también el tipo de día, ya que, al retomar la curva diaria de la Figura 8, a las 19:00 se presenta uno de los picos de demanda, a pesar de que la temperatura promedio para esta hora no corresponde a una de las mayores temperaturas del día, sino que dicha hora coincide con aspectos relacionados con el desarrollo de las actividades cotidianas de los hogares en el país.

Figura 12.
Diagrama de dispersión Demanda horaria Vs Temperatura.

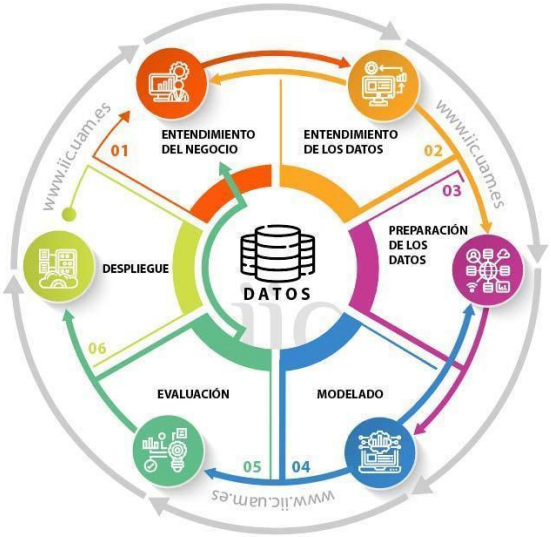


Elaboración propia.

4. Proceso de Analítica.

El presente trabajo se realiza con un enfoque basado en la metodología CRISP-DM, la cual consta de 6 fases, las cuales se ilustran en la Figura 13.

Figura 13.
Fases del modelo de referencia CRISP-DM.



Nota. Gráfico tomado La metodología CRISP-DM en ciencia de datos (2021), por Haya, P. Instituto de Ingeniería del Conocimiento, disponible en https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos. ("Esquema del ciclo CRISP-DM estándar". Figura 1)

Tomando como referencia las definiciones de cada una de las fases de la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) dictadas por Chapman et al (2000), se hace un barrido por cada una de ellas, dando cuenta cómo se aborda cada una de ellas en el presente trabajo.

4.1. Entendimiento del Negocio.

Esta fase se encarga de entender la necesidad de negocio a ser resuelta con el proyecto, el capítulo 1 de este trabajo aborda la problemática a resolver desde la perspectiva de negocio; se refleja cómo es de interés de los agentes involucrados en la cadena de valor de la energía contar con el pronóstico de las demandas operativas para los diferentes mercados de comercialización del país. Así mismo se define en el capítulo 2 los objetivos a ser alcanzados en el presente trabajo, a fin de dar respuesta a los requerimientos de negocio.

4.2. Entendimiento de los Datos.

En esta fase se realiza el proceso de recopilación y exploración de los datos con miras a familiarizarse con los mismos, realizar su respectiva caracterización y de esta manera familiarizarse con la estructura de los datos, así como identificar posibles limitaciones que pueda presentarse ante la incompletitud o características específicas. El numeral 1.3 aborda de manera preliminar los datos que serán considerados para el logro del objetivo planteados, y posteriormente, en los primeros numerales del capítulo 3 se abordará con mayor detalle la estructura de los datos y la fuente de estos.

4.3. Preparación/Prepocesamiento de los Datos.

Esta fase es la encargada de preparar los datos para su posterior análisis, para lo cual se recurre a mecanismos de limpieza, integración, transformación y selección de variables que sean relevantes para afrontar el problema de negocio que se busca resolver. Con miras a contar con un conjunto de datos limpio y adecuado para la posterior aplicación de los modelos predictivos, se explica a partir del numeral 3.2 los procesos llevados a cabo para la transformación e integración de los datos obtenidos de las diferentes fuentes, datos que presentan diferentes niveles de desagregación temporal, por lo que se requiere llegar a un único dataset capaz de reflejar cada una de las variables de interés con una discriminación horaria. Se finaliza el capítulo 3 con un breve ejercicio de análisis descriptivo de los datos finalmente recopilados. Adicionalmente, para la aplicación de los diferentes modelos a abordar en el presente trabajo – siguiente etapa según la metodología CRISP-DM – se requiere la aplicación de transformaciones adicionales con la finalidad de suministrar a los modelos la información adecuada para realizar el proceso de pronóstico de la demanda, en el capítulo 5 se abordarán con un poco de mayor detalle tales transformaciones adicionales.

4.4. Modelado.

La fase de modelado contempla la aplicación de los diferentes modelos con la respectiva calibración de los diferentes parámetros para lograr los valores óptimos. El presente trabajo se enfocará en realizar el modelado empleando técnicas tales como regresiones lineales,

SARIMAX, así como diferentes tipos de redes neuronales. Su aplicación se verá en detalle en el capítulo 5.

A continuación, se presentan los conceptos básicos relacionados con los principales métodos usados para realizar tareas de pronóstico de series de tiempo, para este trabajo corresponde con el pronóstico de la demanda de energía.

4.4.1. **Regresión lineal.**

Corresponde a uno de los modelos más simples para realizar una predicción al calcular simplemente la suma ponderada de las características de entrada, más una constante llamada término de intercepción. Dicho cálculo se rige bajo la siguiente ecuación.

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

Donde:

- $\hat{y}$ es el valor a predecir.
- n el número de características.
- x_i es el valor de la característica i-ésima.
- θ_i es el peso de cada una de las características.
- θ_0 corresponde al término de intercepción.

Ahora bien, para el proceso de entrenamiento de un modelo de regresión lineal, se debe proceder al ajuste de los parámetros, de modo tal que el modelo se ajuste de la mejor forma al conjunto de entrenamiento. Para esto se emplea una medida de ajuste, el cual típicamente se usa el RMSE – La Raíz Cuadrada del Error Cuadrático Medio –, por lo que el entrenamiento velará por identificar los parámetros θ que minimicen el RMSE, aunque en la práctica es más simple el uso del MSE en lugar del RMSE lo que redundará un mismo resultado (Géron, 2019).

4.4.2. **Seasonal Autoregressive Integrated Moving Average with Exogenous regressors (SARIMAX).**

Un modelo SARIMAX corresponde por sus siglas a Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors, lo que traduce Media Móvil Integrada Autorregresiva Estacional con Regresores Exógenos. Este corresponde a un modelo estadístico

empleado para la predicción de series de tiempo que presentan un comportamiento estacional y que dependan de variables externas y se desprende de los modelos ARIMA, por lo que hereda sus principales componentes, a saber (Shumway & Stoffer, 2011):

- Componente AutoRegresivo (AR): Modela la relación entre una observación y un número de observaciones anteriores (autocorrelación).
- Componente Media Móvil (MA): Modela la relación entre una observación y un término de error residual de una regresión que involucra observaciones anteriores.
- Componente Diferenciación Integrada (I): Se aplica para hacer que la serie temporal sea estacionaria, es decir, eliminar tendencias y patrones de estacionalidad.

Por su parte, el componente estacional (S) permite modelar patrones estacionales en la serie temporal y el componente de regresores exógenos (X) permite que además de las características de la propia serie de tiempo, se incluyan variables externas que pueden influir en en la variable a predecir.

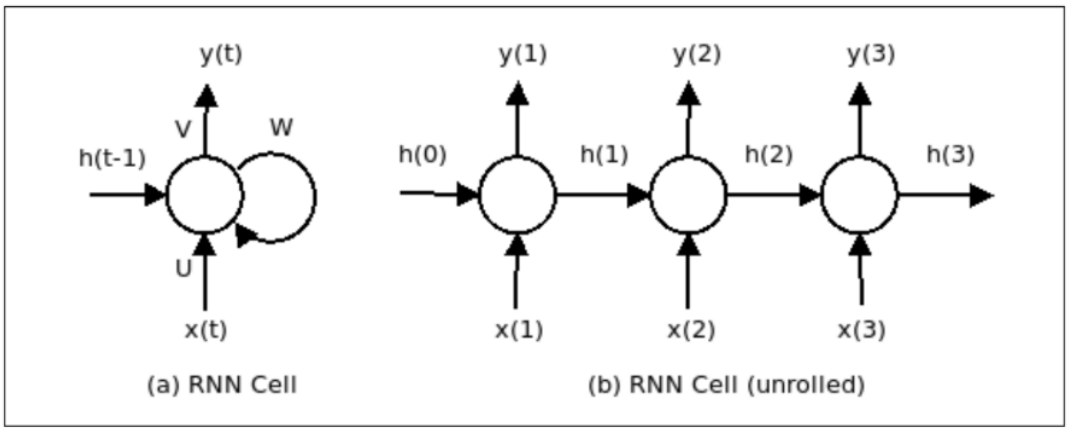
4.4.3. **Redes Neuronales Artificiales – ANN.**

Gerón (2019), describe una red neuronal artificial – ANN por sus siglas en inglés – como un modelo de Machine Learning inspirado en las redes neuronales biológicas presentes en nuestros cerebros. En este sentido se basan en el concepto de Neuronas, su unidad básica, y dichas neuronas se interconectan a través de capas de entrada, capas intermedias o capas ocultas y capas de salida. La conexión entre las capas viene dada por pesos específicos según el nivel de importancia relativa de la entrada de una neurona en la salida de otra. Dentro de los tipos de redes neuronales, según su arquitectura, encontramos las redes neuronales recurrentes – RNN –, las cuales tienen conexiones que forman bucles mediante los cuales agregan retroalimentación y memoria a las redes con el tiempo de modo tal que le permite a la red aprender y generalizar a través de secuencias de entradas en lugar responder a patrones individuales (Brownlee, 2016). Este tipo de redes neuronales son comúnmente empleadas en situaciones en las que existe una dependencia entre una observación y observaciones pasadas.

Tal dependencia se incorpora a través de un estado oculto, o memoria, que guarda la esencia de lo que se ha visto hasta ahora. El valor del estado oculto en cualquier punto en el

tiempo es una función del valor del estado oculto en el paso de tiempo anterior y el valor de la entrada en el paso de tiempo actual, la siguiente imagen describe de mejor manera la incorporación del estado oculto, dependiente del paso de tiempo anterior (Gulli, Kapoor, & Pal, 2019).

Figura 14.
Diagrama esquemático de una celda de una red neuronal recurrente – RNN.



Tomado de Deep Learning with TensorFlow 2 and Keras. Second Edition. (2019). Por Gulli, A., Kapoor, A., & Pal, S.

Aquí, $h_t = \varphi(h_{t-1}, x_t)$ es una función en términos de la entrada en el paso de tiempo actual y el valor oculto del paso anterior, por lo que h_{t-1} puede ser representado en términos de h_{t-2} y así sucesivamente hasta el comienzo de la secuencia.

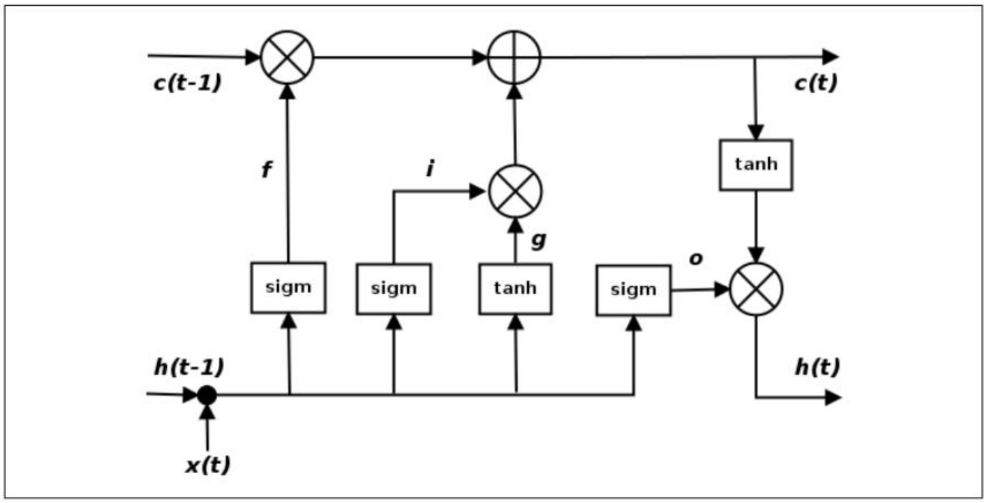
Específicamente, h_t viene dado por $h_t = \tanh(W.h_{t-1} + U.x_t)$ y $y_t = \text{softmax}(V.h_t)$, donde U, V y W corresponden a las matrices de los pesos de la entrada, salida y estados ocultos.

4.4.3.1.Long-Short Term Memory – LSTM.

Corresponde a una variación de la RNN, siendo capaz de aprender dependencias tanto de largo como de corto plazo. Al ser una variante de RNN aplica la recurrencia de una forma similar a la indicada en el numeral anterior, pero en este caso, en lugar de una sola capa con función de activación tanh, se incorporan 4 capas que interactúan para definir el estado de la celda LSTM, así como el estado oculto. Dichas capas corresponden a compuertas denominadas como compuerta de entrada, compuerta de olvido y compuerta de salida, mientras que *g* representa el estado oculto. La compuerta de olvido define cuánto del estado anterior h_{t-1} se desea permitir, la compuerta de entrada define cuánto del estado recién calculado para la entrada actual x_t se

desea dejar pasar, y la compuerta de salida define cuánto del estado interno se desea exponer a la siguiente capa (Gulli, Kapoor, & Pal, 2019).

Figura 15.
Diagrama esquemático de una celda de una red neuronal LSTM

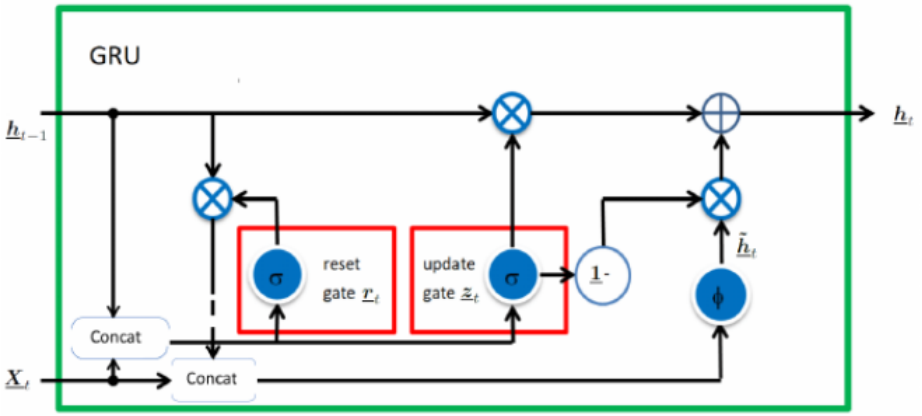


Tomado de Deep Learning with TensorFlow 2 and Keras. Second Edition. (2019). Por Gulli, A., Kapoor, A., & Pal, S.

4.4.3.2.Gated Recurrent Unit – GRU.

Corresponde a una variante de LSTM, pero presenta una estructura interna más simple, siendo las redes GRU más rápidas de entrenar que las LSTM. Para el caso de las redes neuronales GRU se emplean solo 2 compuertas en lugar de las 3 empleadas por las LSTM como compuerta de entrada, compuerta de olvido y compuerta de salida. En este caso se emplean una compuerta de actualización z y una compuerta de reinicio r, la primera define cuánta memoria anterior mantener y la segunda define cómo combinar la nueva entrada con la memoria anterior (Gulli, Kapoor, & Pal, 2019).

Figura 16.
Diagrama esquemático de una celda de una red neuronal GRU



Tomado de On Extended Long Short-term Memory and Dependent Bidirectional Recurrent Neural Network (2019). Por Yuanhang Su y Chung-Chieh Jay Kuo.

4.5. Evaluación / Métricas.

Para esta etapa, ya se han construido los diferentes modelos, obteniendo unos resultados específicos. Dichos resultados han de ser comparados con las premisas establecidas inicialmente, con las observaciones reales, así como también se ha de emplear las métricas definidas para evaluar el desempeño de cada modelo, se deberá realizar ajuste fino a los hiperparámetros a fin de obtener el mejor resultado posible. El capítulo 5 aborda la ejecución de los modelos y las diferentes iteraciones realizadas según aplique a cada modelo para lograr los resultados obtenidos finalmente. Las métricas empleadas, sobre las que serán evaluados los modelos, corresponden a aquellas descritas en el numeral 1.4

4.6. Despliegue.

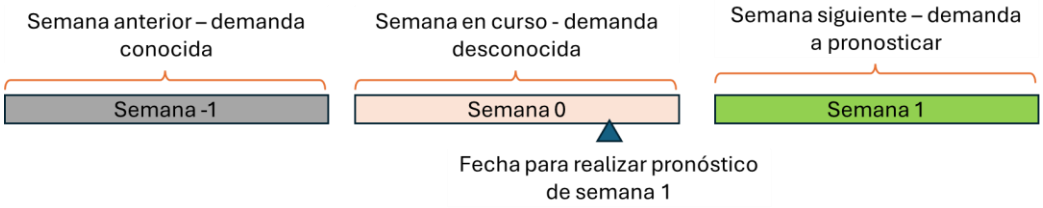
Entendiendo que el modelo en sí mismo no es el fin en sí mismo de un proyecto de analítica y ciencia de datos, sino que este debe ser puesto a disposición del negocio para que sea aplicado en los diferentes procesos de negocio que lo requieren. Esta corresponde a la última etapa de la metodología, la cual, dado el alcance del presente trabajo, no será abordada. El despliegue final dependerá en sí mismo de la arquitectura propia del agente (Distribuidor/comercializador de energía) que desee adoptar el modelo con mejores resultados, adecuándose con las fuentes que tenga a disposición y los sistemas internos.

5. Metodología.

Para el desarrollo del presente trabajo, se hace uso de diferentes modelos que permitan comparar el nivel de ajuste que presentan a los datos, a fin de elegir finalmente aquel que presente mejores respuestas al problema de negocio. Específicamente, se trabajará con modelos basados en Regresiones Lineales, modelos autorregresivos integrados de promedio móvil con componente estacional y factores exógenos como el SARIMAX y por último se evaluarán diferentes configuraciones de modelos de redes neuronales recurrentes como lo son las Simple RNN, modelos LSTM y modelos GRU.

Ahora bien, se considera la siguiente casuística asociada al proceso de pronóstico de la demanda: el proceso de pronóstico de la demanda ha de realizarse la semana anterior a la semana objetivo, denotemos la semana objetivo de pronóstico la semana 1 y la semana en la que se realiza el proceso de pronóstico como semana 0, sin embargo, dado que la semana 0 se encuentra en curso, no se cuenta con la demanda real de dicha semana, por lo que se ha de trabajar con información histórica hasta la semana inmediatamente anterior, es decir la semana -1. Esto implicará que, según el modelo aplicado, se deba realizar el pronóstico para 15 días en lugar del pronóstico de 7 días, es decir pronosticar la semana 0, cuyo pronóstico servirá para el pronóstico de la semana 1. En la Figura 17 ilustra mejor el proceso de preparación de los datos para el modelo de Regresiones lineales.

Figura 17.
Proceso de pronóstico de la demanda para la semana siguiente.



Elaboración propia.

5.1. Línea Base

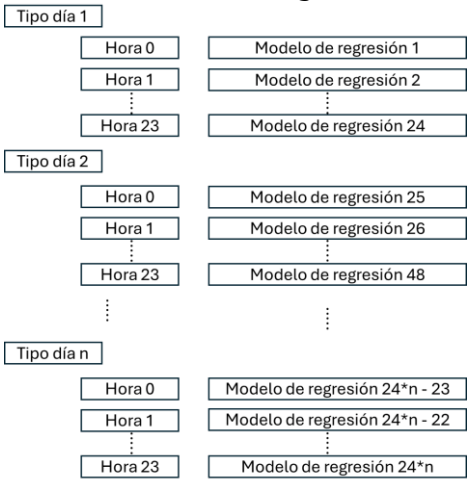
5.1.1. Regresión lineal.

Entendiendo el comportamiento característico de los datos, donde se aprecia que, se presentan patrones cíclicos cada 24 horas y que adicionalmente se tienen magnitudes en la

demanda de energía eléctrica diferentes según el tipo de día, no se trabaja con un modelo único de regresión lineal sino que se establecerá previamente clústers según tipos de día según su afinidad y a partir de esta clasificación se procede con el establecimiento de un modelo de regresión por cada tipo de día para cada una de las 24 horas del día (Ver Figura 18). Lo anterior con el objetivo de encontrar el mejor ajuste posible a los datos de la demanda real de energía eléctrica, de modo tal que se iterará de una forma como se ilustra a continuación.

Figura 18.

Aplicación iterativa de modelos de regresión lineal



Elaboración propia

Ahora bien, considerando que el dataset contempla 43 tipos de día diferentes, se procede a la simplificación o reducción de la cantidad de tipos de día a partir del comportamiento de la demanda de energía en cada una de las 24 horas del día. Para esto se emplea el Análisis de Componentes Principales – PCA – a fin de reducir la dimensionalidad de cada elemento – que para este propósito corresponde a un día con 24 características, una para cada hora del día – dicho análisis arroja una Proporción de varianza explicada por cada componente principal de la siguiente manera para los primeros 3 componentes principales:

Componente principal 1: 0.91

Componente principal 2: 0.06

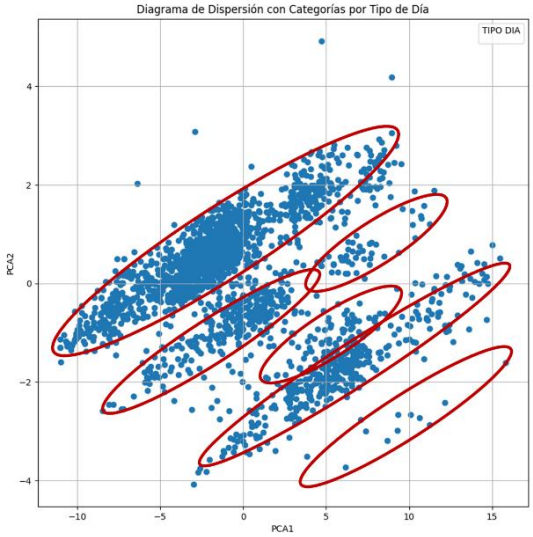
Componente principal 3: 0.01

Si bien, el primer componente principal explica en un 91% la varianza, se trabaja con los 2 componentes principales para realizar el análisis correspondiente para el establecimiento de una nueva agrupación de tipo de día. Al realizar el diagrama de dispersión a partir de los 2

componentes principales, se obtiene un resultado como el ilustrado en la Figura 19; en esta se aprecia la presencia de algunos patrones que pueden orientar la agrupación por tipo de día en una cantidad de categorías inferior a la originalmente presentada en el dataset.

Figura 19.

Diagrama de dispersión – PCA - con 2 componentes principales.

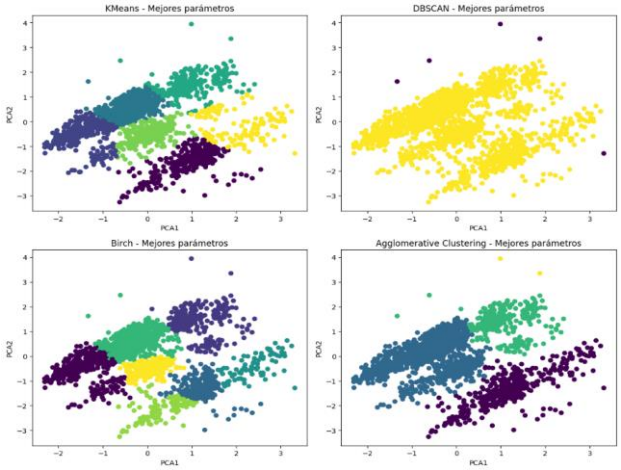


Elaboración propia

Con miras a identificar si en efecto es posible generar una agrupación para disminuir la cantidad de tipos de día, se ejecuta una serie de modelos de agrupamientos o clusters, sobre los cuales se itera para encontrar los mejores parámetros. Los modelos de cluster aplicados son K-MEANS, DBSCAN BIRCH y AGGLOMERATIVE CLUSTERING obteniendo el resultado presentado en la Figura 20.

Figura 20.

Aplicación de modelos de cluster.



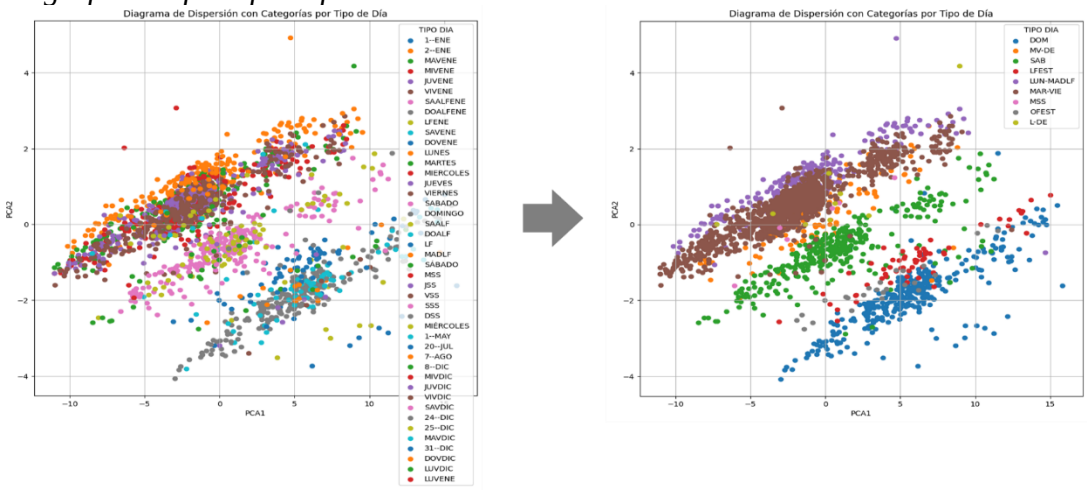
Elaboración propia.

Dado que los resultados obtenidos mediante los métodos de cluster mencionados genera un conjunto de clústers que, para un mismo tipo de día, se incluyen elementos en diferentes grupos, se opta por realizar un análisis visual a partir del diagrama de dispersión, de modo tal que a partir de cada tipo de día en el dataset original, se identifiquen similitudes entre estos que den pie a una agrupación con menor cantidad de categorías. El resultado obtenido se ilustra en la Figura 21, donde se aprecia que se logra una agrupación tal que la cantidad de categorías resultantes de tipos de días es de 8.

Por otra parte, con miras a establecer las características que mayor correlación pueden tener con la demanda de energía para la posterior aplicación de los modelos de regresión lineal, se procede inicialmente con el análisis de autocorrelación, entendiendo que se trata de una serie de tiempo, se pretende identificar en qué medida una observación depende de las observaciones pasadas; posteriormente se realiza el diagrama de correlación con las demás características.

Figura 21.

Agrupación final por tipo de día.



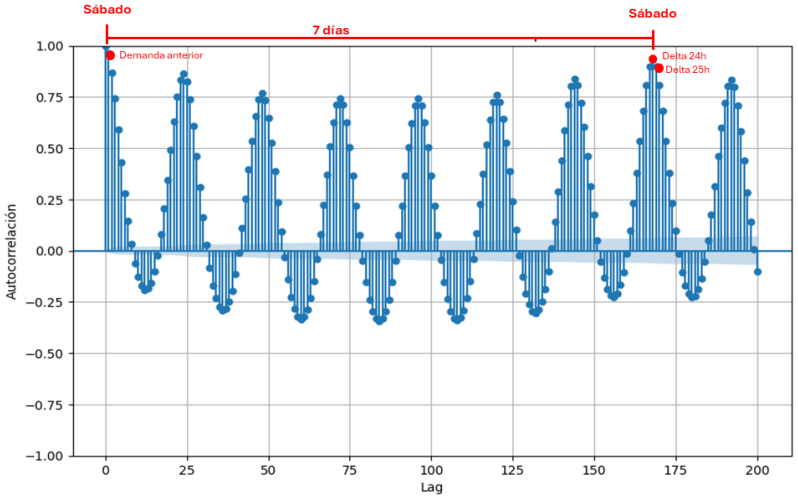
Elaboración propia.

Por otra parte, con el diagrama de autocorrelación aplicado a la variable de la demanda se evidencia que en efecto existe una alta dependencia de las observaciones pasadas, pero que en especial hay alta dependencia con el último día del mismo tipo. La Figura 22 muestra el ejercicio de autocorrelación, donde, a modo de ejemplo se puede observar que la mayor autocorrelación se presenta con la observación inmediatamente anterior seguida de la demanda a la misma hora 7 días atrás, motivo por el cual el dataset se complementa con 3 características adicionales

correspondientes a la demanda inmediatamente anterior, la demanda del mismo período de tiempo del último día del mismo tipo y la demanda observada una hora antes a esta última.

Figura 22.

Análisis de autocorrelación de la demanda de energía.

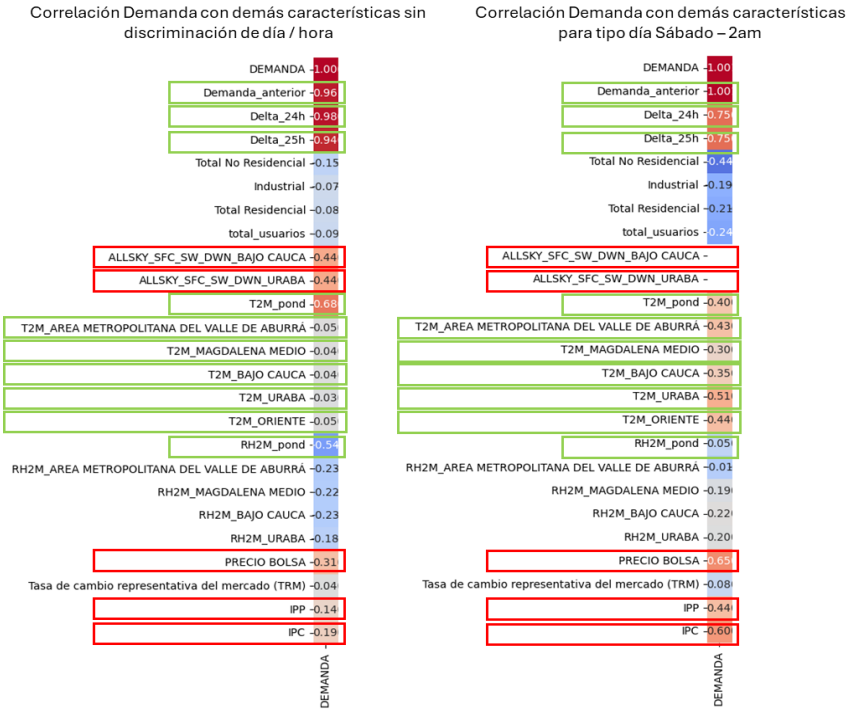


Elaboración propia.

Por su parte, el análisis de correlación arroja, además de una alta correlación con la misma demanda – a partir de las 3 características añadidas de la demanda en momentos anteriores – presenta un nivel de correlación con variables como la temperatura ponderada – cabe recordar que la variable de temperatura ponderada corresponde a un ejercicio de estimación de una temperatura única para el territorio en función de la cantidad de usuarios en cada municipio – así como con variables como la humedad relativa – en este caso una correlación inversa – y con variables macroeconómicas como el IPP y el IPC así como con el precio de energía en bolsa – sin embargo estas últimas serán descartadas para el análisis producto del análisis realizado a partir del comportamiento de las variables presentado en la Figura 24. La Figura 23, parte izquierda, muestra el ejercicio de correlación para la demanda con las demás variables para el dataset completo, es decir, antes de realizar una segmentación por tipo de día / hora, a la derecha, se ilustra el mismo ejercicio, pero esta vez para un tipo de día sábado a las 2 am, en el cual se aprecia que existe una correlación más alta con la medición de la temperatura en unas zonas en particular que con la temperatura ponderada, porque el ejercicio de regresión se abordará en primera instancia con la temperatura ponderada y en un segundo momento con las mediciones de temperatura promedio de las regiones señaladas en verde.

Figura 23.

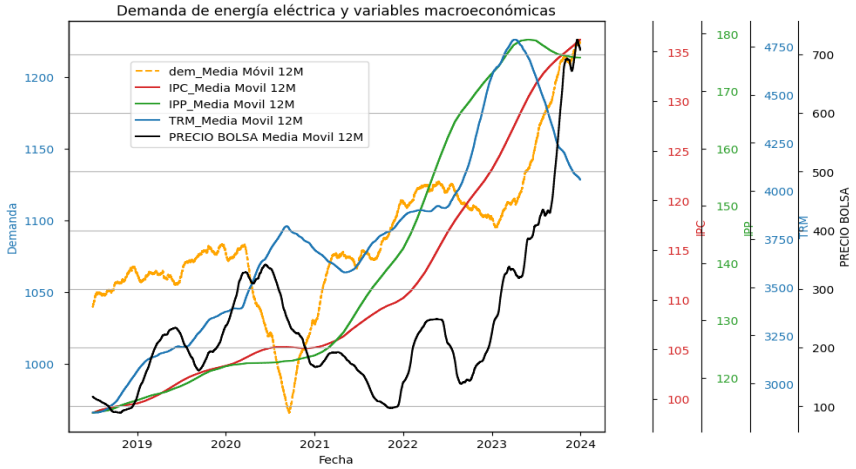
Mapa de correlación de las características numéricas del dataset



Elaboración propia.

Figura 24.

Tendencia Demanda Energía vs variables macroeconómicas.



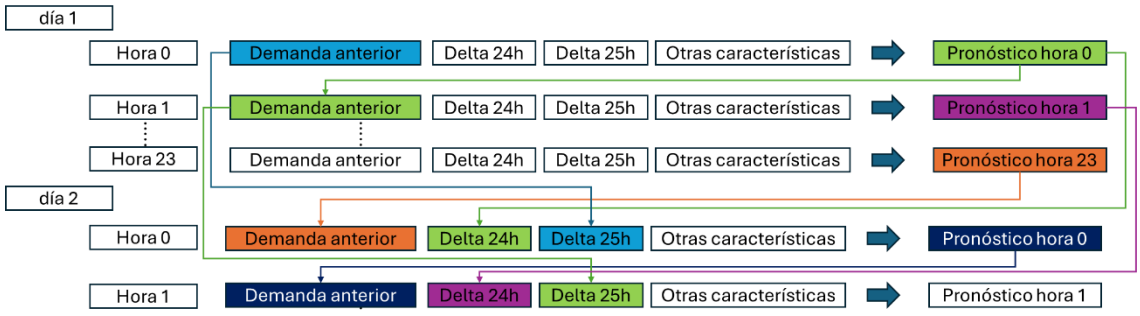
Elaboración propia.

De la Figura 24, se aprecia que si bien, puede existir alguna correlación entre las variables macroeconómicas y la demanda de energía, existen momentos en los que la tendencia cobra comportamientos opuestos ante cambios repentinos. Por su parte, para el caso del precio de energía en bolsa, se infiere que esta última es la que depende de la demanda con un ligero efecto

retardado, dado que de forma general el precio de energía en bolsa tiende a incrementarse en la medida que la demanda se incrementa ocasionando presión sobre los recursos de generación disponibles y en ocasiones condiciones de posible escasez o riesgos de desabastecimiento ante la concurrencia de fenómenos como el niño.

Una vez con la selección de las características a ser consideradas, se procede con un ejercicio iterativo para la creación de cada uno de los modelos de regresión lineal, los cuales son almacenados en un arreglo por tipo de día y por hora de modo tal que posteriormente se aplique el pronóstico. Es importante resaltar que para realizar el pronóstico, se ha de tener en cuenta que, dado que se han incluido características con información de la demanda anterior y 2 observaciones de la demanda del último día del mismo tipo, se hace necesario que se realice el pronóstico de forma ordenada hora a hora y que el pronóstico de cada hora sea llevado como característica al siguiente pronóstico – se convertiría en el atributo "Demanda anterior" del siguiente paso; así mismo, si el siguiente día es del mismo tipo, se deberá alimentar de forma similar las características denotadas como Delta 24h y Delta 25h, en un proceso que se ilustra en la Figura 25.

Figura 25.
Proceso de retroalimentación de las variables predictoras asociadas a la demanda a partir de cada pronóstico.



Nota: la figura supone que el día 1 y el día 2 son del mismo tipo de día, por ejemplo, martes y miércoles, los cuales, según el ejercicio de clusterización quedan agrupados bajo la misma categoría del tipo de día denominado "MAR-VIE". Elaboración propia.

Para el proceso de aplicación de los modelos de regresión lineal, se emplea la librería scikit-learn de Python, específicamente la función LinearRegression, empleando las siguientes variables predictoras: *'Demanda_anterior', 'Delta_24h', 'Delta_25h', 'T2M_pond', 'T2M_AREA*

METROPOLITANA DEL VALLE DE ABURRÁ', 'T2M_URABA', 'T2M_ORIENTE','RH2M_pond'.

En la Tabla 9 se presentan los resultados de las métricas de los modelos de Regresiones Lineales usando el conjunto de prueba.

Tabla 9. Resultados de las métricas de desempeño para el modelo inicial a partir de regresiones lineales.

Modelo	MAPE	MAE	RMSE	MSE	R^2
Regresiones lineales	0.00593	7.32861	10.08507	101.70879	0.99788

Nota. Las métricas anteriores se encuentran en las siguientes unidades: i) MAPE: %, ii) MAE: MW, iii) RMSE: KW, iv) MSE: MW^2 , v) R^2 adimensional.

5.1.2. **SARIMAX.**

La aplicación del modelo SARIMAX, parte la elección de los parámetros

p - Orden del componente autoregresivo: 1
d: Orden de diferenciación: 1
q: Orden del componente de media móvil: 1
P: Orden del componente autoregresivo estacional: 1
D: Orden de diferenciación estacional: 1
Q: Orden del componente de media móvil estacional: 1
s: Longitud del ciclo estacional: 24

El parámetro *s*, asociado a la longitud estacional se selecciona en 24, considerando que el patrón del comportamiento de la demanda responde a un ciclo, en primera instancia que se repite cada 24 horas. La variable endógena corresponde a la DEMANDA y como variables exógenas se seleccionan Delta_24h y Delta_25h (Explicados en el numeral anterior), hora del día, cantidad de usuarios residenciales e industriales, temperatura promedio diferenciada en las regiones Área Metropolitana y Urabá, complementando las variables exógenas con el tipo de día al que corresponden los datos, para esto se hace un proceso de "One Hot Encoding" a fin de llevar la variable tipo de día de un tipo string a tipo numérica.

En este caso se obtienen los siguientes que se presentan en la Tabla 10.

Tabla 10. Resultados de las métricas de desempeño para el modelo inicial SARIMAX.

Modelo	MAPE	MAE	RMSE	MSE	R^2
SARIMAX	0.06411	78.88616	100.08253	10016.51410	0.79149

Nota. Las métricas anteriores se encuentran en las siguientes unidades: i) MAPE: %, ii) MAE: MW, iii) RMSE: KW, iv) MSE: MW^2 , v) R^2 adimensional.

5.1.3. **Redes Neuronales Recurrentes.**

En el caso de Redes Neuronales Recurrentes se emplean redes Recurrentes Simples, Redes LSTM y Redes GRU. Para los tres casos se realiza un ejercicio inicial considerando las arquitecturas indicadas en la Tabla 11, las cuales son propuestas por Géron (2019) y denominadas como redes "Secuencia a Vector".

Tabla 11. Arquitectura empleada inicialmente en las redes neuronales recurrentes. Secuencia a Vector con predicción paso a paso y predicción múltiple.

Secuencia a Vector - Predicción paso a paso	Secuencia a Vector - Predicción múltiples pasos
- GRU(48 , return_sequences=True, input_shape=[none, 1]), - GRU(48), - Dense(24,activation = 'relu'), - Dense(12,activation = 'relu'), - Dense(1)	- GRU(48 , return_sequences=True, input_shape=[none, 1]), - GRU(48), - Dense(96,activation = 'relu'), - Dense(192,activation = 'relu'), - Dense(336)

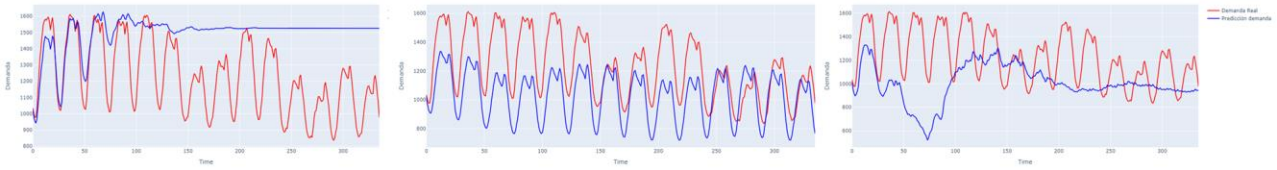
El entrenamiento se hace con una cantidad fija de épocas y sin emplear características adicionales más que la Demanda. La capa de entrada se alimenta con 48 pasos hacia atrás para realizar el pronóstico de 336 pasos hacia adelante, considerando la necesidad inicial de pronosticar tanto la semana en curso como la siguiente (De acuerdo con lo indicado en la figura 17 anteriormente), se eligen 20 épocas y no se elige un tamaño específico de batch, por lo que se emplea el tamaño por defecto. Los resultados iniciales para los 3 modelos se muestran en la Tabla 12 para el período comprendido entre el 18-03-2024 y el 31-03-2024.

Tabla 12. Resultados de las métricas de desempeño para los modelos basados en redes neuronales. Primera iteración.

Modelo	MAPE	MAE	RMSE	MSE	R^2
Simple RNN S2V – 1 step	0.26723	285.08405	351.37301	123462.99391	-1.55683
Simple RNN S2V – 336 steps	0.09129	103.13905	148.69027	22108.79539	0.54214
LSTM S2V – 1 step	0.17029	216.65774	249.97574	62487.87067	-0.29408
LSTM S2V – 336 steps	0.08571	95.62792	140.10005	19628.02309	0.59352
GRU S2V – 1 step	0.20207	267.05155	348.92808	121750.80289	-1.52137
GRU S2S – 336 steps	0.094606	108.41762	153.46899	23552.73367	0.51224

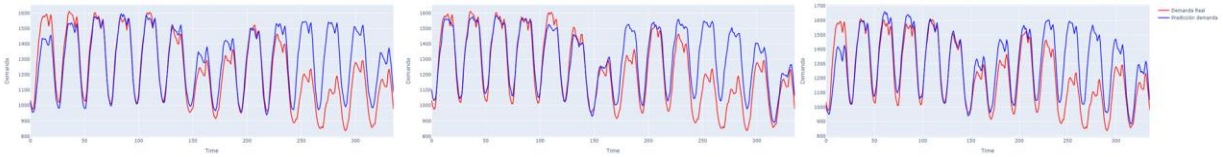
Nota. Las métricas anteriores se encuentran en las siguientes unidades: i) MAPE: %, ii) MAE: MW, iii) RMSE: KW, iv) MSE: MW^2 , v) R^2 adimensional.

Figura 26.
Resultado pronóstico del período 18-03-2024 y el 31-03-2024 con redes neuronales con el pronóstico de un solo paso.



***Nota**: izquierda: Single RNN, centro: LSTM, derecha: GRU.*

Figura 27.
Resultado pronóstico del período 18-03-2024 y el 31-03-2024 con redes neuronales con el pronóstico de 336 pasos simultáneos.



***Nota**: izquierda: Single RNN, centro: LSTM, derecha: GRU. Elaboración Propia.*

Como se puede observar en la Figura 26, los modelos basados en el pronóstico a partir de 1 paso (1 hora), presentan un desempeño muy bajo, por lo que se descarta seguir trabajando con estos, y se abordarán las próximas iteraciones con modelos fundamentados en el pronóstico simultáneo de múltiples pasos adelante como se presentan en la Figura 27.

5.2. Validación.

Para la validación y prueba de los diferentes modelos se emplea la misma partición, esta se ha explicado previamente en el numeral 3.2.1. Si bien se cuenta con un conjunto de datos con información de varios años, se opta por trabajar con un conjunto de validación igual al conjunto de pruebas para el entrenamiento de los modelos basados en redes neuronales; lo anterior, buscando no disminuir la cantidad de datos para el entrenamiento, ya que dadas las características del comportamiento de la demanda según el tipo de día y la hora del día, así como por las diferentes combinaciones que se puedan presentar en una semana (El cual es el período de tiempo a pronosticar), es de especial interés contar con el mayor número de datos posible para que dichas combinaciones puedan presentarse múltiples veces en los datos de entrenamiento.

5.3. Iteraciones y evolución.

5.3.1. Regresiones lineales.

Considerando el ejercicio inicial, el cual arroja métricas que cumplen con la premisa inicial, se procede a realizar una nueva iteración a fin de evaluar una nueva configuración en la cual se tenga en cuenta una menor cantidad de variables predictoras. En este caso, las variables consideradas son aquellas asociadas a la Demanda en momentos anteriores y la temperatura ponderada, es decir *'Demanda_anterior', 'Delta_24h', 'Delta_25h', 'T2M_pond'*; en este caso los resultados obtenidos se presentan en la Tabla 13.

Tabla 13. Resultados de las métricas de desempeño para el segundo modelo construido a partir de regresiones lineales.

Modelo	MAPE	MAE	RMSE	MSE	R^2
Regresiones lineales	0.00572	7.06088	9.55080	91.21794	0.99810

Nota. Las métricas anteriores se encuentran en las siguientes unidades: i) MAPE: %, ii) MAE: MW, iii) RMSE: KW, iv) MSE: MW^2 , v) R^2 adimensional.

En este caso, se aprecia que el segundo modelo presenta una ligera mejoría con respecto al primer modelo creado.

5.3.2. **SARIMAX**

Dados los resultados obtenidos en la primera iteración, se procede a realizar un cambio de enfoque. En este caso, se cambia la variable endógena, y en lugar de emplear la variable DEMANDA, se emplea la variable Delta_24h, con el fin de que la variable endógena cuente con datos específicos de la demanda anterior para el mismo tipo de día; las variables exógenas continúan siendo las mismas, a excepción de Delta_24hl la cual se elimina de las variables exógenas. En el caso de los parámetros del modelo, se trabaja con los mismos del modelo inicial, esto es *p:1, d:1, q:1, P:1, D:1, Q:1, s:24.* Los resultados obtenidos se presentan en la Tabla 14.

Tabla 14. Resultados de las métricas de desempeño para la segunda iteración del modelo SARIMAX.

Modelo	MAPE	MAE	RMSE	MSE	R^2
SARIMAX 24H	0.03647	43.79864	58.65267	3440.13576	0.92838

Nota. Las métricas anteriores se encuentran en las siguientes unidades: i) MAPE: %, ii) MAE: MW, iii) RMSE: KW, iv) MSE: MW^2 , v) R^2 adimensional.

Ahora bien, buscando mejorar los resultados, se procede con un proceso iterativo a través de distintos valores para *p,d,q* sólo para dos semanas, específicamente para las semanas

comprendidas entre el 29/01/2024 y el 11/02/2024, se evalúa con respecto al MAPE, con los resultados presentados en la Tabla 15.

Tabla 15. Resultados de las métricas de desempeño para el modelo SARIMAX variando p,d,q y dejando fijos P, D, Q, s, en 1, 1, 1, 24 respectivamente.

p	d	q	MAPE
1	1	1	0.037672
1	1	2	0.037796
1	1	3	0.037722
1	2	1	0.061699
1	2	2	0.066053
1	2	3	0.064880
2	1	1	0.037661
2	1	2	0.061860
2	1	3	0.039694
2	2	1	0.28133

De la Tabla 15 se aprecia que no existe mejora sustancial con los diferentes parámetros de *p,d,q* probados, por lo que se trabajará con los valores de 1, 1 y 1 para los parámetros del modelo, respectivamente.

5.3.3. **Redes neuronales.**

Considerando el ejercicio inicial realizado con redes neuronales, se deja de lado las iteraciones sobre los modelos basados en el pronóstico de un paso hacia adelante (1 hora) para explorar aquellos basados en el pronóstico de manera simultánea de múltiples pasos. En este sentido, se adopta una arquitectura adicional, la cual corresponde a la descrita por Géron (2019) como “Secuencia a Secuencia”, sin embargo, en esta última se ignoran todas las salidas a excepción de la última, la cual será aquella que contenga la predicción para el rango de tiempo deseado. Para mayor detalle en estos dos enfoques se recomienda remitirse a (Géron, 2019). A

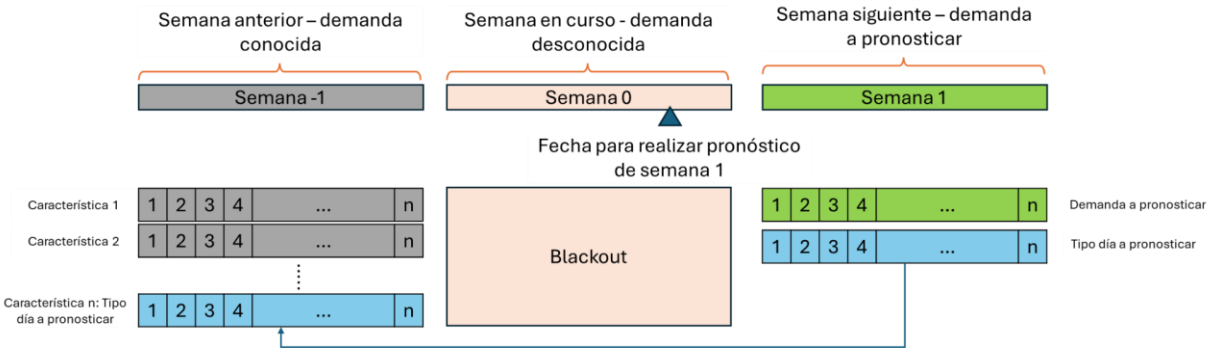
continuación, se ilustra la arquitectura empleada en ambos enfoques que se muestra en la Tabla 16.

Tabla 16. Arquitectura empleada en las redes neuronales recurrentes. Secuencia a Vector y secuencia a Secuencia.

Secuencia a Vector	Secuencia a Secuencia
- GRU(neurons ,activation = activ_func, return_sequences=True, input_shape=input_shape), - GRU(48,activation = activ_func), - Dense(96,activation = activ_func), - Dense(192,activation = activ_func), - Dense(steps_ahead)	- GRU(neurons ,activation = activ_func, return_sequences=True, input_shape=input_shape), - GRU(48,activation = activ_func, return_sequences=True), - TimeDistributed(Dense(96,activation = activ_func)), - Dense(192,activation = activ_func), - Dense(steps_ahead)

Considerando los enfoques señalados en la Tabla 16, se varía entre una red SimpleRNN, una LSTM y una GRU, se prueba con funciones de activación *Relu* y *Sigmoid;* así como también se aplica un early stopping con un número máximo de épocas de 30 y con miras a buscar que el modelo generalice mejor el comportamiento de la demanda se aplica un batch size de 10. Por su parte, se realizan pruebas variando la capa de salida, considerando una salida de 24 horas o una salida de 168 horas (correspondiente a 7 días). La forma de los datos de entrada a la red también se varía, alimentándose con 24 horas y 168 horas respectivamente. El arreglo de entrada es alimentado con la Demanda histórica, y en esta ocasión, se alimenta adicionalmente con las demás características relevantes como temperatura, cantidad de clientes e incluso con el tipo de día al que corresponde cada paso a pronosticar; para esto último se hace necesario que, tanto los pasos a pronosticar coincidan con la cantidad de datos históricos a considerar, de modo que sea posible alimentar la red con esta característica. Ahora bien, de acuerdo con lo ilustrado en la Figura 17, en este caso se organizan los datos para no considerar la información de la semana 0, correspondiente a la semana en curso, por lo que se denomina esta semana como un “blackout”. La Figura 2 ilustra de mejor manera este enfoque.

Figura 28.
Esquema empleado para alimentar las redes neuronales con la característica "Tipo de día a pronosticar".



Elaboración propia.

Al aplicar dichos enfoques, se obtienen mejores resultados al aplicar las funciones de activación de tipo sigmoidal y los resultados se muestran en la Tabla 17.

Tabla 17. Resultados de las métricas de desempeño para los modelos basados en Redes Neuronales.

Modelo	MAPE	MAE	RMSE	MSE	R^2
Simple RNN S2V – 24 steps	0.03975	46.93422	62.31230	3882.82379	0.91917
Simple RNN S2S – 24 steps	0.04299	51.85146	68.78180	4730.93618	0.90151
LSTM S2V – 24 steps	0.04575	55.18659	69.52263	4833.39628	0.89938
LSTM S2S – 24 steps	0.03688	44.44053	64.11581	4110.83744	0.91442
GRU S2V – 24 steps	0.03136	38.10264	52.44211	2750.17531	0.94275
GRU S2S – 24 steps	0.04265	50.79737	63.74504	4063.43138	0.91541
Simple RNN S2V – 168 steps	0.05261	63.95714	86.90541	7552.55038	0.84278
Simple RNN S2S – 168 steps	0.05264	63.85665	82.04113	6730.74837	0.85988
LSTM S2V – 168 steps	0.03631	43.10526	58.30262	3399.19582	0.9292
LSTM S2S – 168 steps	0.03073	36.43716	56.85891	3232.93658	0.93270
GRU S2V – 168 steps	0.02634	31.60541	42.70185	1823.44859	0.96204
GRU S2S – 168 steps	0.03210	38.73854	56.14585	3152.35724	0.93437

Nota. Las métricas anteriores se encuentran en las siguientes unidades: i) MAPE: %, ii) MAE: MW, iii) RMSE: KW, iv) MSE: MW^2 , v) R^2 adimensional. S2V hace referencia a los modelos tipo Secuencia a vector y S2S a aquellos basados en Secuencia a Secuencia.

5.4. Herramientas.

El desarrollo de los diferentes modelos y herramientas de clasificación han sido empleados de las diferentes funcionalidades disponibles en librerías de python como sklearn,

keras y tensorflow y los diferentes códigos han sido ejecutados empleando la herramienta de jupyter notebooks de google Colaboratory.

6. Resultados y Discusión

6.1. Métricas.

Una vez se han aplicado los modelos anteriormente descritos, se muestran a continuación los resultados correspondientes a los modelos con mejor desempeño en cada tipo, ver Tabla 18, considerando que los resultados corresponden al pronóstico semana a semana al período comprendido entre el 01-01-2024 hasta el 31-03-2024:

Tabla 18. Resultados de las métricas de desempeño para los modelos con mejores resultados.

Modelo	MAPE	MAE	RMSE	MSE	R^2
Regresiones lineales	0.00572	7.06088	9.55080	91.21794	0.99810
SARIMAX 24H	0.03647	43.79864	58.65267	3440.13576	0.92838
GRU S2V – 24 steps	0.03136	38.10264	52.44211	2750.17531	0.94275
LSTM S2S – 168 steps	0.03073	36.43716	56.85891	3232.93658	0.93270
GRU S2V – 168 steps	0.02634	31.60541	42.70185	1823.44859	0.96204

Nota. Las métricas anteriores se encuentran en las siguientes unidades: i) MAPE: %, ii) MAE: MW, iii) RMSE: KW, iv) MSE: MW^2 , v) R^2 adimensional.

Los resultados anteriores reflejan que el modelo con una mayor exactitud en el pronóstico es aquel desarrollado a partir de múltiples regresiones lineales con un MAPE de 0.00572, seguido por las redes GRU en su configuración de Secuencia a Vector con pronósticos de 168 horas con un MAPE de 0.02634; en tercer lugar se encuentra una red LSTM configurada para pronosticar 168 pasos con un MAPE de 0.03073; en cuarto lugar tenemos una red GRU configurada en Secuencia a Vector, pero en esta ocasión con un pronóstico de 24 horas, es decir, con el pronóstico de los 7 días de manera simultánea con un MAPE de 0.03136.

6.2. Evaluación cualitativa.

De acuerdo con las métricas indicadas en la sección anterior, se puede decir que estos 5 modelos cumplen con la premisa inicial ya que al analizar el MAPE, se presenta una desviación inferior al +/- 4% en el pronóstico con respecto a la demanda real, sin embargo, solo el modelo construido a partir de regresiones lineales presenta una consistencia durante todo el período de tiempo, conservando en su gran mayoría desviaciones muy por debajo del valor objetivo, esto debido a al enfoque adoptado de múltiples modelos discriminados por tipo de día y hora del día, lo que permite caracterizar con mucho detalle el comportamiento específico de cada período de

tiempo, a diferencia de los demás modelos en los que se construye un único modelo que debe estar en capacidad de generalizar el comportamiento de la demanda para todos los tipos de día y hora del día a partir de las características con las que se entrenan los modelos. Ahora bien, con miras a evaluar los resultados obtenidos con respecto a la exactitud del pronóstico actual realizarse para el Mercado de Comercialización de Antioquia, se procede a realizar un breve análisis descriptivo del porcentaje de error tanto para los modelos evaluados como para el pronóstico vigente, dicho análisis se presenta en la Tabla 19.

Tabla 19. Análisis descriptivo del porcentaje de desviación en el pronóstico.

Métrica	**PRONÓSTICO OR**	**Regresiones lineales**	**GRU S2V – 168 steps**	**LSTM S2S – 168 steps**	**GRU S2V – 24 steps**	**SARIMAX 24H**
1er cuartil	1.07%	0.19%	1.01%	1.26%	1.01%	1.26%
2do cuartil	2.51%	0.43%	2.06%	2.67%	2.20%	2.64%
3er cuartil	4.68%	0.77%	3.52%	4.52%	4.29%	4.76%
Min	0.0012%	0.000078%	0.0035%	0.00001%	0.0036%	0.00022%
Max	18.83%	6.83%	15.84%	51.13%	33.33%	28.45%
Media	3.38%	0.57%	2.55%	3.63%	3.14%	3.65%
Desv est	3.08%	0.56%	2.17%	4.57%	3.13%	3.58%
% horas superan el 4%	30.54%	0.27%	18.77%	30.40%	27.06%	31.96%

Nota: El % de horas superan el 4% corresponde a la cantidad de pronósticos que superan un error del 4% con respecto a la demanda real en el período de tiempo del 01-01-2024 al 31-03-2024.

De la tabla 19 se evidencia que el modelo basado en regresiones lineales y el de redes neuronales GRU secuencia a vector con el pronóstico de 168 pasos presentan un mejor comportamiento en cuanto a los porcentajes de error reportados por el Operador de Red (OR), y en gran medida el modelo GRU secuencia a vector con el pronóstico de 24 pasos también presenta un mejor desempeño en cuanto al porcentaje de error. Ahora bien, entendiendo que las métricas anteriores corresponden a todo el período comprendido entre el 01/01/2024 y el 31/03/2024, se procede a realizar el análisis puntual del pronóstico para 3 semanas específicas y de este modo determinar si los modelos pueden tener un mejor comportamiento que el pronóstico vigente para dichas semanas. Las semanas bajo las cuales se realiza este análisis son:

- Semana 1: 01/01/2024 - 07/01/2024
- Semana 6: 05/02/2024 - 11/02/2024

- Semana 13: 25/03/2024 - 31/03/2024

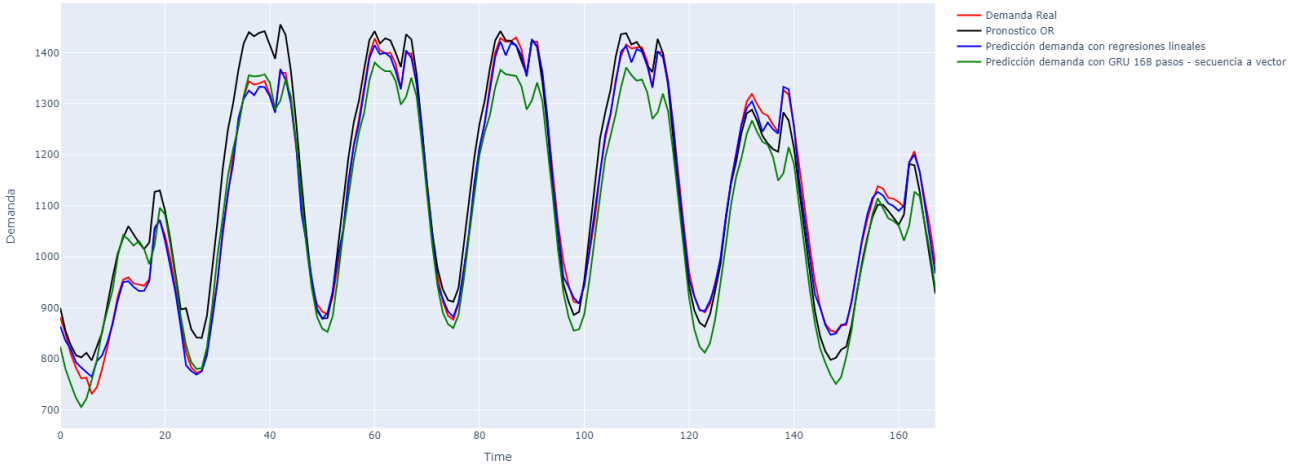
El análisis se basa en la comparación del MAPE para dichas semanas y los resultados que se obtienen se muestran en la Tabla 20.

Tabla 20. MAPE evaluado sobre el pronóstico de las semanas 1, 6 y 13.

Métrica	**PRONÓSTICO OR**	**Regresiones lineales**	**GRU S2V – 168 steps**	**LSTM S2S – 168 steps**	**GRU S2V – 24 steps**	**SARIMAX 24H**
Semana 1	0.04033	0.00866	0.04416	0.07816	0.03897	0.07347
Semana 6	0.03843	0.00542	0.02307	0.01916	0.04112	0.03442
Semana 13	0.08119	0.00863	0.04660	0.06559	0.07298	0.08284

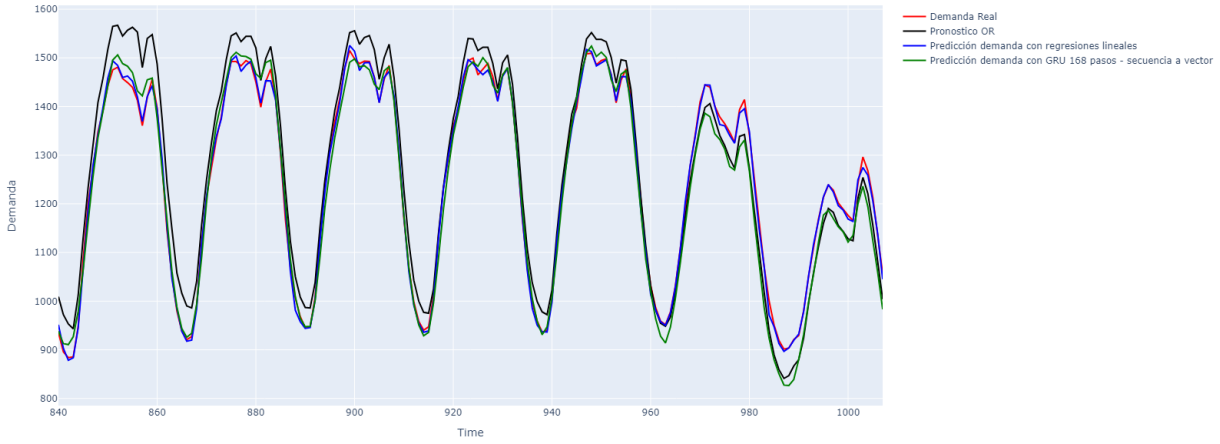
De la Tabla 20 se evidencia que, para las 3 semanas en específico el modelo basado en regresiones lineales presenta un mejor desempeño y que los demás modelos presentan mejor comportamiento para algunas de las semanas mientras que en algunas otras el pronóstico actual para el mercado de comercialización de Antioquia presenta una menor desviación; sin embargo, para aquellos en los que los modelos de redes neuronales o SARIMAX presentan un mejor desempeño, no existe una mejora demasiado significativa en comparación con la que arroja el basado en regresiones lineales. A continuación, se muestran en las Figuras 29, 30 y 31 el comparativo entre el pronóstico actual del OR y el pronóstico realizado con regresiones lineales y con la red GRU secuencia a vector con el pronóstico de 168 pasos.

Figura 29.
Comparativo del pronóstico de la semana 1 para diferentes modelos y pronóstico actual.



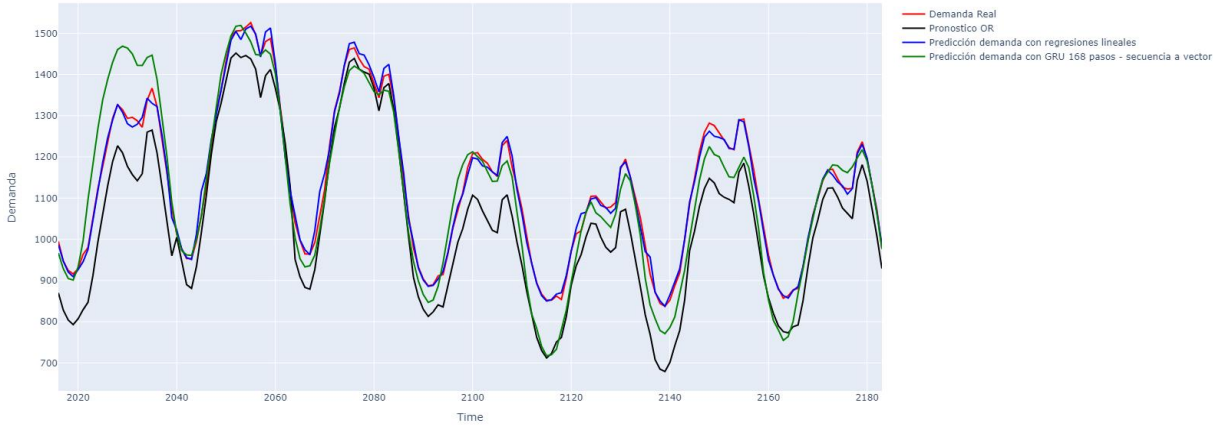
Elaboración propia.

Figura 30. *Comparativo del pronóstico de la semana 6 para diferentes modelos y pronóstico actual.*



Elaboración propia.

Figura 31.
Comparativo del pronóstico de la semana 13 para diferentes modelos y pronóstico actual.



Elaboración propia.

Se evidencia que, los 2 modelos graficados presentan un comportamiento más cercano a la demanda real que el pronóstico actual, salvo en algunos períodos de tiempo en los que el pronóstico actual es más acertado que el arrojado por la red neuronal GRU.

6.3. Consideraciones de producción.

Un futuro despliegue a producción de alguno de los modelos aquí planteados ha de buscar que sea posible automatizar el proceso de principio a fin, partiendo de la obtención de la información relevante como la demanda histórica, como información asociada a las variables climatológicas y la cantidad de usuarios por municipio, por lo que se debe desplegar los scripts

necesarios para el uso de la API dispuesta por la NASA para la obtención de las variables climatológicas o si por el contrario se tienen servicios climatológicos alternativos que brinden acceso a la información histórica para la geografía objeto de estudio, en este caso Antioquia. Así mismo se ha de desarrollar las respectivas integraciones con los sistemas corporativos específicos para el acceso a información relativa a la cantidad de usuarios por municipio.

Adicionalmente, es importante tener en consideración los tiempos requeridos correr tanto el entrenamiento del modelo, como para el pronóstico, así como definir, si, por ejemplo para el caso de las redes neuronales, se realizará el entrenamiento del modelo con los nuevos datos con una periodicidad superior a semanal considerando los nuevos datos disponibles o si se opta por realizar el reentrenamiento cada vez que se deba realizar el pronóstico, esto debido a los altos tiempos y consumo de recursos para dicho proceso de entrenamiento en las redes con un bajo "Batch size" y gran cantidad de pasos hacia atrás para alimentar la red como también gran cantidad de pasos a predecir. A continuación los tiempos requeridos para el entrenamiento y pronóstico de los modelos con mejores resultados.

- Regresiones lineales: Aproximadamente 1.8 segundos para el entrenamiento y el pronóstico de una semana. Corriendo en una máquina de Google Colab con 12.7 GB de Ram, 78.2 GB de Disco y GPU 15 GB.
- GRU24 steps: 1809 segundos para el entrenamiento y 0.6 segundos para el pronóstico. Corriendo en una máquina de Google Colab con 12.7 GB de Ram, 107.7 GB de Disco.
- GRU168 steps: Más de 900 segundos por época, en un proceso que puede tomar alrededor de 15 a 20 épocas para converger. Corriendo en una máquina de Google Colab con 53 GB de Ram, 78.2 GB de Disco y GPU 22.5 GB.

7. Conclusiones

El enfoque que se ha imprimido a la solución del problema a partir de modelos sencillos como lo son las regresiones lineales, los cuales por sí solos no brindarían una precisión adecuada dadas las características de la variable a pronosticar, el hecho de construir un esquema iterativo con múltiples modelos de regresión lineal para cada tipo de día y hora del día ha mostrado brindar una buena exactitud en el pronóstico de la demanda de energía a la vez que presentan un buen desempeño en consumo de recursos, lo que lo hacen una solución adecuada para su operativización o despliegue final, toda vez que implica una mínima demanda de recursos informáticos permitiendo así el reentrenamiento periódico del modelo en la medida que se cuenta con la información de la demanda real para realizar los pronósticos futuros. Sin bien, inicialmente se contemplaron múltiples características a ser consideradas como variables predictoras, finalmente aquella con mayor precisión corresponde a aquella que tan solo tuvo como variables predictoras aquellas relacionadas a la demanda en periodos anteriores y la temperatura ponderada en función de los usuarios de cada municipio.

Si bien las redes neuronales presentaron un desempeño inferior al modelo de regresiones lineales, amerita continuar con la exploración de diferentes arquitecturas, en especial con LSTM y GRU ya que se obtuvo un pronóstico con un nivel de precisión similar al que se realiza en la actualidad para el mercado de comercialización de Antioquia, e incluso con una ligera mejoría para el caso de la red GRU, con lo que se puede adoptar un enfoque similar al empleado con regresiones lineales, en el que se discrimine de mejor manera el tipo de día según sus características o profundizar en arquitecturas tipo funcionales con esquemas más complejos a los empleados en el presente trabajo. Ahora bien, es de importancia evaluar el desempeño de estos modelos con respecto al consumo de recursos informáticos y hasta qué punto es posible sacrificar la precisión del modelo a fin de tener un consumo de recursos razonable y oportuno para la dinámica del pronóstico que se debe realizar.

8. Recomendaciones

La solución planteada a partir de regresiones lineales presenta un desempeño que supera de momento a los demás modelos planteados, así como también presenta una mejora considerable con respecto al pronóstico empleado en la actualidad para el mercado de comercialización de Antioquia, por lo que se plantea su adopción para realizar el pronóstico semanal de la demanda de energía.

Adicionalmente, se recomienda evaluar arquitecturas y esquemas adicionales para las redes neuronales recurrentes de modo tal que se pueda lograr un nivel de precisión similar a logrado con regresiones lineales y que pueda ser empleado como alternativa a dicho modelo.

Se ha de evaluar el mecanismo más apropiado para su implementación a nivel de arquitectura de modo que sea posible automatizar el proceso, contemplando los scripts requeridos para el uso de la API dispuesta por la NASA para la obtención de los históricos de las variables climatológicas o el empleo de fuentes alternativas a las que se tenga acceso para tal fin. Del mismo modo, la adecuación a las consultas a sistemas internos que permitan el acceso a información de la cantidad de usuarios conectados a la red de energía en cada municipio, ya que para el presente trabajo se emplea la información disponible en el SUI, la cual es tan solo puesta a disposición al público con más de un mes de retraso.

Referencias

Brownlee, J. (2016). *Deep Learning With Python. Develop Deep Learning Models On Theano And TensorFlow Using Keras.*

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. Concepts, Tools, and Techniques to Build Intelligent Systems. Second Edition.* Canadá: O'Reilly Media, Inc.,.

Gulli, A., Kapoor, A., & Pal, S. (2019). *Deep Learning with TensorFlow 2 and Keras. Second Edition.* Birmingham: Packt Publishing.

Montgomery, D., Peck, E., & Vining, G. (2012). *Introduction to Linear Regression Analysis (5th Edition).* New Jersey: Wiley.

Shumway, R., & Stoffer, D. (2011). *Time series analysis and its applications with R examples. Third edition.* Springer.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0. Step-by-step data mining guide.*

Consejo Nacional de Operación [CNO]. (2020). *Acuerdo 1303, Por el cual se actualizan los procedimientos para la gestión integral de la demanda*.

Escobar, Luis; Valdés, Julio; Zapata, Santiago. (2010). *Redes Neuronales Artificiales en predicción de Series de Tiempo. Una aplicación a la Industria*. Obtenido de Universidad de Palermo: https://www.palermo.edu/ingenieria/Pdf2010/CyT9/02.pdf

Haya, P. (2021). *La metodología CRISP-DM en ciencia de datos*. Obtenido de Instituto de Ingeniería del Conocimiento: https://www.iic.uam.es/innovacion/metodologia-crisp-dm-ciencia-de-datos/

Ortuño, J. M., Ramos, J. A., & Senent, J. C. (2018). *Modelo Arima. Análisis Estadístico de Series Económicas*. Obtenido de https://rstudio-pubs-static.s3.amazonaws.com/384039_cc37e393f643455bb01ad4b392a081bd.html

Radečić, D. (26 de Julio de 2021). *Time Series From Scratch — Train/Test Splits and Evaluation Metrics*. Obtenido de Towards data science: https://towardsdatascience.com/time-series-from-scratch-train-test-splits-and-evaluation-metrics-4fd654de1b37

Su, Y., & Chung-Chieh, J. (2019). *On Extended Long Short-term Memory and Dependent. Bidirectional Recurrent Neural Network*.