Predicción de la Demanda de Energía Eléctrica Usando Modelos de Inteligencia Artificial para Series Temporales

Anderson Sebastián Torres Sánchez

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
2023

Cita	Torres Sánchez, A
Referencia **Estilo IEEE (2020)**	A.S. Torres Sánchez, "Predicción de la demanda de Energía Eléctrica Usando Modelos de Inteligencia Artificial para Series Temporales", Trabajo de grado especialización e Analítica y Ciencia de Datos, Universidad de Antioquia, Medellín, Colombia, 2023.

creative commons BY NC SA

Especialización en Analítica y Ciencia de Datos, Cohorte V.

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

El contenido de esta obra corresponde al derecho de expresión de los autores y no compromete el pensamiento institucional de la Universidad de Antioquia ni desata su responsabilidad frente a terceros. Los autores asumen la responsabilidad por los derechos de autor y conexos.

Tabla de contenido

Resumen 10

Abstract 11

1. Descripción del problema 12

1.1. Problema de negocio 14

1.2. Aproximación desde la analítica de datos 14

1.3. Origen de los datos 14

1.4. Métricas de desempeño 14

2. Objetivos 15

2.1. Objetivo general 15

2.2. Objetivos específicos 15

3. Marco Teórico 16

3.1. Autoregressive Integrated Moving Average (ARIMA) 16

3.2. Seasonal Autoregressive Integrated Moving Average with eXogenous regressors (SARIMAX) 17

3.3. Linear regressor autoregressive 18

3.4. Random Forest autoregressive 19

3.5. Redes neuronales recurrentes (RNN) 19

4. Metodología 22

4.1. Recopilación de datos: 22

4.2. Preprocesamiento de datos: 22

4.3. Análisis de datos: 22

4.4. Búsqueda de modelos: 22

4.5. Desarrollo de los modelos: 22

4.6. Evaluación y comparación: 22

4.7. Optimización y ajuste: 22

5. Datos 24
5.1. Extracción de dato originales 24
5.1.1. API XM: 24
5.1.2. API Socrata: 26
5.2. Preprocesamiento 28
5.2.1. Demanda de energía eléctrica en Antioquia: 28
5.2.2. Temperatura: 29
5.2.3. Combinación de datasets: 29
5.3. Limpieza de datos 31
5.4. Análisis de datos 32
5.4.1. Distribución de los datos 32
5.4.2. Descomposición estacional 33
5.4.3. Autocorrelación 36
5.5. División del datasets 37
6. Descripción de los modelos 38
6.1. Modelos 38
6.1.1. Autoregressive Integrated Moving Average (ARIMA) 38
6.1.2. Seasonal Autoregressive Integrated Moving Average Exogenous Variables (SARIMAX) 38
6.1.3. Linear Regressor Autoregressive 39
6.1.4. Random Forest Regressor Autoregressive 39
6.1.5. Recurrent Neural Network (RNN-LSTM) 40
6.2. Ajuste de hiperparámetros 42
6.2.1. Autoregressive Integrated Moving Average (ARIMA) 42
6.2.2. Seasonal Autoregressive Integrated Moving Average Exogenous Variables (SARIMAX) 42

6.2.3. Random Forest Regressor Autoregressive 43
6.2.4. RNN-LSTM 43
6.3. Métricas 44
7. Resultados y discusión 45
7.1. Métricas 45
7.2. Métricas modelos con ajuste de hiperparámetros 57
7.3. Consideraciones de producción 66
8. Conclusiones 67

Lista de tablas

Tabla 1 Resultados obtenidos de las métricas de validación para train y test 46

Tabla 2 Resultados obtenidos de las métricas de validación simplificadas 48

Tabla 3 Resultados obtenidos de las métricas de validación para train y test con ajuste de hiperparámetros 58

Tabla 4 Resultados obtenidos de las métricas de validación simplificadas con ajuste de hiperparámetros 59

Lista de figuras

Figura 1: Mapa de la distribución de recursos de generación del país, donde el tamaño representa la capacidad en MW de los recursos de generación. Elaboración propia a partir de los datos del API de XM. 12

Figura 2: Mapa de calor según el consumo promedio diario de energía eléctrica en las regiones del país. Elaboración propia a partir de los datos del API de XM. 13

Figura 3: Arquitectura de red neuronales recurrente [12]. 19

Figura 4: Arquitectura de red neuronales recurrente LSTM [12]. 20

Figura 5: Diagrama de la metodología general. Elaboración propia. 22

Figura 6: Estructura del dataset de la demanda de energía eléctrica en Antioquia. 26

Figura 7: Estructura del dataset de la temperatura en Antioquia. 27

Figura 8: Ejemplo del proceso de pivot y unpivot [14] 28

Figura 9: Estructura del dataset de la demanda de energía eléctrica en Antioquia luego de la transformación unpivot. 28

Figura 10: Estructura del dataset de temperatura en Antioquia luego de las transformaciones. 29

Figura 11: Estructura del dataset final. 30

Figura 12: Perfil de distribución de la demanda de energía eléctrica por hora. 32

Figura 13: Perfil de distribución de la demanda de energía eléctrica por día de la semana. 33

Figura 14: Grafico de visualización de la demanda de energía eléctrica en Antioquia de todo el dataset final. 34

Figura 15: Descomposición estacional de la demanda de energía eléctrica en Antioquia para todo el dataset. 34

Figura 16: Descomposición estacional de la demanda de energía eléctrica en Antioquia para los primeros meses del año del 2021. 35

Figura 17: Autocorrelación de la demanda de energía eléctrica en Antioquia para los primeros $n = 30$ instantes anteriores. 36

Figura 18: Separación en train y test de la demanda de energía eléctrica en Antioquia. 37

Figura 19: Ejemplo de cómo funciona la recursividad en los modelos autorregresivos [15] 49

Figura 20: Resultado de la predicción de la demanda de energía eléctrica en Antioquia del modelo ARIMA con variable exógena del temperatura promedio. 50

Figura 21: Resultado de la predicción de la demanda de energía eléctrica en Antioquia del modelo SARIMAX con variable exógena de temperatura promedio. 51

Figura 22: Resultado de la predicción de la demanda de energía eléctrica en Antioquia del modelo Linear Regressor Autoregressive con variable exógena del tipo de día.....52

Figura 23: Resultado de la predicción de la demanda de energía eléctrica en Antioquia del modelo Random Forest Autoregressive con variable exógena del tipo de día. 53

Figura 24: Resultado de la predicción de la demanda de energía eléctrica en Antioquia del modelo RNN LSTM con variable exógena del tipo de semana. 54

Figura 25: Resultado de la predicción de la demanda de energía eléctrica en Antioquia del modelo ARIMA con ajuste de hiperparámetros con variable exógena del tipo de semana. 60

Figura 26: Resultado de la predicción de la demanda de energía eléctrica en Antioquia del modelo SARIMAX con ajuste de hiperparámetros con variable exógena de temperatura promedio.....61

Figura 27: Resultado de la predicción de la demanda de energía eléctrica en Antioquia del modelo Random Forest Autoregressive con ajuste de hiperparámetros con variable exógena del tipo de día. 62

Figura 28: Resultado de la predicción de la demanda de energía eléctrica en Antioquia del modelo RNN LSTM con ajuste de hiperparámetros con variable exógena del tipo de día. 63

Siglas, acrónimos y abreviaturas

Esp.	Especialista
IEEE	Institute of Electrical and Electronics Engineers
IDEAM	Instituto de Hidrología, Meteorología y Estudios Ambientales
MAE	Mean Absolute Error, Error medio absoluto
MAPE	Mean Absolute Percentage Error, Error porcentual absoluto medio
MSc	Magister Science
PhD	Philosophize Doctor
RMSE	Root Mean Absolute Error, Raíz del error cuadrático medio
UdeA	Universidad de Antioquia
XM	Administradores del mercado eléctrico

Resumen

La predicción de la demanda de energía eléctrica es esencial en el sector energético eléctrico y puede tener un impacto significativo en las decisiones operativas y estratégicas en la operación y despacho de los recursos de generación de electricidad de los sistemas de suministro de energía eléctrica, un adecuado pronóstico de la demanda de energía evita incurrir en sobre costos operativos considerables y disminuye el riesgo de desabastecimiento de la demanda de energía. Esto es particularmente crucial en Antioquia, el segundo departamento más poblado de Colombia y el segundo PIB más alto. Además, Antioquia posee la mayor capacidad de generación instalada en Colombia, con 5.789 MW, lo que equivale al 30% de la capacidad nacional y términos de demanda de energía eléctrica, Antioquia se ubica como la cuarta región más grande en el país, superada solo por la región caribe, la región centro y la región oriente.

En este documento se presenta una solución al problema de pronóstico de la demanda de la energía eléctrica para el departamento de Antioquia como parte importante del sector eléctrico colombiano, se desarrollan modelos de predicción de demanda de energía eléctrica horaria para la planeación de corto plazo, es decir, un horizonte de pronóstico de una semana considerando periodos de una hora. El informe se divide en secciones que abarcan la extracción y limpieza de datos, un análisis exploratorio de los mismos, la creación de modelos de predicción de demanda utilizando métodos estadísticos clásicos, métodos de aprendizaje automático (machine learning) autorregresivos y métodos de aprendizaje profundo (Deep learning) enfocados en series de tiempo, y concluye con las principales conclusiones.

Repositorio de GitHub [1]:
A.S. Torres Sánchez, especialización-analitica-monografia,
https://github.com/astorress/especializacion-analitica-monografia

Palabras clave: demanda de energía eléctrica, series temporales, machine learning, Deep learning.

Abstract

Predicting electricity demand is essential in the energy sector and can have a significant impact on operational and strategic decisions in the operation and dispatch of power generation resources of electric power supply systems. Adequate power demand forecasting avoids incurring considerable operational cost overruns and decreases the risk of power demand shortages. This is particularly crucial in Antioquia, Colombia's second most populous department and second highest GDP. In addition, Antioquia has the largest installed generation capacity in Colombia, with 5,789 MW, equivalent to 30% of the national capacity and in terms of electricity demand, Antioquia ranks as the fourth largest region in the country, surpassed only by the Caribbean region, the central region and the eastern region.

This paper presents a solution to the problem of electric power demand forecasting for the department of Antioquia as an important part of the Colombian electricity sector. Hourly electric power demand forecasting models are developed for short-term planning, i.e., a one-week forecast horizon considering one-hour periods. The report is divided into sections covering data extraction and cleaning, an exploratory analysis of the data, the creation of demand prediction models using classical statistical methods, autoregressive machine learning methods and deep learning methods focused on time series and concludes with the main conclusions.

*Keywords***:** electricity demand, time series, machine learning, deep learning.

1. Descripción del problema

El pronóstico de la demanda de energía eléctrica es un proceso fundamental para la toma de decisiones operativas y estratégicas en el mercado energético del país, cuya falta de exactitud puede generar altos costos económicos y riesgo de desabastecimiento de la demanda de electricidad [2]. Dichas predicciones pueden determinar de forma previa la carencia de energía eléctrica en el sistema y evitar de esta forma razonamientos de energía o desde otro punto de vista permite con anticipación comprometer plantas de generación de energía para cubrir la demanda de energía en una región o regiones del país. Antioquia, al ser el segundo departamento con más población del país, que según cifras el DANE [3] representa el 13,5% de la población total del país, además de representar el 5,44% del territorio nacional [4] y poseer el segundo PIB más alto del país [5], representa, es uno de los focos más importantes del sector eléctrico del país.

Antioquia es el departamento con mayor capacidad de generación instalada en Colombia, teniendo según datos analizados del API de XM [6], una capacidad instalada de 5.789 MW, que constituye alrededor del 30% de la capacidad total instalada del país.

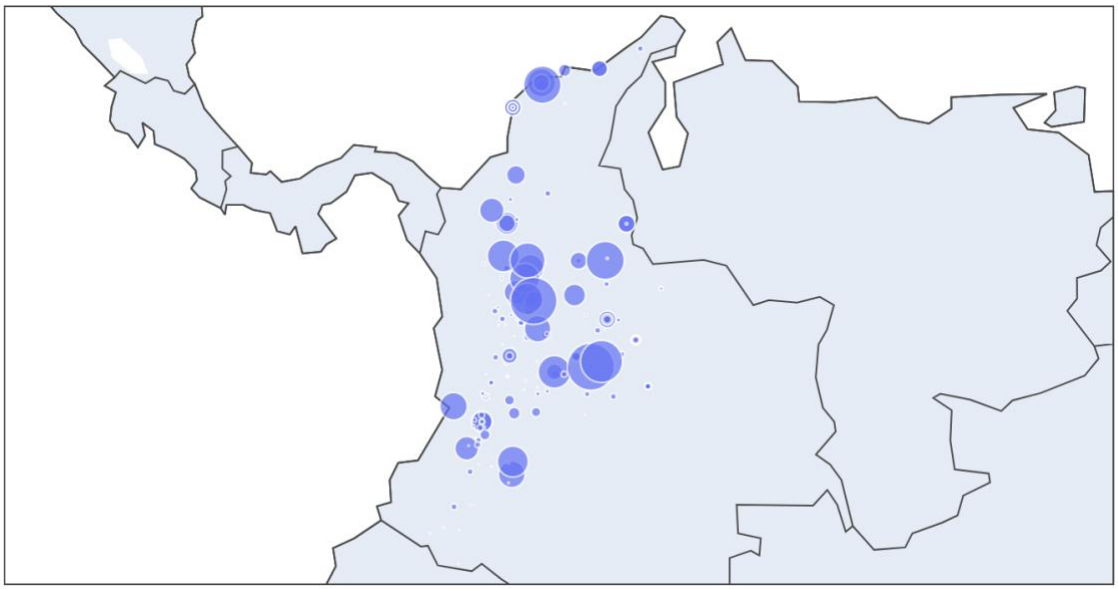


Figura 1: Mapa de la distribución de recursos de generación del país, donde el tamaño representa la capacidad en MW de los recursos de generación. Elaboración propia a partir de los datos del API de XM.

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

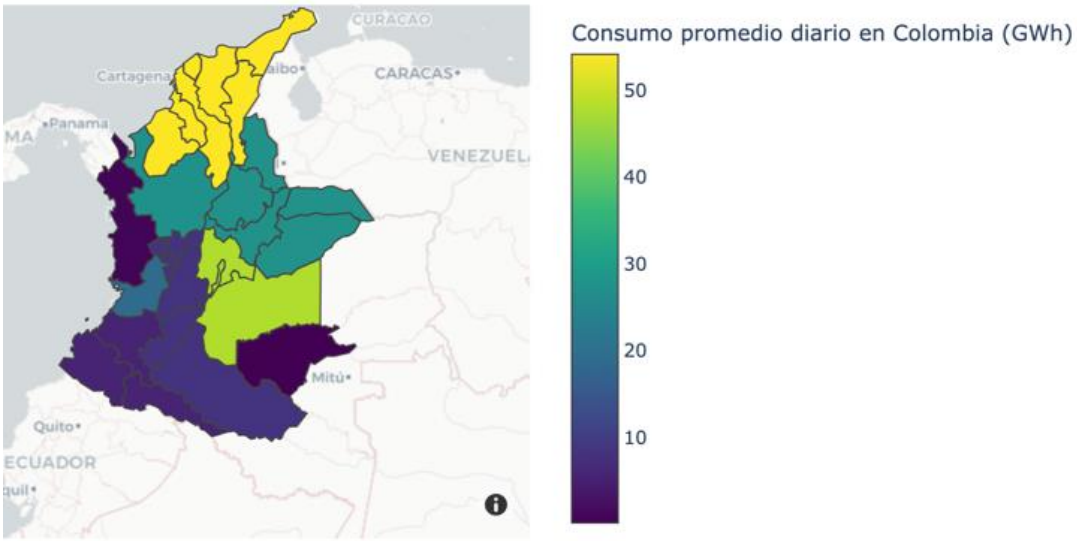


Figura 2: Mapa de calor según el consumo promedio diario de energía eléctrica en las regiones del país. Elaboración propia a partir de los datos del API de XM.

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

Debido a lo expuesto anteriormente, se puede evidenciar la gran importancia del departamento de Antioquia en el sector eléctrico del país y su gran relevancia en la elaboración de modelos de predicción de la demanda de energía eléctrica más exactos aumentando la confiabilidad, seguridad y operatividad del sistema eléctrico colombiano [7].

1.1. Problema de negocio

Para aumentar la exactitud de las predicciones de la demanda de energía eléctrica, es necesarios construir métodos de predicción que puedan obtener valores lo más cercanos a la realidad, dichos métodos de predicción deben ser capaces de estar en un rango del ± 5 % del valor real de la demanda de energía eléctrica para determinada hora. Por parte del operador regional y nacional del sistema eléctrico esto corresponde a un máximo valor aceptable que debe tener una predicción para la demanda de energía eléctrica; además desarrollar modelos de pronóstico capaces de realizar las predicciones con la posibilidad de utilizar otras variables exógenas que se relacionan con el comportamiento de la demanda de energía, con las cuales se busca disminuir los errores en los pronósticos de la demanda de energía eléctrica.

1.2. Aproximación desde la analítica de datos

Este trabajo se enfocará en los pronósticos con resolución horaria de la demanda de energía eléctrica en Antioquia con un horizonte de pronóstico de corto plazo, que corresponde a predecir una semana hacia el futuro; estos pronósticos de la demanda eléctrica se utilizan en la planeación operativa de corto plazo del despacho de las plantas de generación para garantizar su disponibilidad en el suministro de la electricidad que objetivo primordial en el mercado de energía eléctrica.

1.3. Origen de los datos

Los datos para este trabajo son obtenidos a través de dos fuentes, la primera a través del API de XM [6], en esta API se alojan los datos de todo el mercado de energía eléctrico colombiano y por supuesto la demanda de energía eléctrica; la segunda fuente de datos es del API de Socrata [8], en esta API de código abierto tiene acceso directo a los datos del repositorio abierto de datos del Gobierno Colombiano "Datos Abiertos" [9], acá se encuentran datos de las mediciones de temperatura realizadas por el IDEAM en todo el país.

1.4. Métricas de desempeño

Las métricas usadas comúnmente en la literatura para la predicción de series de tiempo son las siguientes:

- Error absoluto medio (Mean Absolute Error o MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (1)$$

  Donde, y_i es el valor real y $\hat{y}_i$ es el valor predicho.

- Raíz del error cuadrático medio (Root Mean Square Error o RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad (2)$$

  Donde, y_i es el valor real y $\hat{y}_i$ es el valor predicho.

- Error porcentual absoluto medio (Mean Absolute Percentage Error o MAPE):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \qquad (3)$$

  Donde, y_i es el valor real y $\hat{y}_i$ es el valor predicho.

2. Objetivos

2.1. Objetivo general

El objetivo general de este proyecto es desarrollar y comparar modelos de predicción de la demanda de energía eléctrica en Antioquia, usando métodos de inteligencia artificial adecuadas para el manejo de series temporales, obteniendo un mejor desempeño en la predicción que contribuya a una gestión eficiente y sostenible del suministro eléctrico.

2.2. Objetivos específicos

- Construir y preparar una base de datos históricos de la demanda de energía eléctrica en Antioquia, tomando los datos disponibles por el operador del mercado eléctrico Colombiano (XM) y complementando con datos climáticos de temperatura, para la estimación de la demanda de energía eléctrica usando técnicas de inteligencia artificial.
- Recopilar datos históricos de demanda de energía eléctrica en Antioquia, usando el API de XM, así como variables exógenas relevantes, como variables climatológicas que caracterizan las condiciones climáticas.
- Preprocesar y limpiar los datos para asegurar la calidad de estos, aplicando métodos de normalización, detección de atípicos y transformación según sea necesario.
- Realizar un análisis exploratorio de datos, con el fin de encontrar patrones, periodos cíclicos, periodos estacionarios, fluctuaciones, tendencias, entre otras.
- Revisar al menos tres técnicas de inteligencia artificial que sean más adecuadas para la predicción de la demanda de energía eléctrica, como redes neuronales profundas y herramientas de aprendizaje automático.
- Desarrollar e implementar modelos de predicción utilizando las técnicas de inteligencia artificial seleccionadas, ajustando los hiperparámetros de manera óptima.
- Evaluar y comparar el desempeño de los modelos de inteligencia artificial, utilizando métricas como el error medio absoluto (MAE), error porcentual absoluto medio (MAPE), el error cuadrado medio (RMSE), entre otros.

3. Marco Teórico

3.1. Autoregressive Integrated Moving Average (ARIMA)

Es un modelo estadístico ampliamente utilizado para pronósticos de series temporales, a continuación, se explica a detalle su funcionamiento [10]:

- Componente Autorregresivo (AR): Esta componente dentro del modelo establece que el valor actual de la serie de tiempo depende linealmente de valores anteriores, es decir, se modela como una regresión lineal de sus propios valores pasados. El parámetro que representa esta componente es "p" y se refiere a la cantidad de pasos de tiempos anteriores considerados en la regresión.

$$Y(t) = C + \varphi_1 * Y(t-1) + \varphi_2 * Y(t-2) + \cdots + \varphi_p * Y(t-p) + \varepsilon_t \quad (4)$$

  Donde:

  - $\varphi_1, \varphi_2, \ldots, \varphi_p$ son los coeficientes autoregresivos.
  - ε_t es el término de error en el tiempo t.
  - C es una constante.

- Componente de media móvil (MA): Esta componente dentro del modelo implica que el valor actual de la serie de tiempo depende de los errores pasados, es decir, de los residuos. El parámetro que representa esta componente es "q" y representa la cantidad de errores pasados considerados.

$$Y(t) = \mu - \theta_1 * \varepsilon_{t-1} - \theta_2 * \varepsilon_{t-2} - \cdots - \theta_q * \varepsilon_{t-q} + \varepsilon_t \quad (5)$$

  Donde:

  - $\theta_1, \theta_2, \ldots, \theta_q$ son los coeficientes de media móvil
  - ε_t es el término de error en el tiempo t.
  - μ es una constante.

- Componente diferenciación: Esta componente dentro del modelo se refiere a la diferenciación integrada, que es el proceso de calcular las diferencias entre valores sucesivos de la serie de tiempo. Este procedimiento se utiliza para lograr que la serie de tiempo sea estacionaria, es decir, con una media y varianza contante en el tiempo. El parámetro que representa esta componente es "d" y representa la cantidad de veces que sea diferencia la serie

$$Y'(t) = Y(t) - Y(t-1) \quad (6)$$

  Donde:

  - $Y'(t)$ es la serie diferenciada

El proceso de ajuste y selección de estos parámetros se realiza a través de técnicas como la función de autocorrelación parcial (PACF) y la función de autocorrelación (ACF).

3.2. Seasonal Autoregressive Integrated Moving Average with eXogenous regressors (SARIMAX)

Es un modelo estadístico ampliamente utilizado para pronósticos de series temporales, combina componentes de series de tiempo estacionarias y no estacionarias, junto con variables exógenas para lograr una presentación completade la dinámica de la serie temporal, a continuación, se explica a detalle su funcionamiento [10]:

- Componente Autorregresivo (AR): Esta componente dentro del modelo establece que el valor actual de la serie de tiempo depende linealmente de valores anteriores, es decir, se modela como una regresión lineal de sus propios valores pasados. El parámetro que representa esta componente es "p" y se refiere a la cantidad de pasos de tiempos anteriores considerados en la regresión.

$$Y(t) = C + \varphi_1 * Y(t-1) + \varphi_2 * Y(t-2) + \cdots + \varphi_p * Y(t-p) + \varepsilon_t \qquad (7)$$

  Donde:

  - $\varphi_1, \varphi_2, \ldots, \varphi_p$ son los coeficientes autoregresivos.
  - ε_t es el término de error en el tiempo t.
  - C es una constante.

- Componente de media móvil (MA): Esta componente dentro del modelo implica que el valor actual de la serie de tiempo depende de los errores pasados, es decir, de los residuos. El parámetro que representa esta componente es "q" y representa la cantidad de errores pasados considerados.

$$Y(t) = \mu - \theta_1 * \varepsilon_{t-1} - \theta_2 * \varepsilon_{t-2} - \cdots - \theta_q * \varepsilon_{t-q} + \varepsilon_t \qquad (8)$$

  Donde:

  - θ_1, θ_2, …, θ_q son los coeficientes de media móvil
  - ε_t es el término de error en el tiempo t.
  - μ es una constante.

- Componente diferenciación: Esta componente dentro del modelo se refiere a la diferenciación integrada, que es el proceso de calcular las diferencias entre valores sucesivos de la serie de tiempo. Este procedimiento se utiliza para lograr que la serie de tiempo sea estacionaria, es decir, con una media y varianza contante en el tiempo. El parámetro que representa esta componente es "d" y representa la cantidad de veces que sea diferencia la serie.

$$Y'(t) = Y(t) - Y(t-1) \qquad (9)$$

Donde:

- $Y'(t)$ es la serie diferenciada

- Componente Seasonal (S): Esta componente dentro del modelo se refiere a la estacionalidad de los datos de la serie de tiempo, en un modelo SARIMAX, se identifica la estacionalidad de los datos y se modela utilizando parámetros estacionales como el autorregresivo (AR) estacional y componente de media móvil (MA) estacional, lo cual permite capturar patrones recurrentes en los datos a lo largo del tiempo.

$$Y'(t) = Y'(t-s) + \varepsilon_t \qquad (10)$$

Donde:

- s es la estacionalidad.

- Variable Exógena (X): El "X" en SARIMAX indica que se pueden incorporar variables exógenas, es decir, variables adicionales que no son parte de la serie de tiempo principal, pero que pueden influir en la serie de tiempo principal.

El modelo SARIMAX implica la estimación de los parámetros "p", "d" y "q" de los componentes AR, MA y de integración, respectivamente, además, se deben identificar los parámetros estaciones "P", "D" y "Q" para las componentes estaciones de AR, MA y de integración. El proceso de ajuste y selección de estos parámetros se realiza a través de técnicas como la función de autocorrelación parcial (PACF) y la función de autocorrelación (ACF).

3.3. Linear regressor autoregressive

Un modelo de machine learning autorregresivo como el linear regressor, es una herramienta fundamental para el análisis de series de tiempo, su funcionamiento se basa principalmente en la componente AR de modelos como ARIMA o SARIMAX, es decir, se basa en que el valor actual de la serie de tiempo depende linealmente de valores anteriores de la propia seria de tiempo que pueden proporcionar información útil para predecir el valor futuro.

$$Y(t) = C + \varphi_1 * Y(t-1) + \varphi_2 * Y(t-2) + \cdots + \varphi_p * Y(t-p) + \varepsilon_t \qquad (11)$$

3.4. Random Forest autoregressive

Un modelo de Random Forest autorregresivo para series de tiempo es una técnica de aprendizaje automático que combina los principios de Random Forest, una técnica de ensamble, con la modelización autorregresiva para abordar problemas de series temporales. El modelo Random Forest autorregresivo combina varios elementos. En primer lugar, utiliza la idea de un Random Forest, que es un ensamble de árboles de decisión. Cada árbol de decisión se entrena con una muestra aleatoria de los datos y toma decisiones basadas en características específicas. Esto permite que el modelo capture la complejidad y la no linealidad de los datos.

En el contexto de series de tiempo, cada árbol de decisión en el Random Forest autorregresivo se entrena para predecir un valor futuro basado en observaciones pasadas, utilizando técnicas autorregresivas, como ARIMA o SARIMA, como base. Luego, el conjunto de árboles de decisión trabaja en paralelo para predecir múltiples pasos en el futuro. Esto permite que el modelo maneje múltiples horizontes de predicción y capture las tendencias y estacionalidades presentes en la serie de tiempo.

La ventaja clave de un Random Forest autorregresivo es su capacidad para lidiar con la no estacionalidad y la complejidad de las series de tiempo, ya que cada árbol se entrena de forma independiente y puede abordar diferentes patrones y comportamientos. Además, el ensemble reduce el riesgo de sobreajuste y mejora la generalización del modelo [11].

3.5. Redes neuronales recurrentes (RNN)

Las redes neuronales son modelos matemáticos inspirados en el funcionamiento del cerebro humano, diseñados para aprender y reconocer patrones en datos. Consisten en capas de neuronas interconectadas que procesan información. En el contexto de series de tiempo, las redes neuronales se utilizan para realizar tareas como predicción, clasificación y generación. Dentro de las redes neuronales existen varios tipos de redes, entre ellas se encuentras las redes neuronales recurrentes (RNN) que son un tipo especializado de red neuronal que se emplea en secuencias temporales. A diferencia de las redes neuronales feedforward, las RNN tienen conexiones retroalimentadas, lo que les permite mantener una memoria a corto plazo de estados anteriores. Esta memoria es especialmente útil para capturar dependencias temporales en series de tiempo. A continuación, se muestra la arquitectura de una RNN [12].

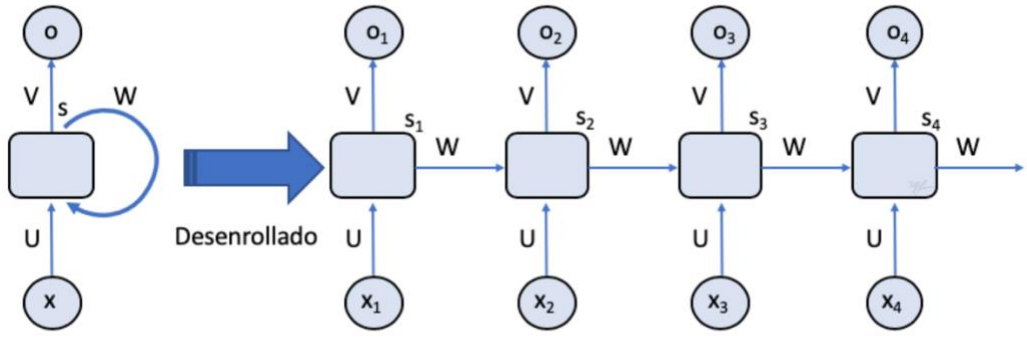


Figura 3: Arquitectura de red neuronales recurrente [12].

La Figura 3 muestra a la izquierda la arquitectura a una red neuronal recurrente y la derecha la arquitectura de la misma red neuronales expandida. La expansión de la RNN se extenderá tanto como el número de datos en el vector de la serie de tiempo x_t.

Donde:

- x_t es el vector de entrada en el tiempo t.
- s_t es un estado oculto de la red en el tiempo t. $s_t = \varphi_1(Ux_t + Ws_{t-1})$
- O_t es la salida de la neurona en el tiempo t. $O_t = \varphi_2(Vs_t)$
- φ_1 es una función de activación

Las redes neuronales LSTM (Long Short-Term Memory) son una variante avanzada de las RNN que abordan uno de sus problemas fundamentales: la desaparición del gradiente. Las LSTM tienen una arquitectura especializada con compuertas que controlan el flujo de información y permiten que la red aprenda dependencias tanto de corto como de largo plazo en series temporales, ver Figura 4. Esto las hace muy efectivas en aplicaciones donde las dependencias temporales son cruciales, como el análisis de series de tiempo financieras o la traducción automática [12].

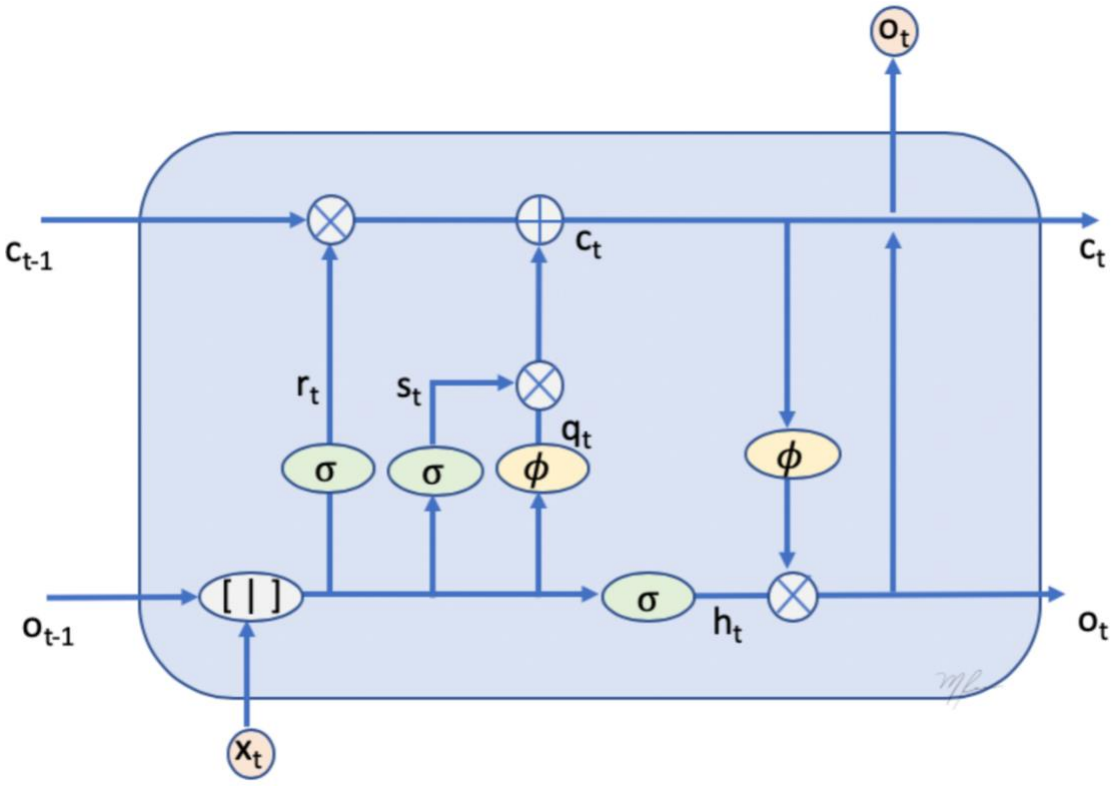


Figura 4: Arquitectura de red neuronales recurrente LSTM [12].

Donde:

- x_t es el vector de entrada en el tiempo t.
- C_t es la información (memoria) que es pasada a la siguiente etapa.
- O_t es la salida de la neurona en el tiempo t.

- C_{t-1} es la información (memoria) que es pasada por la etapa anterior en el tiempo t-1.
- O_{t-1} es la salida de la neurona en el tiempo t-1.
- $\emptyset$ son las funciones de activación.
- [||] es la operación que denota como la concatenación de las entradas.
- $\oplus$ es el operador que denota la suma de vectores.
- σ son signoides que aporximan a una respuesta binaria, su operación en combinación con el producto $\otimes$ pueden entenderse como switches que permiten el flujo, o no, de la información

4. Metodología

La metodología para la elaboración de este trabajo consta de 7 etapas que son ilustradas en la Figura 5.



Figura 5: Diagrama de la metodología general. Elaboración propia.

4.1. Recopilación de datos:

En esta etapa se recopilarán los datos históricos horarios de la demanda de energía eléctrica en Antioquia, además, como uno de los objetivos es complementar las predicciones a realizar con datos climáticos, también se recopilarán datos históricos horarios de la temperatura en Antioquia.

4.2. Preprocesamiento de datos:

En esta etapa se realizarán la limpieza de los datos, tanto de la demanda horaria de energía eléctrica en Antioquia, como de los datos horarios de la temperatura en Antioquia, para ello, se aplicarán métodos de normalización, detección de atípicos, imputación de datos, creación de nuevas columnas a partir de las ya existentes, entre otras transformaciones a la estructura de los datos según sea necesario.

4.3. Análisis de datos:

En esta etapa se realizará un análisis de datos, enfocado en series de tiempo, donde se realizará la búsqueda de patrones y perfiles típicos de la demanda de energía eléctrica, tendencias, periodos cíclicos, periodos estacionarios, fluctuaciones, entre otros.

4.4. Búsqueda de modelos:

En esta etapa se realizará una búsqueda en la literatura de cuáles son los métodos de la estadística clásica, machine learning y deep learning, más adecuados para las series de tiempo o datos con características semejantes a los datos que se tienen.

4.5. Desarrollo de los modelos:

En esta etapa de aplicarán los modelos obtenidos en la etapa anterior a los datos que se tienen. Para esta etapa se estipula realizar alrededor de 4 iteraciones para cada uno de los modelos, utilizando diversas variables exógenas que faciliten la predicción de la demanda de energía eléctrica horaria en el corto plazo de una semana.

4.6. Evaluación y comparación:

En esta etapa se evaluará la eficiencia y precisión de los resultados de los modelos, aplicando diversas métricas de validación, con el objetivo de encontrar el mejor modelo para la predicción de la demanda de energía eléctrica horaria en el corto plazo de una semana.

4.7. Optimización y ajuste:

En esta etapa se ajustarán y optimizarán los hiperparámetros y parámetros de los modelos utilizados y a cada una de sus iteraciones, con el objetivo de mejora aún más las predicciones de la

demanda de energía eléctrica, aplicando las mismas métricas de validación de la etapa anterior. Obteniendo así el mejor modelo para la predicción de la demanda de energía eléctrica.

5. Datos

5.1. Extracción de dato originales

Los datos para este trabajo, como ya se mencionó anteriormente son obtenidos a través de dos fuentes, la primera a través del API de XM [6] y la segunda fuente de datos es del API de Socrata [8].

5.1.1. API XM:

En esta API se alojan múltiples variables del mercado eléctrico colombiano, tales como, el precio de bolsa, volumen de los embalses, demanda de comercial de energía eléctrica, precio de ofertas de energía, listados de agentes, listado de embalses, entre otros, en total se disponen de más de 140 variables del mercado eléctrico colombiano.

Para este caso, la variable a utilizar es la demanda comercial de energía eléctrica, esta variable presenta datos con granularidad horaria, además de un código, que permite identificar de forma sencilla cual es el agente del mercado al cual corresponde dicha demanda comercial, a continuación, se presenta un listado con las columnas del dataset de la variable.

- Id: Corresponde a la identificación del tipo de variables del mercado eléctrico colombiano, este puede tener valor de 'Agente' o 'Sistema', dependiendo de si la variable es dada por el sistema general o por un agente especifico.
- Values_code: Código de identificación del agente y su etapa productiva dentro mercado de energía eléctrica, este código consta de cuatro letras y se crea de la siguiente forma:
    - Las primeras 3 letras en mayúscula, corresponde al identificador único del agente, por ejemplo, Empresas Públicas de Medellín o 'EPM', Empresa de energía de Cundinamarca 'EEC'.
    - Las ultima letra corresponde a la etapa productiva del agente, teniendo así, 'G' para generación, 'T' para transmisión, 'D' para distribución y 'C' para comercializador.

  Teniendo así que 'EPMC' corresponde a Empresas Públicas de Medellín como comercializador
- Values_Hour01: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 1 del día.
- Values_Hour02: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 2 del día.
- Values_Hour03: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 3 del día.
- Values_Hour04: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 4 del día.
- Values_Hour05: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 5 del día.
- Values_Hour06: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 6 del día.
- Values_Hour07: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 7 del día.

- Values_Hour08: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 8 del día.
- Values_Hour09: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 9 del día.
- Values_Hour10: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 10 del día.
- Values_Hour11: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 11 del día.
- Values_Hour12: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 12 del día.
- Values_Hour13: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 13 del día.
- Values_Hour14: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 14 del día.
- Values_Hour15: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 15 del día.
- Values_Hour16: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 16 del día.
- Values_Hour17: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 17 del día.
- Values_Hour18: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 18 del día.
- Values_Hour19: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 19 del día.
- Values_Hour20: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 20 del día.
- Values_Hour21: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 21 del día.
- Values_Hour22: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 22 del día.
- Values_Hour23: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 23 del día.
- Values_Hour24: Corresponde al valor de la demanda comercial de energía eléctrica en kWh correspondiente a la hora 24 del día.
- Date: Corresponde a la fecha de la demanda comercial de energía eléctrica en formato ‘yyyy-mm-dd’.

Nota: Para este caso la hora 1, corresponde las 00:00h o 12:00am y la hora 24 corresponde a las 23:00h o 11:00pm

Como se pudo observar, no existe como una forma directa de obtener los datos de la demanda de energía eléctrica de Antioquia, por ello, se realizó la suposición de que la demanda comercial del comercializador EPMC está fuertemente influenciado por el departamento de Antioquia, debido a que este es el comercializador natural del departamento, además se extrajeron los datos desde el 1 de enero del 2015 hasta el 1 de noviembre del 2023, obteniendo así un total de 3.224 registros y 27 columnas.

A continuación, se presenta la estructura del dataset.

	Id	Values_code	Values_Hour01	Values_Hour02	Values_Hour03	...	Values_Hour21	Values_Hour22	Values_Hour23	Values_Hour24	Date
0	Agente	EPMC	874902.37	811130.06	768748.60	...	997156.93	939934.82	846098.10	737264.97	2015-01-01
1	Agente	EPMC	699781.13	709869.54	696670.76	...	1237953.79	1130694.38	1027821.94	929550.82	2015-01-02
2	Agente	EPMC	845367.77	806900.27	777052.40	...	1231779.11	1108847.83	1001956.93	973063.72	2015-01-03
3	Agente	EPMC	848403.10	804607.94	774110.78	...	1155918.51	1060257.60	962598.22	830725.98	2015-01-04
4	Agente	EPMC	811468.23	795797.35	753654.14	...	1303091.56	1192559.28	1060380.63	988940.56	2015-01-05
...	...	...	...	...	...	...	...	...	...	...	...
3178	Agente	EPMC	1229609.81	1185567.00	1145843.72	...	1636027.36	1522228.84	1391455.95	1290731.83	2023-09-14
3179	Agente	EPMC	1199396.09	1167002.89	1138902.15	...	1596169.60	1498153.92	1396988.50	1292183.28	2023-09-15
3180	Agente	EPMC	1214220.79	1172907.93	1140582.42	...	1481850.54	1390996.57	1302452.12	1215020.07	2023-09-16
3181	Agente	EPMC	1149436.04	1098979.94	1053502.06	...	1354029.63	1298861.37	1223416.92	1166799.61	2023-09-17
3182	Agente	EPMC	1093388.11	1062575.11	1044373.05	...	1590831.02	1465352.44	1338550.97	1246540.76	2023-09-18

Figura 6: Estructura del dataset de la demanda de energía eléctrica en Antioquia.

5.1.2. API Socrata:

En esta API se tiene acceso directo a diversos repositorios de datos abierto de todo el mundo, incluyendo el repositorio de datos abiertos del Gobierno Colombia “Datos Abiertos”, a través de esta API se puede acceder a todos los datos alojados en este repositorio, tales como, informes de procesos judiciales de algunas ciudades del país, datos sobre la exploración minera, límites geográficos del país, mediciones de temperatura, casos de cáncer en el país, entre otros, en total se disponen de alrededor de 7974 de bases de datos de alrededor de 1.443 entidades públicas.

Para este caso, la base datos a utilizar corresponde a las mediciones de temperatura realizadas por el IDEAM, la cual se encuentra con una granularidad cada 5 minutos, el código de identificación de las estaciones de medición alrededor del país y los municipios donde se encuentran dichas estaciones, a continuación, se presenta un listado con las columnas del dataset de la variable.

- codigoestacion: Corresponde al código de identificación de las estaciones medición de temperatura alrededor del país.
- fechaobservacion: Corresponde a la fecha y hora de la medición de temperatura en formato ISO 86601 [12], este formato se utiliza para representar fechas y horas de manera estandarizada, por el ejemplo, el valor “2015-06-02T19:05:59.111” representa la fecha del 2 de junio del 2015 a las 7:05pm con 59 segundos y 111 milisegundos.
- valorobservado: Corresponde a la medición de temperatura obtenido por las estaciones de medición en ℃.
- municipio: Corresponde al municipio donde se encuentra la estación de medición,

Esta base de datos si bien consta de una grande cantidad de información sobre las temperaturas en Colombia, no cuenta con datos de todos los municipio del país, ya que no en todos hay estaciones de medición, por lo cual, se seleccionaron 4 municipio de Antioquia presentes dentro de la base de datos para la obtención del valor de temperatura en Antioquia, los cuales son Medellín, Bello, Rionegro y Santuario, además se extrajo la base de datos desde el 1 de enero el 2015 hasta el 1 noviembre del 2023, obteniendo así un total 74.684 registros y 4 columnas.

A continuación, se presenta la imagen de la estructura del dataset.

	codigoestacion	fechaobservacion	valorobservado	municipio
0	0023085260	2015-01-16 01:00:00	13.7	RIONEGRO
1	0027015310	2015-02-12 23:00:00	20.2	BELLO
2	0027015330	2015-03-25 23:00:00	21.0	MEDELLÍN
3	0023085260	2015-03-02 16:00:00	23.6	RIONEGRO
4	0023085270	2015-03-12 04:00:00	16.1	RIONEGRO
...	...	...	...	...
278208	0027015330	2023-10-10 23:54:00	21.2	MEDELLÍN
278209	0023085270	2023-10-10 23:56:00	16.5	RIONEGRO
278210	0027015330	2023-10-10 23:56:00	21.2	MEDELLÍN
278211	0023085270	2023-10-10 23:58:00	16.5	RIONEGRO
278212	0027015330	2023-10-10 23:58:00	21.1	MEDELLÍN

Figura 7: Estructura del dataset de la temperatura en Antioquia.

5.2. Preprocesamiento

A continuación, se mostrarán los procesos más relevantes en el preprocesamiento de datos para ambos dataset explicados anteriormente.

5.2.1. Demanda de energía eléctrica en Antioquia:

Como se explicó anteriormente, estos datos fueron extraídos a través del API de XM, y como se pude observar, el dataset que arroja dicha API tiene un formato pelicular, ya que normalmente las bases de datos y en especial las bases de datos de series temporales, debería crecer hacia abajo y no hacia los lados como estas, es decir, los registros nuevos de datos de demanda horaria de energía eléctrica deberían ser almacenado como filas y no como columnas, algo similar a como está la Figura 7. Por esto se debido transformar el dataset original, para que tenga una estructura adecuada, similar a la de la temperatura, haciendo así el dataset mucho más manejable.

Para esto, se utiliza un proceso de procesamiento llamado unpivot [14] el cual consiste en convertir valores ubicados en las columnas de una tabla en filas, haciendo así que las tablas o dataset aumente sus números de registros hacia abajo.

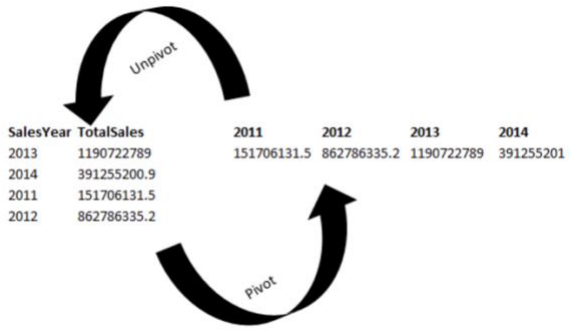


Figura 8: Ejemplo del proceso de pivot y unpivot [14].

Haciendo este proceso y aplicando algunas transformaciones adicionales el número de registros aumenta y disminuye el número de columnas, dando así, se obtiene el siguiente dataset con 77.376 registros y 2 columnas.

	fecha_hora	consumo
0	2015-01-01 00:00:00	874902.37
1	2015-01-01 01:00:00	811130.06
2	2015-01-01 02:00:00	768748.60
3	2015-01-01 03:00:00	744469.99
4	2015-01-01 04:00:00	735272.46
...	...	...
76387	2023-09-18 19:00:00	1651808.39
76388	2023-09-18 20:00:00	1590831.02
76389	2023-09-18 21:00:00	1465352.44
76390	2023-09-18 22:00:00	1338550.97
76391	2023-09-18 23:00:00	1246540.76

Figura 9: Estructura del dataset de la demanda de energía eléctrica en Antioquia luego de la transformación unpivot.

5.2.2. Temperatura:

Como se mencionó anteriormente, no existen datos de temperatura para todos los municipios de Antioquia, razón por la cual se decidió tomar solo 4 municipio, estos municipio fueron seleccionados debido se encuentran dentro de las 2 subregiones más importantes para el departamento a nivel de demanda, las cuales son el valle de Aburrá (Medellín y Bello) y el valle de San Nicolas o subregión Oriente (Rionegro y Santuario), estas dos subregiones representan la mayor parte de la población residencial e industrial del departamento, lo cual, hace que sean un muestra muy representativa de la variable de temperatura en Antioquia.

No obstante, debido a esta suposición y también, debido a que esta variable posee una mayor granularidad que los datos de demanda de energía eléctrica, decidió obtener 3 valores distintos de temperatura horarios, "temperatura_prom", "temperatura_mediana" y "temperatura_max", agrupando los datos que se tenían cada 5 minutos en datos horarias de los valores mencionados, esto con el objetivo de tener distintos valores de temperatura con los cuales poder realizar las predicciones, haciendo que exista una mayor diversidad en los datos, reduciendo de esta forma el error incurrido por las suposiciones realizadas, además de esto, debido a que ya no es necesario, ni el "codigoestacion" ni "municipio" se decide eliminar dichas columnas del dataset, obteniendo así el siguiente dataset con 77.376 registros y 4 columnas.

	fecha	temperatura_mediana	temperatura_prom	temperatura_max
0	2015-01-01 00:00:00	20.800	19.250000	21.3
1	2015-01-01 01:00:00	20.700	18.900000	20.8
2	2015-01-01 02:00:00	20.400	18.683333	20.5
3	2015-01-01 03:00:00	20.200	18.566667	20.4
4	2015-01-01 04:00:00	19.700	18.216667	19.9
...	...	...	...	...
74186	2023-10-10 19:00:00	20.150	20.148333	23.2
74187	2023-10-10 20:00:00	19.350	19.431667	22.9
74188	2023-10-10 21:00:00	19.150	19.160476	21.8
74189	2023-10-10 22:00:00	19.125	19.103889	21.7
74190	2023-10-10 23:00:00	18.900	18.873218	21.5

Figura 10: Estructura del dataset de temperatura en Antioquia luego de las transformaciones.

5.2.3. Combinación de datasets:

Con las transformaciones anteriores realizadas a ambos datasets, se procede a realizar la combinación de ambos para así tener una base de datos unificada. Además de esto y debido a la importancia de características asociadas a las fechas en la demanda de energía eléctrica, se crean columnas adicionales como lo son:

- hora: Corresponde al valor numérico de la hora, va desde las 0 (12:00am) hasta 24 (11:00pm).
- mes: Corresponde al valor numérico del mes, va desde 1 (enero) hasta 12 (diciembre).
- fin_de_semana: Corresponde a un valor binario que indica si es el día fue fin de semana (1) o no (0).
- festivo: Corresponde a un valor binario que indica si es el día fue festivo (1) o no (0).

- tipo_dia: Corresponde a un valor categórico numérico que indica si el día es un día de lunes a sábado (1), domingo (2) o si es festivo (3).
- "tipo_semana": Corresponde a un valor categórico numérico que indica si la semana es semana sin festivo (1), semana con festivo (2) o semana santa (3).

	fecha_hora	consumo	hora	mes	dia_semana	fin_de_semana	festivos	temperatura_mediana	temperatura_prom	temperatura_max	tipo_dia	tipo_semana
0	2015-01-01 00:00:00	874902.37	0	1	4	0	1	20.800	19.250000	21.3	1	2
1	2015-01-01 01:00:00	811130.06	1	1	4	0	1	20.700	18.900000	20.8	1	2
2	2015-01-01 02:00:00	768748.60	2	1	4	0	1	20.400	18.683333	20.5	1	2
3	2015-01-01 03:00:00	744469.99	3	1	4	0	1	20.200	18.566667	20.4	1	2
4	2015-01-01 04:00:00	735272.46	4	1	4	0	1	19.700	18.216667	19.9	1	2
...	...	...	...	...	...	...	...	...	...	...	...	...
76387	2023-09-18 19:00:00	1651808.39	19	9	1	0	0	21.375	21.440000	25.7	1	1
76388	2023-09-18 20:00:00	1590831.02	20	9	1	0	0	20.500	20.151667	24.3	1	1
76389	2023-09-18 21:00:00	1465352.44	21	9	1	0	0	19.650	19.558750	23.3	1	1
76390	2023-09-18 22:00:00	1338550.97	22	9	1	0	0	18.850	18.938333	22.7	1	1
76391	2023-09-18 23:00:00	1246540.76	23	9	1	0	0	18.750	18.665000	22.0	1	1

Figura 11: Estructura del dataset final.

5.3. Limpieza de datos

En el proceso de limpieza de datos, se observó que si bien algunos algoritmos de detección de atípicos y gráficos como el de bigotes arrojaban que existían datos atípicos en algunas variables, en especial en las de temperatura, estos datos son lo que se conocen como atípicos naturales, es decir, datos que se salen de los patrones o tendencias típicas de los datos eran atípicos, pero que no eran fallas de los sensores por ejemplo, ya que, los valores se encontraban dentro de los rangos permitidos y eliminarlos hacia que se pierda variabilidad en los datos. También, se evidencia datos faltantes en las variables de temperatura, alrededor de un 3,5% de registros del dataset final, por esto, se decidido realizar una imputación de datos con el algoritmo KNN, con K=5.

5.4. Análisis de datos

Con los datos ya procesados y limpios, se procede a realizar un análisis de datos, enfocados en series de tiempo, el objetivo de este análisis es encontrar patrones, tendencias, periodos, fluctuaciones, entre otros, en los datos, estas características en los datos ayudarán mejor a comprender los comportamientos que no pueden observarse a simple vista, lo cual, hará mucho más fácil la selección, análisis de desarrollo de los modelos más precisos que permitan predecir de manera efectiva la demanda de energía eléctrica en Antioquia. A continuación, se mostrarán los resultados más relevantes del análisis de datos para el dataset final mostrado anteriormente.

5.4.1. Distribución de los datos

Comprender la distribución de los datos, es un punto de partida muy útil para comenzar a comprender los compartimientos y patrones ocultos en los datos, en este caso se utilizó gráficos de bigotes para poder de forma rápida y sencilla visualizar dichas distribuciones. A continuación, se presenta algunos de estos gráficos realizados.

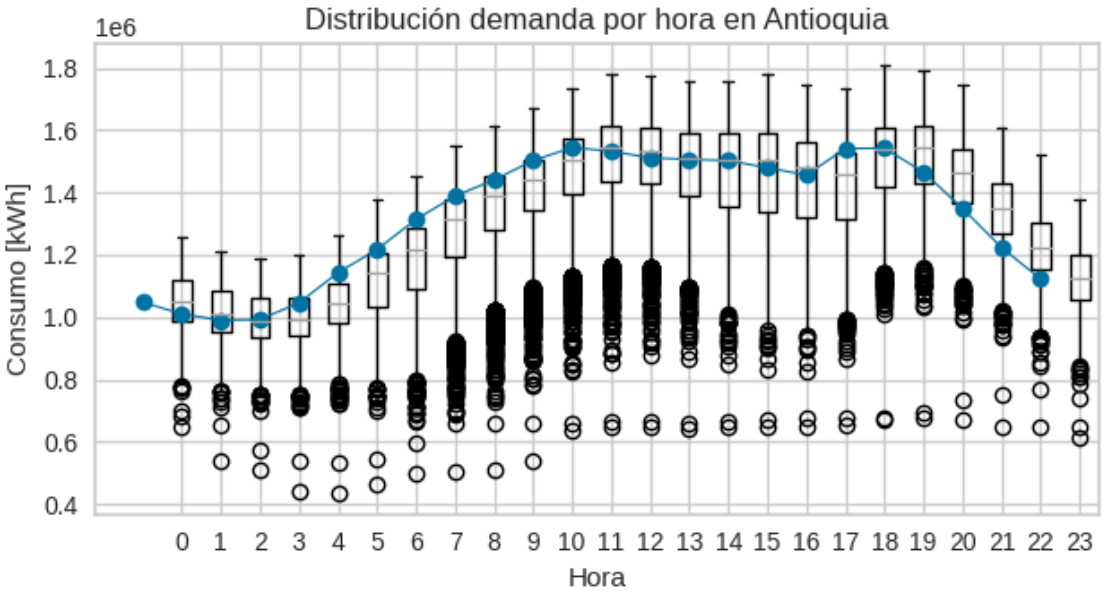


Figura 12: Perfil de distribución de la demanda de energía eléctrica por hora.

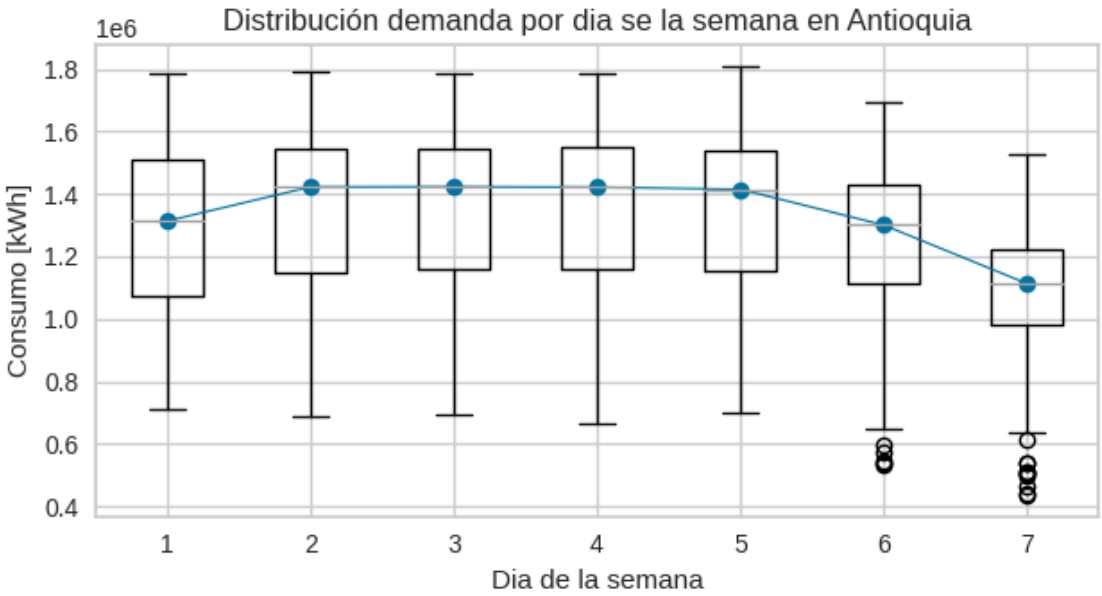


Figura 13: Perfil de distribución de la demanda de energía eléctrica por día de la semana.

De las Figura 12 y la Figura 13 al prestar atención en los valores promedio de cada hora, de puede notar que existe un perfil típico de demanda de energía diaria, en dicho perfil, existen valores de alta demanda alrededor del medio de día y al finalizar la tarde alrededor de las 6:00pm, además existen valores de baja demanda las 9:00 pm y las 4 am.

5.4.2. Descomposición estacional

La descomposición estacional [15], es un análisis muy útil para las series de tiempo, este análisis consiste en describir una serie de tiempo en tres componentes que describen cada una un comportamiento de esta, esta presentación de la serie puede ser de forma aditiva (es decir, mediante sumas) o multiplicativa (es decir, mediante multiplicaciones).

$$x_t = T_t + S_t + R_t \qquad (12)$$

$$x_t = T_t * S_t * R_t \qquad (13)$$

Donde, x_t es la serie de tiempo, T_t es la tendencia de la serie de tiempo, S_t es la estacionalidad de la serie de tiempo y R_t residuo.

Con esta esta descomposición se pueden encontrar información dentro de la serie que no puede notar a simple vista, para ilustrar esto, se muestra a continuación, un gráfico de la serie de tiempo completa de la demanda de energía eléctrica en Antioquia.

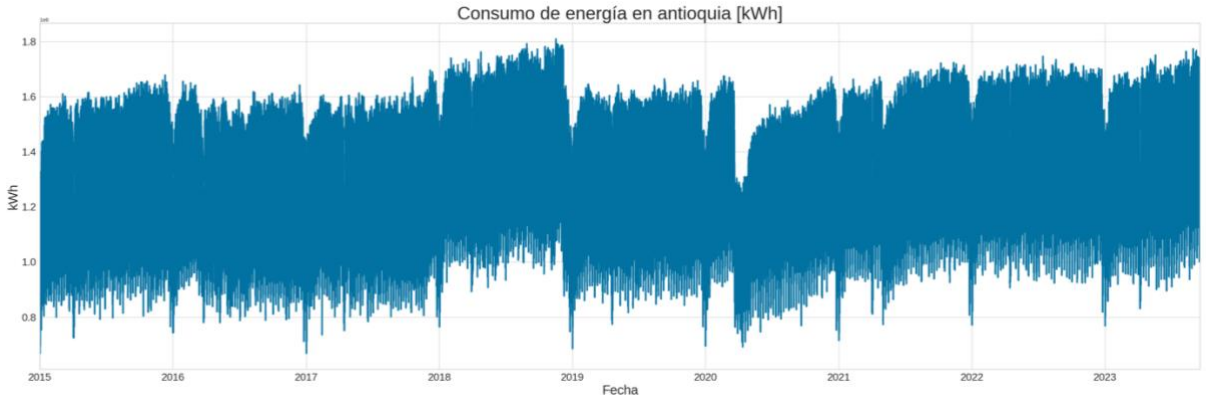


Figura 14: Grafico de visualización de la demanda de energía eléctrica en Antioquia de todo el dataset final.

Como se puede observar en la Figura 14, es difícil encontrar patrones a simple vista en la serie de tiempo, solo pudiéndose observar picos de alto y bajo demanda en algunos periodos de tiempo. Debido a esta dificultad de encontrar dichos patrones, se evidencia la necesidad de realizar la descomposición estacional de la serie de tiempo. A continuación, en la Figura 15 se muestra la descomposición estacional aditiva para la demanda de energía eléctrica en Antioquia.

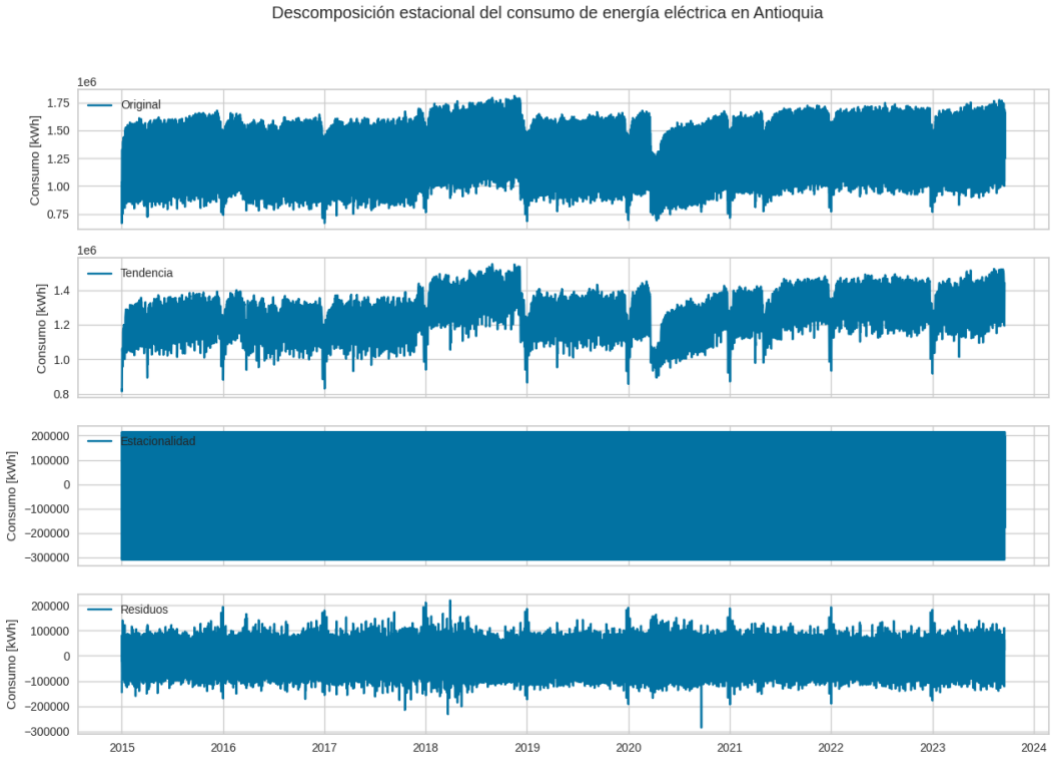


Figura 15: Descomposición estacional de la demanda de energía eléctrica en Antioquia para todo el dataset.

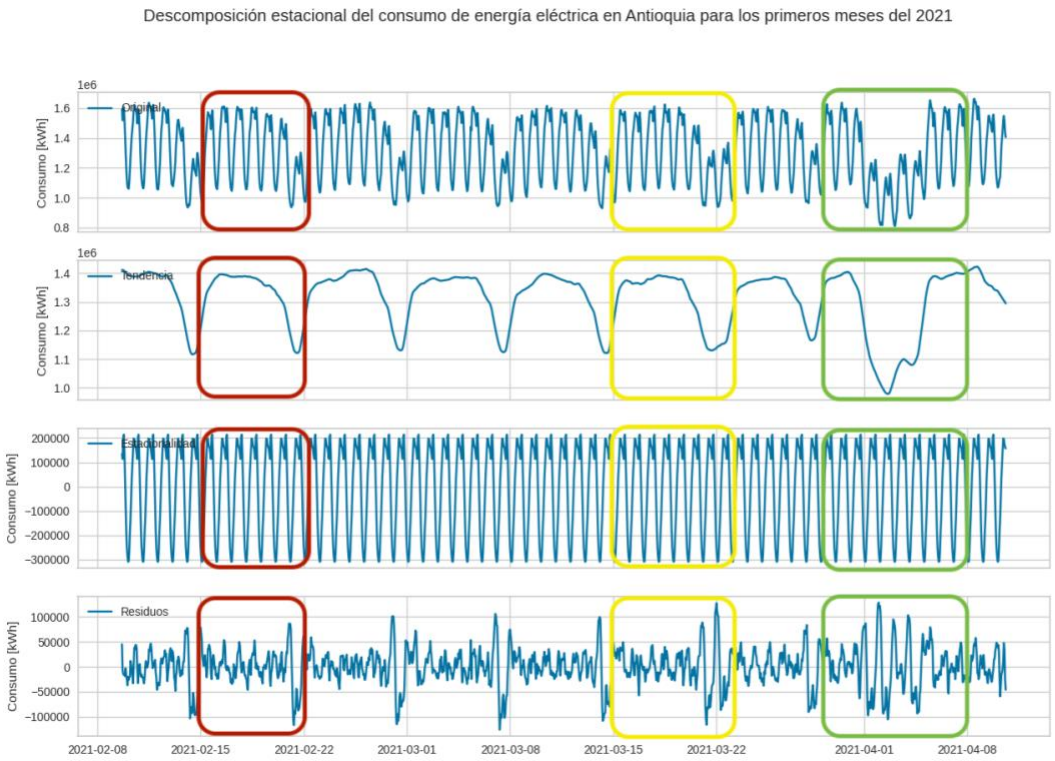


Figura 16: Descomposición estacional de la demanda de energía eléctrica en Antioquia para los primeros meses del año del 2021.

La Figura 16 consta de cuatro sub gráficos, el primero a la serie de tiempo original, el segundo corresponde a la tendencia de la serie de tiempo, el tercero a la estacionalidad o ciclos de la serie de tiempo y cuarto al residuo. De modo general en la Figura 16 se puede observar que la estacionalidad de los datos es diaria, es decir, los patrones de demanda de energía eléctrica se repiten cada día, además, la tendencia de los datos en si misma también posee cierta estacionalidad, por esto al observar dicha componente se puede notar que cada cierto intervalo de tiempo se repite la tendencia, esta estacionalidad en la tendencia es semanal, teniendo los valores más alto en mitad de semana y descayendo hasta el domingo, en el cual se presenta la tendencia más baja de demanda de energía eléctrica.

Ahora, haciendo énfasis en los elementos visuales para facilitar el análisis de la Figura 16, se puede notar que existen tres momentos relevantes en los gráficos, el primero momento correspondiente el rectángulo rojo, corresponde a periodos de demanda de energía eléctrica de semanas normales, es decir sin festivos, en la cual los valores alto de demanda de encuentran en mitad de semana y los valores más bajos al finalizar la semana, es decir, los domingos. Este tipo de comportamientos son los que más se repiten, siendo esto normal. El segundo momento, correspondiente al rectángulo amarillo, correspondiente a periodos de demanda de energía eléctrica de semanas con días festivos, esto se puede evidenciar a que existen dos días de baja demanda de forma consecutiva, comportamiento que se da por tener un lunes festivo. El tercer y último momento correspondiente al rectángulo verde, corresponde a periodos de demanda de energía eléctrica en semana santa, esto se puede evidenciar a que existen cuatro días de baja demanda, esto sucede ya que en semana santa existen varios días festivos consecutivos.

Con lo expuesto anteriormente, se puede evidenciar que existe una relación directa entre la demanda de energía eléctrica con los tipos de día, haciendo que día entre semana presenten un comportamiento diferente a días domingo o días festivos.

5.4.3. Autocorrelación

La autocorrelación [10], es una medida de similaridad que poseen las series de tiempo, esta medida tiene el mismo principio que las medidas correlación, es decir, se basa en encontrar que tanto se relaciona una variable respecto a otra, en este caso, la autocorrelación se basa en encontrar la relación de un variable en un instante x_t respecto a ella misma en un instante anterior x_{t-n}, donde n representa la cantidad de datos anteriores de la correlación. Al igual, que la correlación, la autocorrelación presenta valores -1 y 1, donde el valor -1 representa que los instantes están inversamente correlacionados, el valor 0 representa que los instantes no están correlacionados y el valor 1 representa que los instantes están directamente correlacionados. A continuación, se muestra el grafico de autocorrelación de la demanda de energía eléctrica en Antioquia para los primeros $n = 30$, instantes anteriores.

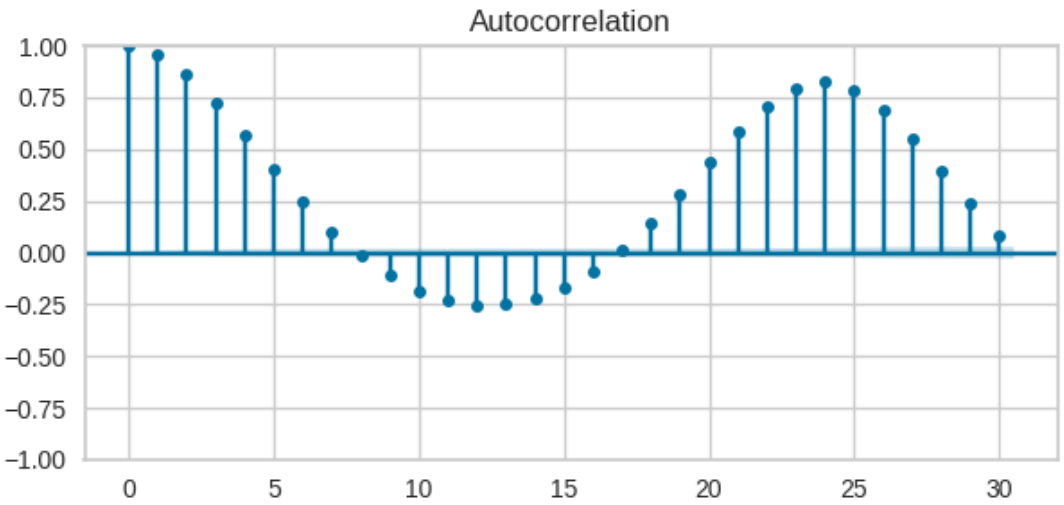


Figura 17: Autocorrelación de la demanda de energía eléctrica en Antioquia para los primeros $n = 30$ instantes anteriores.

La Figura 17 consta de dos ejes, el eje de ordenadas (y) corresponde al valor de la correlación, que como ya se mencionó tiene un rango de -1 a 1, y el eje de abscisas (x) corresponde a los n instantes anteriores. De la Figura 17 se puede notar que existe un ciclo que se repite cada 24 instantes o datos anteriores, esto tiene relación con lo que se observó en los análisis anteriores ya que se valida la existencia un ciclo diario en los datos, para los cuales se presenta una autocorrelación alta, es decir, se observa que se presta atención a los valores cercanos al instante anterior a las 24 horas antes, se nota que este es el mayor valor de autocorrelación que se presenta (si tener en cuenta los instantes anteriores que corresponde con las horas 1 y 2 antes del periodo a predecir). Lo anterior indica que la mejor forma de predecir un valor de la demanda en el instante "t", x_t , es usando los 24 valores previos de la demanda, x_{t-24}. Es importante aclarar que para predecir la demanda de energía eléctrica en un periodo dado se debe incluir el valor de la demanda en ese periodo del día anterior, por ejemplo, para predecir la demanda de la 1:00 pm de un día cualquiera lo mejor es usar la demanda de la 1:00 pm del día anterior. Igualmente, se deben tener en cuenta los valores de las demandas en los periodos previos al instante que se desea predecir.

5.5. División del datasets

Para todos los modelos desarrollados, de realizó una separación de 80% y 20% para los datasets de entrenamiento y validación respectivamente, cuyos valores en fechas del 1 de enero del 2015 hasta el 21 de diciembre del 2021, para el entrenamiento y del 21 de diciembre del 2021 hasta el 18 de septiembre del 2023 para la validación. A continuación, en la Figura 18 se muestra la división de los datasets en entrenamiento (train) y validación (test) de la demanda de energía eléctrica en Antioquia para los primeros.

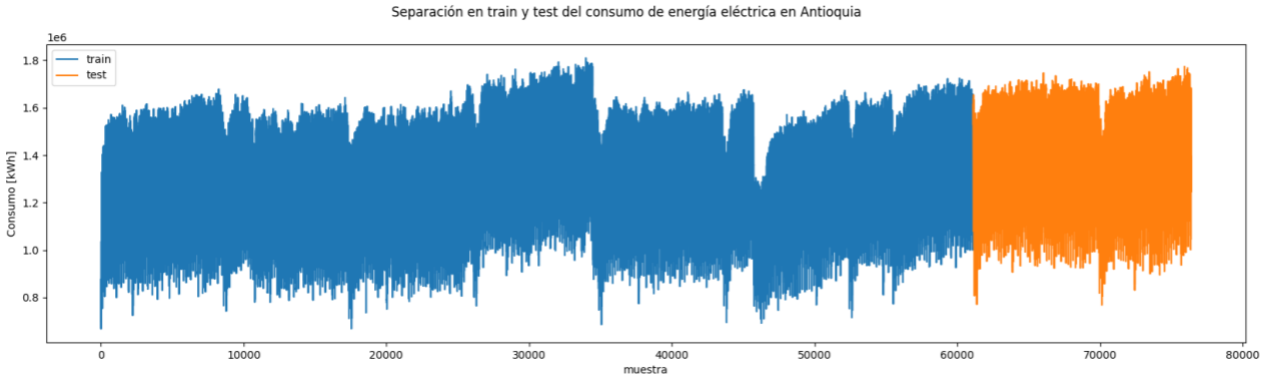


Figura 18: Separación en train y test de la demanda de energía eléctrica en Antioquia.

6. Descripción de los modelos

6.1. Modelos

Para realizar la predicción de la demanda de energía eléctrica en Antioquia, se utilizaron en total 5 modelos, dos ellos basados en modelos clásicos de la estadística para la predicción de series de tiempo, ARIMA y SARIMAX, también se usaron dos modelos de machine learning usando procesos autorregresivos para series tiempo, Linear Regressor y Random Foreste Regressor. También se utiliza una herramienta de aprendizaje profundo, se usó un modelo de redes neuronales recurrentes (RNN-LSTM), esto se realizó con el objetivo de comparar distintas técnicas de predicción provenientes de conceptos y filosofías distintas. A continuación, se presenta una descripción de los modelos utilizados y sus configuraciones.

6.1.1. Autoregressive Integrated Moving Average (ARIMA)

Para este modelo, se realizaron 4 iteraciones, que serán explicadas a continuación.

- ARIMA simple: En este modelo, debido a su funcionamiento, solo se ingresó la demanda de energía eléctrica sin algunas de las variables del datasets, esto con el objetivo de validar el comportamiento del autorregresivo del modelo ARIMA sin la ayuda de variables externas. A este modelo se le ingresó la siguiente configuración $p_{value} = 5$, $d_{value} = 1$ y $q_{value} = 0$.
- ARIMA con variable exógena de temperatura promedio: En este modelo, aparte de la variable de demanda de energía eléctrica, se le ingresó como variable exógena la temperatura promedio del departamento, como un predictor adiciona que facilite la predicción. A este modelo se le ingresó la siguiente configuración $p_{value} = 5$, $d_{value} = 1$, $q_{value} = 0$ y $X_{exog} = temperatura_prom.$
- ARIMA con variable exógena del tipo de día: En este modelo, aparte de la variable de demanda de energía eléctrica, se le ingresó como variable exógena del tipo de día como un predictor adiciona que facilite la predicción. A este modelo se le ingresó la siguiente configuración $p_{value} = 5$, $d_{value} = 1$, $q_{value} = 0$ y $X_{exog} = tipo_dia.$
- ARIMA con variable exógena del tipo de semana: En este modelo, aparte de la variable de demanda de energía eléctrica, se le ingresó como variable exógena del tipo de semana como un predictor adiciona que facilite la predicción. A este modelo se le ingresó la siguiente configuración $p_{value} = 5$, $d_{value} = 1$, $q_{value} = 0$ y $X_{exog} = tipo_semana.$

6.1.2. Seasonal Autoregressive Integrated Moving Average Exogenous Variables (SARIMAX)

Para este modelo, se realizaron 4 iteraciones, que serán explicadas a continuación.

- SARIMAX simple: En este modelo, debido a su funcionamiento, solo se ingresó la demanda de energía eléctrica sin datos de las variables del datasets, esto con el objetivo de validar el comportamiento del autorregresivo y estacional del modelo SARIMAX sin la ayuda de variables externas. A este modelo se le ingresó la siguiente configuración $p_{value} = 1$, $d_{value} = 1$, $q_{value} = 1$, $P_{value} = 1$, $D_{value} = 1$, $Q_{value} = 1$ y $S_{value} = 12$
- SARIMAX con variable exógena de temperatura promedio: En este modelo, aparte de la variable de demanda de energía eléctrica, se le ingresó como variable exógena la

temperatura promedio del departamento como un predictor adiciona que facilite la predicción. A este modelo se le ingresó la siguiente configuración $p_{value} = 1$, $d_{value} = 1$, $q_{value} = 1$, $P_{value} = 1$, $D_{value} = 1$, $Q_{value} = 1$, $S_{value} = 12$ y $X_{exog} = temperatura_prom.$

- SARIMAX con variable exógena del tipo de día: En este modelo, aparte de la variable de demanda de energía eléctrica, se le ingresó como variable exógena del tipo de día como un predictor adiciona que facilite la predicción. A este modelo se le ingresó la siguiente configuración $p_{value} = 1$, $d_{value} = 1$, $q_{value} = 1$, $P_{value} = 1$, $D_{value} = 1$, $Q_{value} = 1$, $S_{value} = 12$ y $X_{exog} = tipo_dia.$
- SARIMAX con variable exógena del tipo de semana: En este modelo, aparte de la variable de demanda de energía eléctrica, se le ingresó como variable exógena del tipo de semana como un predictor adiciona que facilite la predicción. A este modelo se le ingresó la siguiente configuración $p_{value} = 1$, $d_{value} = 1$, $q_{value} = 1$, $P_{value} = 1$, $D_{value} = 1$, $Q_{value} = 1$, $S_{value} = 12$ y $X_{exog} = tipo_semana.$

6.1.3. Linear Regressor Autoregressive

Para este modelo, se realizaron 4 iteraciones, que serán explicadas a continuación.

- Linear Regressor Autoregressive simple: En este modelo, debido a su funcionamiento, solo se ingresó la demanda de energía eléctrica sin datos de las variables del datasets, esto con el objetivo de validar el comportamiento del autorregresivo del modelo Linear Regressor Autoregressive sin la ayuda de variables externas. A este modelo se le ingresó la siguiente configuración $lag = 24$.
- Linear Regressor Autoregressive con variable exógena de temperatura promedio: En este modelo, aparte de la variable de demanda de energía eléctrica, se le ingresó como variable exógena la temperatura promedio del departamento como un predictor adiciona que facilite la predicción. A este modelo se le ingresó la siguiente configuración $lag = 24$ y $X_{exog} = temperatura_prom.$
- Linear Regressor Autoregressive con variable exógena del tipo de día: En este modelo, aparte de la variable de demanda de energía eléctrica, se le ingresó como variable exógena del tipo de día como un predictor adiciona que facilite la predicción. A este modelo se le ingresó la siguiente configuración $lag = 24$ y $X_{exog} = tipo_dia.$
- Linear Regressor Autoregressive con variable exógena del tipo de semana: En este modelo, aparte de la variable de demanda de energía eléctrica, se le ingresó como variable exógena del tipo de semana como un predictor adiciona que facilite la predicción. A este modelo se le ingresó la siguiente configuración $lag = 24$ y $X_{exog} = tipo_semana.$

6.1.4. Random Forest Regressor Autoregressive

Para este modelo, se realizaron 4 iteraciones, que serán explicadas a continuación.

- Random Forest Regressor Autoregressive simple: En este modelo, debido a su funcionamiento, solo se ingresó la demanda de energía eléctrica sin datos de las variables del datasets, esto con el objetivo de validar el comportamiento del autorregresivo del modelo Random Forest Regressor Autoregressive sin la ayuda de

variables externas. A este modelo se le ingresó la siguiente configuración $lag = 24$ y $random_state = 123$.

- Random Forest Regressor Autoregressive con variable exógena de temperatura promedio: En este modelo, aparte de la variable de demanda de energía eléctrica, se le ingresó como variable exógena la temperatura promedio del departamento como un predictor adiciona que facilite la predicción. A este modelo se le ingresó la siguiente configuración $lag = 24$, $random_state = 123$ y $X_{exog} = temperatura_prom.$
- Random Forest Regressor Autoregressive con variable exógena del tipo de día: En este modelo, aparte de la variable de demanda de energía eléctrica, se le ingresó como variable exógena del tipo de día como un predictor adiciona que facilite la predicción. A este modelo se le ingresó la siguiente configuración $lag = 24$, $random_state = 123$ y $X_{exog} = tipo_dia$.
- Random Forest Regressor Autoregressive con variable exógena del tipo de semana: En este modelo, aparte de la variable de demanda de energía eléctrica, se le ingresó como variable exógena del tipo de semana como un predictor adiciona que facilite la predicción. A este modelo se le ingresó la siguiente configuración $lag = 24$, $random_state = 123$ y $X_{exog} = tipo_semana$

6.1.5. Recurrent Neural Network (RNN-LSTM)

Para este modelo, se realizaron 4 iteraciones, que serán explicadas a continuación.

- RNN-LSTM simple: En este modelo, debido a su funcionamiento, solo se ingresó la demanda de energía eléctrica sin datos de las variables del datasets, esto con el objetivo de validar el comportamiento del recurrente del modelo RNN sin la ayuda de variables externas. A este modelo se le ingresó la siguiente configuración $model = 'sequential'$, $hidden_layer = 1$, $hidden_layer_neuron = 'LSTM'$, $output_layer_neuron = 'Dense'$ $lag = 24$, $n_features = 1$, $neurons = 50$, $activation = 'relu'$, $optimizer = 'adam'$, $epoch = 5$, $batch_size = 32$ y $loss_function = 'mse'$.
- RNN-LSTM con variable exógena de temperatura promedio: En este modelo, aparte de la variable de demanda de energía eléctrica, se le ingresó como variable exógena la temperatura promedio del departamento como un predictor adiciona que facilite la predicción. A este modelo se le ingresó la siguiente configuración $model = 'sequential'$, $hidden_layer = 1$, $hidden_layer_neuron = 'LSTM'$, $output_layer_neuron = 'Dense'$ $lag = 24$, $n_features = 1$, $neurons = 50$, $activation = 'relu'$, $optimizer = 'adam'$, $epoch = 5$, $batch_size = 32$, $loss_function = 'mse'$ y $X_{exog} = temperatura_prom.$
- RNN-LSTM con variable exógena del tipo de día: En este modelo, aparte de la variable de demanda de energía eléctrica, se le ingresó como variable exógena del tipo de día como un predictor adiciona que facilite la predicción. A este modelo se le ingresó la siguiente $model = 'sequential'$, $hidden_layer = 1$, $hidden_layer_neuron = 'LSTM'$, $output_layer_neuron = 'Dense'$ $lag = 24$, $n_features = 1$, $neurons = 50$, $activation = 'relu'$, $optimizer = 'adam'$, $epoch = 5$, $batch_size = 32$, $loss_function = 'mse'$ y $X_{exog} = tipo_dia$.

- RNN-LSTM con variable exógena del tipo de semana: En este modelo, aparte de la variable de demanda de energía eléctrica, se le ingresó como variable exógena del tipo de semana como un predictor adiciona que facilite la predicción. A este modelo se le ingresó la siguiente configuración $model = 'sequential'$, $hidden_layer = 1$, $hidden_layer_neuron = 'LSTM'$, $output_layer_neuron = 'Dense'$ $lag = 24$, $n_features = 1$, $neurons = 50$, $activation = 'relu'$, $optimizer = 'adam'$, $epoch = 5$, $batch_size = 32$, $loss_function = 'mse'$ y $X_{exog} = tipo_semana$

6.2. Ajuste de hiperparámetros

Para realizar el ajuste de hiperparámetros de los modelos antes mencionados, se utilizo el método de grid search, definiendo un espacio de modelos en donde buscar el modelo que mejore las métricas de validación respecto a los modelos base. A continuación, se presenta una descripción de los hiperparámetros usados en el método del grid search.

6.2.1. Autoregressive Integrated Moving Average (ARIMA)

Para el método del grid search en este modelo, se utilizaron un espacio de modelos con los siguientes hiperparámetros.

$$p_{values} = [0, 1, 2, 3, 4, 5]$$
$$d_{values} = [0, 1, 2, 3]$$
$$q_{values} = [0, 1, 2, 3, 4, 5]$$
$$metric_error = MAPE$$

Obteniendo como resultado el mejor modelo dado por los siguientes hiperparámetros:

$$p_{values} = 4$$
$$d_{values} = 1$$
$$q_{values} = 5$$

Estos hiperparámetros fueron reemplazados en cada uno de los modelos descritos en la Sección 6.1.1. de este documento para realizar nuevas predicciones a la demanda de energía eléctrica en Antioquia.

6.2.2. Seasonal Autoregressive Integrated Moving Average Exogenous Variables (SARIMAX)

Para el método del grid search en este modelo, se utilizaron un espacio de modelos con los siguientes hiperparámetros.

$$p_{values} = [0, 1, 2, 3, 4, 5]$$
$$d_{values} = [0, 1, 2, 3]$$
$$q_{values} = [0, 1, 2, 3, 4, 5]$$
$$P_{values} = [0, 1, 2, 3, 4, 5]$$
$$D_{values} = [0, 1, 2, 3]$$
$$Q_{values} = [0, 1, 2, 3, 4, 5]$$
$$s_{values} = [12, 23]$$
$$metric_error = MAPE$$

Obteniendo como resultado el mejor modelo dado por los siguientes hiperparámetros:

$$p_{values} = 4$$
$$d_{values} = 1$$
$$q_{values} = 5$$
$$P_{values} = 1$$
$$D_{values} = 1$$
$$Q_{values} = 1$$
$$s_{values} = 12$$

Estos hiperparámetros fueron reemplazados en cada uno de los modelos descritos en la Sección 6.1.2. de este documento para realizar nuevas predicciones a la demanda de energía eléctrica en Antioquia.

6.2.3. Random Forest Regressor Autoregressive

Para el método del grid search en este modelo, se utilizaron un espacio de modelos con los siguientes hiperparámetros.

$$lags = [3, 7, 10, 15, 20, 24, 25, 30]$$
$$n_{estimators} = [10, 20, 30, 40, 50, 100, 150, 200]$$
$$max_{depth} = [2, 4, 6, 8, 10, 50, 100]$$
$$metric_error = MSE$$

Obteniendo como resultado el mejor modelo dado por los siguientes hiperparámetros:

$$lags = 24$$
$$n_{estimators} = 10$$
$$max_{depth} 100$$

Estos hiperparámetros fueron reemplazados en cada uno de los modelos descritos en la Sección 6.1.4. de este documento para realizar nuevas predicciones a la demanda de energía eléctrica en Antioquia.

6.2.4. RNN-LSTM

Para el método del grid search en este modelo, se utilizaron un espacio de modelos con los siguientes hiperparámetros.

$$neurons = [20, 50, 100, 150, 200]$$
$$activation = ['relu', 'tanh', 'sigmoid']$$
$$optimizer = ['adam', 'sgd', 'rmsprop']$$
$$batch_size = [16, 32, 64, 128]$$
$$epoch = [5, 10, 25, 50, 100, 200]$$

Obteniendo como resultado el mejor modelo dado por los siguientes hiperparámetros:

$$neurons = 50$$
$$activation = 'relu'$$
$$optimizer = 'adam'$$
$$batch_size = 32$$
$$epoch = 25$$

Estos hiperparámetros fueron reemplazados en cada uno de los modelos descritos en la Sección 6.1.5. de este documento para realizar nuevas predicciones a la demanda de energía eléctrica en Antioquia.

6.3. Métricas

Para las métricas de desempeño de cada uno de los veinte anteriores modelos, se calcula las métricas de validación de tres formas diferentes:

- Error de test: Para este caso se calculan las respectivas métricas de error a todo el conjunto de datos de test con el objetivo de validar el modelo, obteniendo así MAE_{test}, $RMSE_{test}$ y $MAPE_{test}$
- Error de train: Para este caso se calculan las respectivas métricas de error a todo el conjunto de datos de train con el objetivo de comparar estas métricas con las métricas de test para validar que no exista sobre ajuste en los modelos, obteniendo así MAE_{train}, $RMSE_{train}$ y $MAPE_{train}$
- Error de test simplificado: Para este caso se calculas las respectivas métricas de error a todo el conjunto de datos de test, pero solo para un periodo de corto plazo de una semana, estás son las métricas de desempeño reales con las cuales se tomará la decisión final del mejor modelo, ya que, este es el parámetro y objetivo principal de este trabajo, encontrar el modelo más preciso para el corto plazo de una semana de predicciones horarias de la demanda de energía eléctrica en Antioquia, obteniendo así $MAE_{test_simplificado}$, $RMSE_{test_simplificado}$ y $MAPE_{test_simplificado}$

7. Resultados y discusión

7.1. Métricas

En la Tabla 1, están registrados los resultados obtenidos de las métricas de validación tanto para test como para train de los veinte diferentes modelos implementados para este proyecto, además, como bien se mencionó en la Sección 6.3, el objetivo principal de este proyecto es la mejora la exactitud en la predicción de la demanda a corto plazo de una semana, también se presentas las métricas simplificadas de los modelos en la Tabla 2, en ambas tablas, se marcarán con negrita los mejores modelos de cada tipo dándole según las métricas de validación.

Tabla 1
Resultados obtenidos de las métricas de validación para train y test

Modelos	Train			Test		
	MAE [kWh]	RMSE [kWh]	MAPE [p.u.]	MAE [kWh]	RMSE [kWh]	MAPE [p.u.]
- ARIMA Simple	2.6x10^5	3.2x10^5	0.1910	3.2x10^5	3.8x10^5	0.2209
- ARIMA con variable exógena de temperatura promedio	2.2x10^5	2.7x10^5	0.1653	2.8x10^5	3.3x10^5	0.1909
- ARIMA con variable exógena del tipo de día	2.6x10^5	3.1x10^5	0.1889	3.2x10^5	3.8x10^5	0.2176
- ARIMA con variable exógena del tipo de semana	**3.2x10^5**	**3.8x10^5**	**0.2206**	**2.4x10^5**	**2.9x10^5**	**0.1698**
- SARIMAX Simple	2.1x10^4	2.9x10^5	0.0178	4.3x10^6	5.0x10^6	3.2631
- SARIMAX con variable exógena de temperatura promedio	**3.3x10^4**	**4.5x10^4**	**0.0272**	**1.7x10^6**	**1.9x10^6**	**1.2761**
- SARIMAX con variable exógena del tipo de día	2.1x10^4	2.9x10^4	0.0178	4.2x10^6	4.9x10^6	3.2001
- SARIMAX con variable exógena del tipo de semana	2.1x10^4	2.9x10^4	0.0178	4.2x10^6	4.9x10^6	3.2001
- Linear regressor Autoregressive Simple	1.9x10^5	2.2x10^5	0.1325	2.1x10^5	2.4x10^5	0.1529
- Linear regressor Autoregressive con variable exógena de temperatura promedio	**1.5x10^5**	**2.0x10^5**	**0.1295**	**1.8x10^5**	**2.2x10^5**	**0.1357**
- Linear regressor Autoregressive con variable exógena del tipo de día	1.5x10^5	2.1x10^5	0.1362	1.9x10^5	2.2x10^5	0.1444
- Linear regressor Autoregressive con variable exógena del tipo de semana	1.9x10^5	2.2x10^5	0.1346	2.1x10^5	2.4x10^5	0.1526
- Random Forest Autoregressive Simple	1.2x10^5	1.3x10^5	0.0985	1.4x10^5	1.8x10^5	0.1093
- Random Forest Autoregressive con variable exógena de temperatura promedio	9.7x10^4	1.2x10^5	0.0689	1.1x10^5	1.4x10^5	0.0890
- Random Forest Autoregressive con variable exógena del tipo de día	1.0x10^5	1.1x10^5	0.0826	1.1x10^5	1.2x10^5	0.0846
- Random Forest Autoregressive con variable exógena del tipo de semana	**8.9x10^4**	**1.1x10^5**	**0.0647**	**9.3x10^4**	**1.3x10^5**	**0.0742**
- RNN LSTM Simple	8.8x10^4	1.2x10^5	0.0717	2.5x10^4	3.3x10^4	0.0194
- RNN LSTM con variable exógena de temperatura promedio	9.2x10^4	1.3x10^5	0.0744	2.7x10^4	3.3x10^4	0.0206
- RNN LSTM con variable exógena del tipo de día	9.0x10^4	1.3x10^5	0.0735	2.0x10^4	2.7x10^4	0.0155

- RNN LSTM con variable exógena del tipo de semana	**8.8x10^4**	**1.3x10^5**	**0.0726**	**2.0x10^4**	**2.7x10^4**	**0.0153**

Tabla 2
Resultados obtenidos de las métricas de validación simplificadas

Modelos	Métricas		
	MAE [kWh]	RMSE [kWh]	MAPE [p.u.]
- ARIMA Simple	3.3x10^5	3.9x10^5	0.2257
- ARIMA con variable exógena de temperatura promedio	**2.1x10^5**	**2.8x10^5**	**0.1478**
- ARIMA con variable exógena del tipo de día	3.3x10^5	3.8x10^5	0.2222
- ARIMA con variable exógena del tipo de semana	3.3x10^5	3.9x10^5	0.2257
- SARIMAX Simple	2.0x10^5	2.6x10^5	0.1346
- SARIMAX con variable exógena de temperatura promedio	**1.7x10^5**	**2.0x10^5**	**0.1211**
- SARIMAX con variable exógena del tipo de día	2.0x10^5	2.7x10^5	0.1330
- SARIMAX con variable exógena del tipo de semana	2.0x10^5	2.5x10^5	0.1306
- Linear regressor Autoregressive Simple	1.4x10^5	1.6x10^5	0.0994
- Linear regressor Autoregressive con variable exógena de temperatura promedio	1.4x10^5	1.6x10^5	0.0970
- Linear regressor Autoregressive con variable exógena del tipo de día	**1.1x10^5**	**1.3x10^5**	**0.0772**
- Linear regressor Autoregressive con variable exógena del tipo de semana	1.3x10^5	1.5x10^5	0.0942
- Random Forest Autoregressive Simple	1.3x10^5	1.8x10^5	0.0985
- Random Forest Autoregressive con variable exógena de temperatura promedio	1.3x10^4	1.7x10^4	0.0967
- Random Forest Autoregressive con variable exógena del tipo de día	**4.7x10^4**	**5.2x10^4**	**0.0343**
- Random Forest Autoregressive con variable exógena del tipo de semana	1.1x10^5	1.5x10^5	0.0840
- RNN LSTM Simple	2.6x10^4	3.3x10^4	0.0193
- RNN LSTM con variable exógena de temperatura promedio	2.6x10^4	3.2x10^4	0.0194
- RNN LSTM con variable exógena del tipo de día	1.9x10^4	2.4x10^4	0.0140
- RNN LSTM con variable exógena del tipo de semana	**1.8x10^4**	**2.3x10^4**	**0.0132**

Lo primero que se debe analizar en la Tabla 1, es que ninguno de los modelos entrenados y sus iteraciones presentan sobreajustes considerables, ya que, si bien en modelos como el ARIMA o el Linear Regressor Autorregresive existen métricas de test que presentan valores un poco mayores en comparación a las métricas de train, no son valores que generen sospecha de sobreajuste, por otro lado, como ya se mencionó, si bien el objetivo de este proyecto es encontrar un modelos para mejorar la exactitud de las predicciones de la demanda de energía eléctrica a corto, con un horizonte de la predicción de una semana, es necesario mencionar que algunos modelos se comportan mejor que otros dependiendo del horizonte de la predicción, lo cual nos puede dar indicios de cuáles son los mejores modelos a largo, medio o corto plazo, en este caso si no se tiene en cuenta las métricas simplificadas, es decir, las métricas para poner el modelo en producción, los mejores modelos son en primer lugar las RNN LSTM, seguido por los obtenidos con técnicas de machine learning con comportamiento autorregresivos, es decir, Random Forest Autoregressive y Linear Regressor Autoregressive, y por último, los modelos estadísticos clásicos, ARIMA y SARIMAX.

Ahora, al analizar los resultados de la Tabla 2, se puede evidenciar que en la mayoría de los tipos de modelos, los mejores resultado se obtienen cuando se ingresan variables exógenas que ayudan a predecir la demanda de energía, salvo en los modelos RNN LSTM, donde el mejor modelo es donde no se usan variables exógenas, además, se puede notar que las métricas de test presentadas entre las Tabla 1 y Tabla 2, mejoran considerablemente en todo los tipos de modelos, esto es debido a que todos los modelos realizan un proceso recursivo, donde para calcular un valor en Y_t utilizan valores anteriores de Y_t (Y_{t-1}, Y_{t-2}, etc.) por lo cual, al avanzar en las predicciones Y_t, va llegar el punto donde las predicciones dependerán de valores ya predichos, lo cual hace que los modelos luego de un instante de tiempo t las predicciones pierdan variabilidad, esto se puede visualizar con más claridad en la Figura 19 se visualiza este proceso, esto no sucede en modelos como la RNN LSTM, debido a que estas redes neuronales poseen celdas de memoria como se puede apreciar en la Figura 4, en las cuales se puede almacenar información a largo plazo, estas celdas de memoria permite que las redes aprendan patrones de tendencia y estacionalidad a lo largo de todo la secuencia de datos, por lo cual en el proceso de predicción se toma el dato actual y la información de datos anteriores para predecir el dato futuro [12].

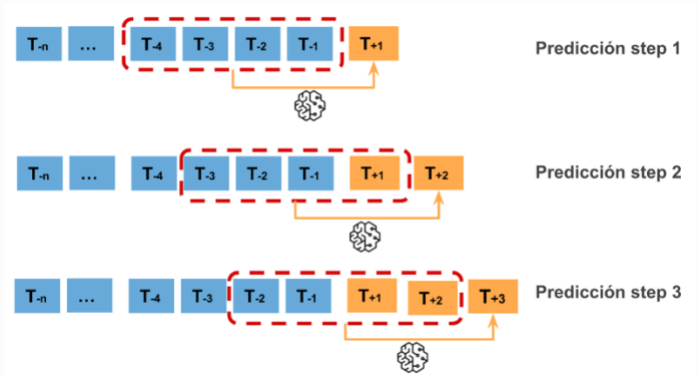


Figura 19: Ejemplo de cómo funciona la recursividad en los modelos autorregresivos [15]

De todos los tipos de modelo presentados los que mejores métricas presentan en general para la predicción de la demanda de energía eléctrica a corto plazo con un horizonte de predicción de una semana son los RNN LSTM, seguido por Random Forest Autoregressive, Linear Regressor Autorregresive, SARIMAX y por último ARIMA, teniendo métricas por debajo del 5% de error según el MAPE para algunos modelos como el Random Forest Autoregressive con variable exógena de tipo de día o tipo de semana, RNN LSTM simple o con variable exógena de tipo de día, tipo de semana o temperatura promedio, lo cual, es el error mínimo esperado para estos modelos. De estos estos veinte modelos el que mejores métricas presenta es el RNN LSTM sin variables exógenas teniendo un error de 1.3%.

A continuación, se presentan los gráficos de los resultados de los mejores modelos de las predicciones de la demanda de energía eléctrica para cada tipo de modelo.

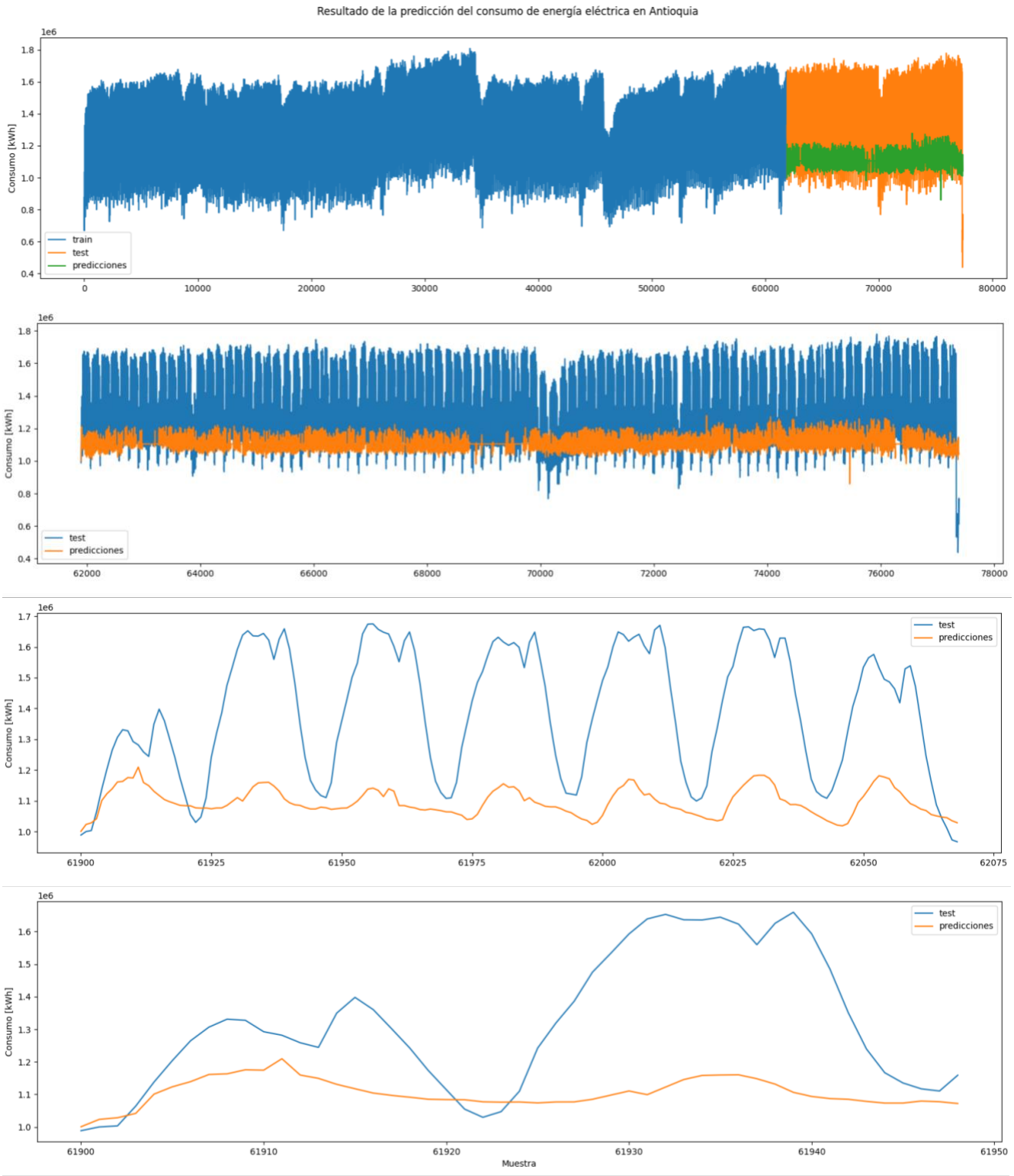


Figura 20: Resultado de la predicción de la demanda de energía eléctrica en Antioquia del modelo ARIMA con variable exógena del temperatura promedio.

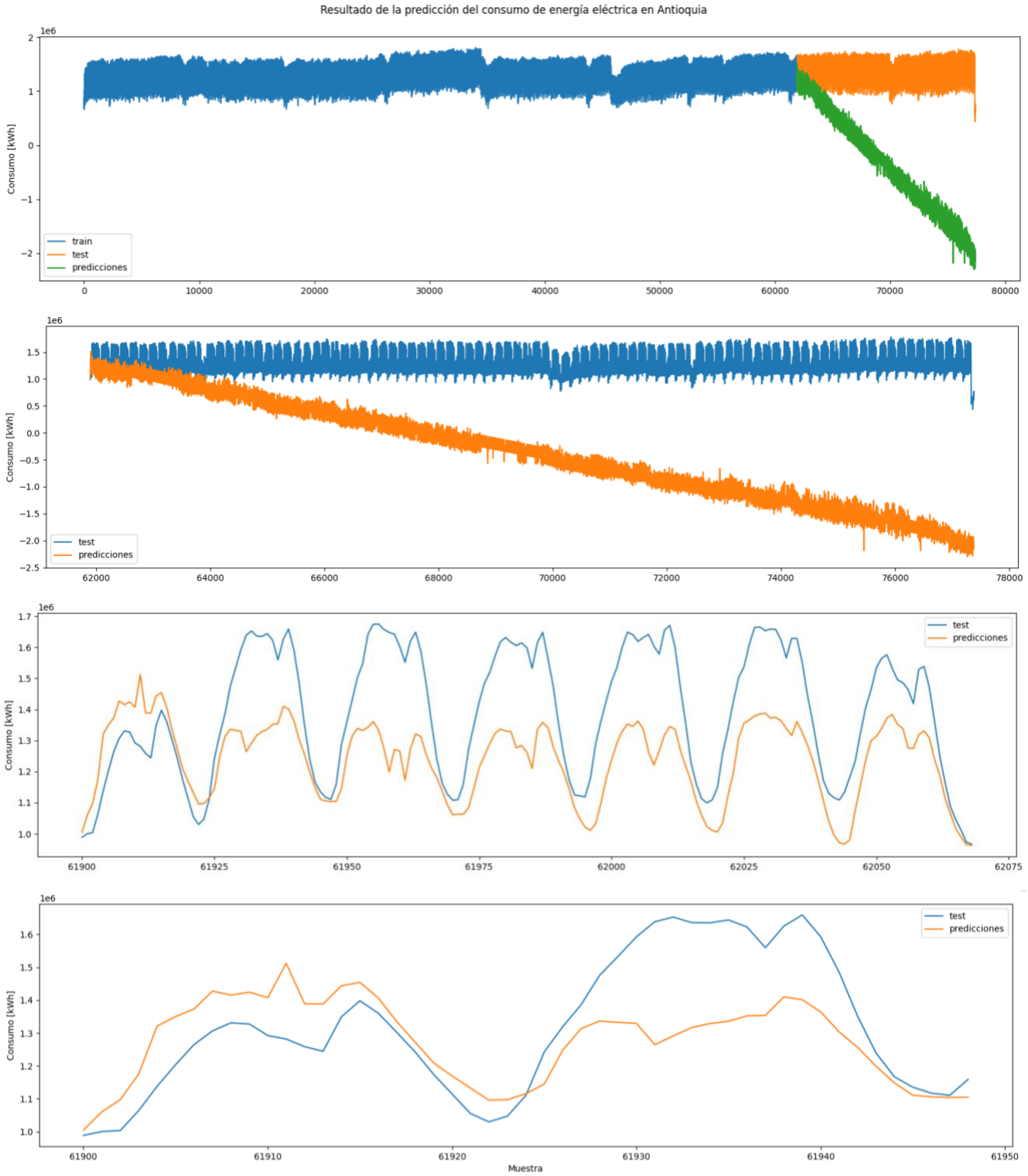


Figura 21: Resultado de la predicción de la demanda de energía eléctrica en Antioquia del modelo SARIMAX con variable exógena de temperatura promedio.

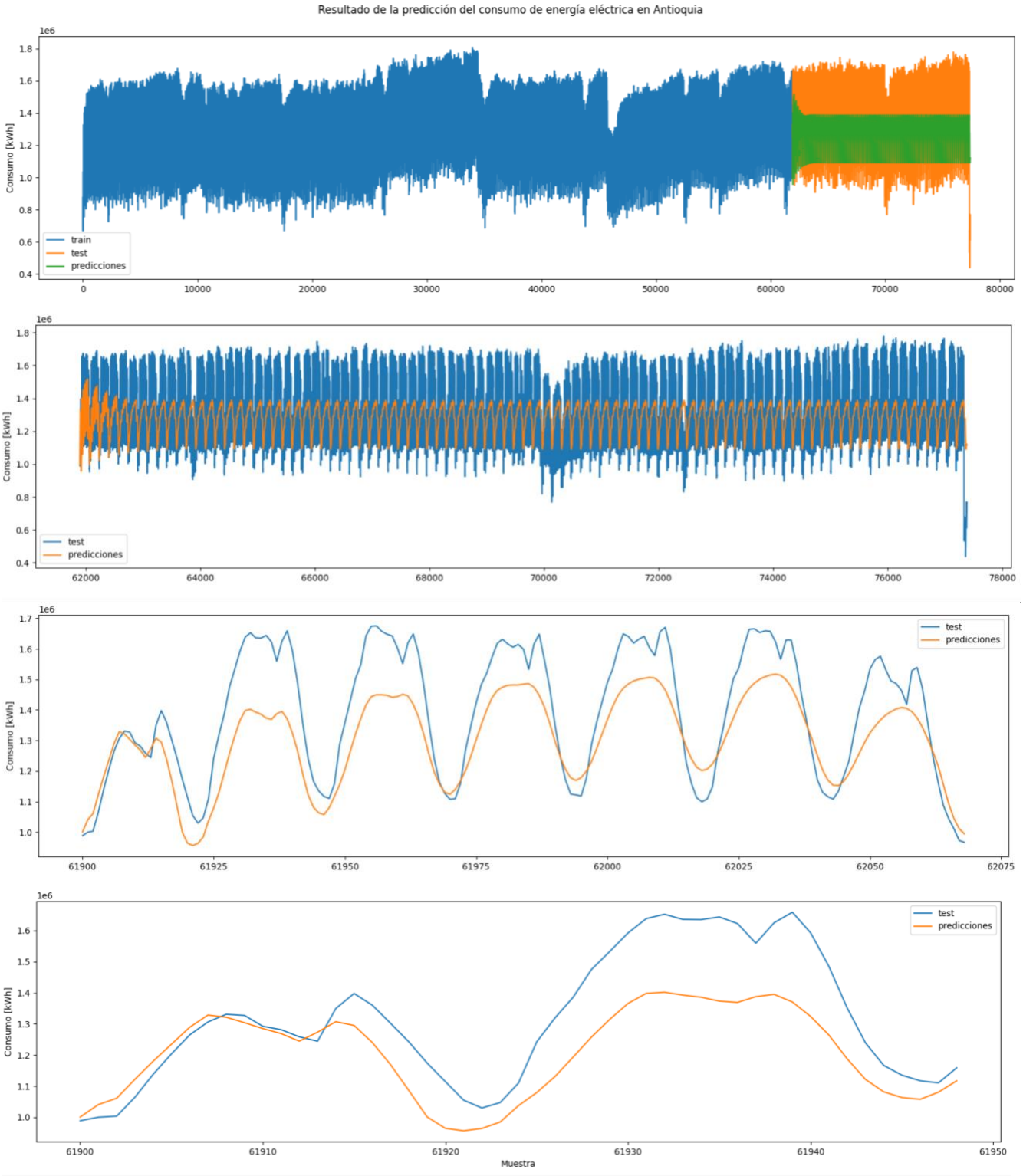


Figura 22: Resultado de la predicción de la demanda de energía eléctrica en Antioquia del modelo Linear Regressor Autoregressive con variable exógena del tipo de día.

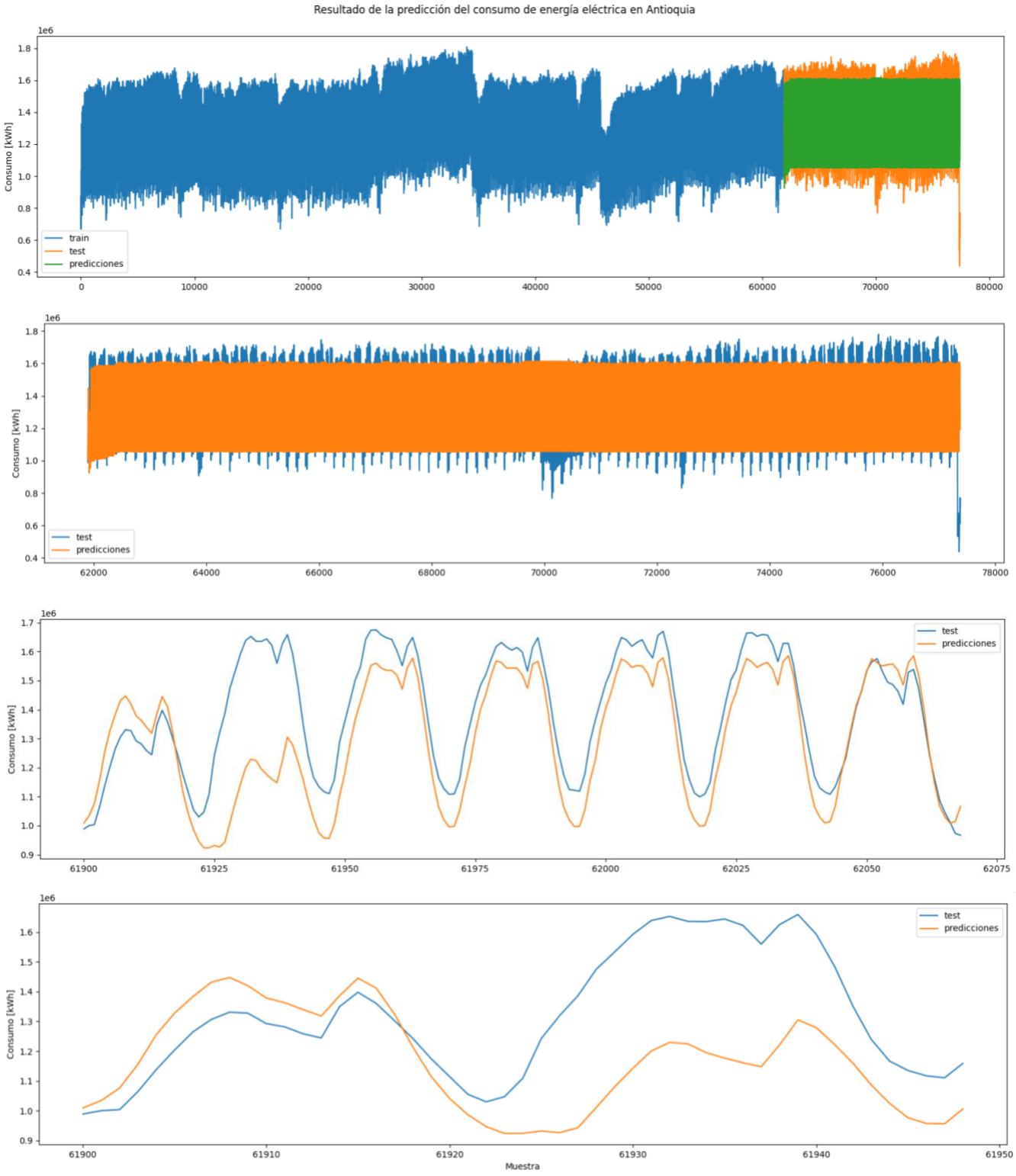


Figura 23: Resultado de la predicción de la demanda de energía eléctrica en Antioquia del modelo Random Forest Autoregressive con variable exógena del tipo de día.

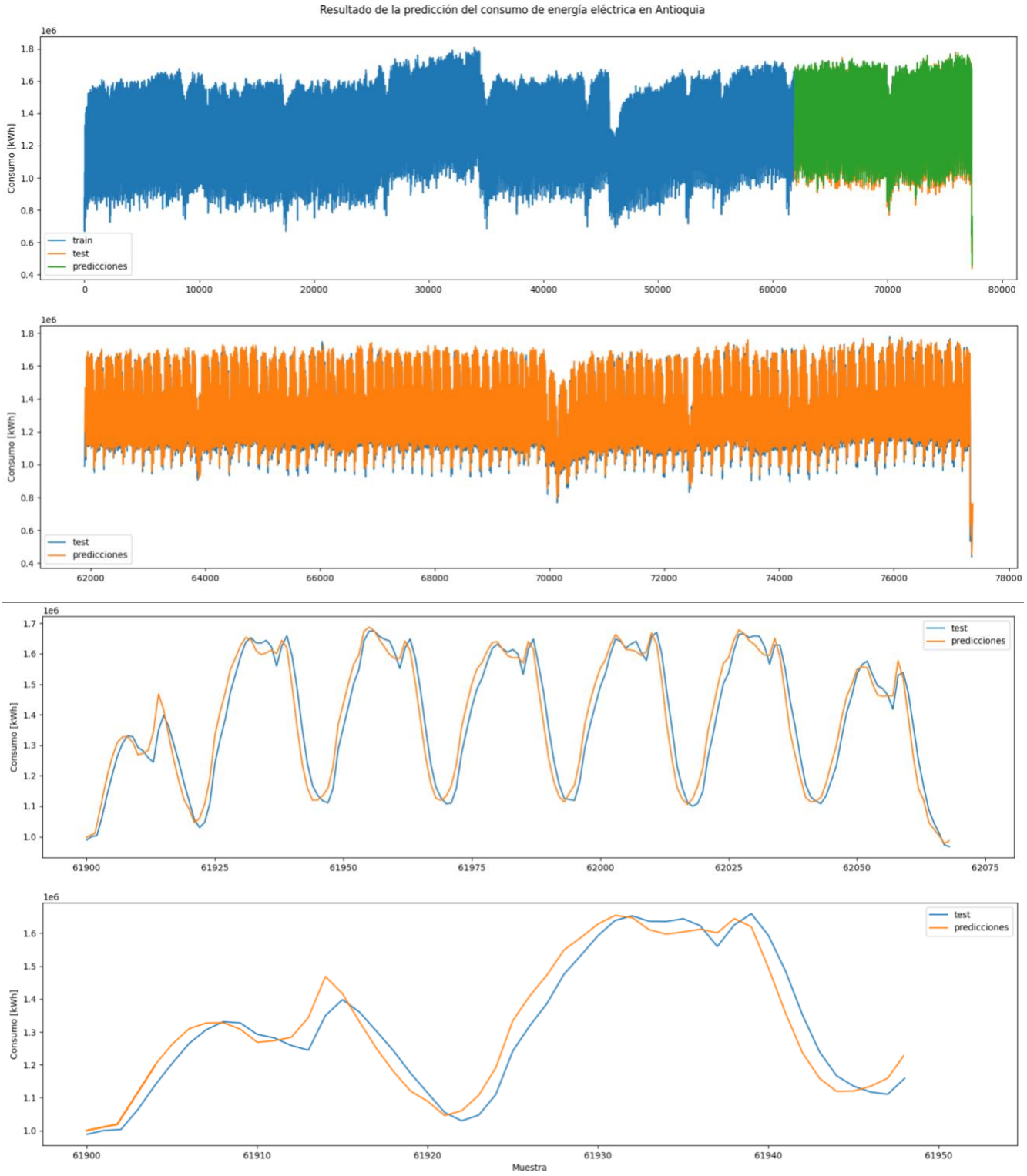


Figura 24: Resultado de la predicción de la demanda de energía eléctrica en Antioquia del modelo RNN LSTM con variable exógena del tipo de semana.

En las Figura 20, Figura 21, Figura 22, Figura 23 y Figura 24 se presentan 4 sub gráficos, los cuales, representan lo siguiente, el primer sub grafico es la visualización del resultado del modelos para todo el conjunto de datos de test, incluyendo el conjunto de datos de train; El segundo sub grafico es la visualización del resultado del modelo para el conjunto de datos de test, haciendo énfasis solo en el conjunto de test; El tercer sub grafico es la visualización del resultado del modelo para el conjunto de datos de test, haciendo énfasis en la primer semana de la predicción; El cuarto sub grafico es la visualización del resultado del modelo para el conjunto de datos de test, haciendo énfasis en los primer dos días de la predicción.

En la Figura 20 donde se muestra la predicción de la demanda de energía eléctrica con un modelo ARIMA, en este caso usando la temperatura promedio como variable exógena, se puede observar en los cuatro sub gráficos, que este modelo no realizó una predicción de forma adecuada de la demanda de energía, debido a que este no logra ajustarse correctamente ni a la tendencia ni estacionalidad de los datos, además, se puede observar con más detalle en el tercer sub gráfico, que la predicción intenta replicar la estacionalidad de los datos sin poder predecir los valores indicados, manteniéndose siempre en un valor promedio donde se presenta una ligera estacionalidad.

En la Figura 21 donde se muestra la predicción de la demanda de energía eléctrica con un modelo SARIMAX, en este caso se usando la temperatura promedio como variable exógena, se puede apreciar que en este modelo luego de aproximadamente ocho semanas, el valor medio de la predicción comienza a decaer y no logra mantenerse en los valores límites de los datos, sin embargo al observar con más detalle el tercer y cuarto sub gráfico, se puede notar como a pesar de esto, el modelo intenta predecir de mejor forma que el modelo ARIMA la estacionalidad de los datos, aunque sin lograr predecir los valores adecuados. Este modelo si bien presenta el problema de luego de un periodo de tiempo decae la tendencia de los datos, al centrarse solo en el horizonte de la predicción establecida, el modelo replica la estacionalidad de los datos y, por ende, este modelo es adecuado solo en el corto plazo.

En la Figura 22 donde se muestra la predicción de la demanda de energía eléctrica con un modelo Linear Regressor Autorregresive, en este caso usando el tipo de día como variable exógena, se puede apreciar en el primer sub grafico que luego de cierto periodo de tiempo la predicción comienza a oscilar intentando mantener la estacionalidad de los datos, manteniéndose siempre entre dos valores limites, casi como si fueran cotas, si bien esto sucede, al centrarse en el horizonte de la predicción, que se muestra en el tercer y cuarto sub gráfico, se puede apreciar que la predicción se ajuste de mejor manera a los modelos anteriores de ARIMA y SARIMAX a la estacionalidad de los datos, incluso, ajustándose de forma adecuada en algunos momentos, sin embargo, este modelo no lograr ajustar por completo, dando la sensación de la predicción tuviese un comportamiento senoidal durante todo el día, sin embargo, la demanda de energía durante un día cualquier presenta periodos de máximo y mínima demanda que el modelo no logra predecir.

En la Figura 23 donde se muestra la predicción de la demanda de energía eléctrica con un modelo Random Forest Autorregresive, en este caso usando el tipo de día como variable exógena, se puede notar que al igual que con el modelo Linear Regressor Autoregressive, luego de un cierto periodo de tiempo la predicción comienza a oscilar intentando mantener la estacionalidad de los datos, manteniéndose entre dos cotas, sin embargo, estas cotas o limites son valores que en el modelo anterior mencionado, por lo cual, esto hace que la predicción se ajuste mejor a los datos, además, como se puede apreciar en el tercer y cuarto sub gráfico, la predicción aparte de ajustarse

a la estacionalidad de los datos, también logra ajustarse correctamente a los máximos y mínimos de la demanda de energía durante el día.

En la Figura 24 donde muestra la predicción de la demanda de energía eléctrica con un modelo RNN LSTM, en este caso usando el tipo de semana como variable exógena, se puede notar que este modelo es el que mejores predicciones realiza de los datos, logrando mantener la estacionalidad y tendencia de los datos sin importar en qué momento de la predicción se encuentre, logrando predicciones que se ajustan con exactitud tanto en el corto, mediano y largo plazo.

7.2. Métricas modelos con ajuste de hiperparámetros

En la Tabla 3, están registrados los resultados obtenidos de las métricas de validación tanto para test como para train de los veinte diferentes modelos implementados para este proyecto con el ajuste de hiperparámetros, mostrada en la Sección 6.2, además, también se presentas las métricas simplificadas de los modelos en la Tabla 4, en ambas tablas, se marcarán con negrita los mejores modelos de cada tipo dándole según las métricas de validación.

Tabla 3
Resultados obtenidos de las métricas de validación para train y test con ajuste de hiperparámetros

Modelos	Train			Test		
	MAE [kWh]	RMSE [kWh]	MAPE [p.u]	MAE [kWh]	RMSE [kWh]	MAPE [p.u.]
- ARIMA Simple	2.0x10^5	2.3x10^5	0.1650	2.2x10^5	2.5x10^5	0.1560
- ARIMA con variable exógena de temperatura promedio	1.8x10^5	2.1x10^5	0.1615	1.8x10^5	2.0x10^5	0.1407
- ARIMA con variable exógena del tipo de día	**2.0x10^5**	**2.4x10^5**	**0.1822**	**2.0x10^5**	**2.2x10^5**	**0.13547**
- ARIMA con variable exógena del tipo de semana	2.0x10^5	2.3x10^5	0.1651	2.1x10^5	2.5x10^5	0.1556
- SARIMAX Simple	2.1x10^4	2.9x10^5	0.0175	3.8x10^6	4.4x10^6	2.8671
- SARIMAX con variable exógena de temperatura promedio	**3.3x10^4**	**4.4x10^4**	**0.02688**	**2.0x10^5**	**2.5x10^5**	**0.1689**
- SARIMAX con variable exógena del tipo de día	2.0x10^4	2.9x10^4	0.0170	2.1x10^6	2.4x10^6	1.5612
- SARIMAX con variable exógena del tipo de semana	2.0x10^4	2.9x10^4	0.0170	2.1x10^6	2.4x10^6	1.512
- Random Forest Autoregressive Simple	1.3x10^5	1.7x10^5	0.0995	1.5x10^5	2.0x10^5	0.1182
- Random Forest Autoregressive con variable exógena de temperatura promedio	1.3x10^4	1.4x10^5	0.0889	1.6x10^5	1.9x10^5	0.1174
- Random Forest Autoregressive con variable exógena del tipo de día	**1.1x10^5**	**1.3x10^5**	**0.0825**	**1.3x10^5**	**1.5x10^5**	**0.0983**
- Random Forest Autoregressive con variable exógena del tipo de semana	1.1x10^5	1.5x10^5	0.0892	1.4x10^5	1.8x10^5	0.1092
- RNN LSTM Simple	8.8x10^4	1.3x10^5	0.0726	1.3x10^4	1.8x10^4	0.0105
- RNN LSTM con variable exógena de temperatura promedio	8.7x10^4	1.3x10^5	0.0716	1.7x10^4	2.2x10^4	0.0130
- RNN LSTM con variable exógena del tipo de día	8.9x10^4	1.3x10^5	0.0741	1.8x10^4	2.3x10^4	0.0140
- RNN LSTM con variable exógena del tipo de semana	**8.9x10^4**	**1.3x10^5**	**0.0732**	**1.1x10^4**	**1.5x10^4**	**0.0088**

Tabla 4
Resultados obtenidos de las métricas de validación simplificadas con ajuste de hiperparámetros

Modelos	Métricas		
	MAE [kWh]	RMSE [kWh]	MAPE [p.u.]
- ARIMA Simple	1.7x10^5	2.0x10^5	0.1198
- ARIMA con variable exógena de temperatura promedio	1.5x10^5	1.7x10^5	0.1185
- ARIMA con variable exógena del tipo de día	1.8x10^5	2.0x10^5	0.1344
- ARIMA con variable exógena del tipo de semana	**1.6x10^5**	**1.9x10^5**	**0.1151**
- SARIMAX Simple	1.9x10^5	2.4x10^5	0.1295
- SARIMAX con variable exógena de temperatura promedio	**1.2x10^5**	**1.4x10^5**	**0.0856**
- SARIMAX con variable exógena del tipo de día	1.6x10^5	2.0x10^5	0.1115
- SARIMAX con variable exógena del tipo de semana	1.6x10^5	2.0x10^5	0.1115
- Random Forest Autoregressive Simple	1.3x10^5	1.7x10^5	0.0988
- Random Forest Autoregressive con variable exógena de temperatura promedio	1.4x10^5	1.8x10^5	0.1033
- Random Forest Autoregressive con variable exógena del tipo de día	**4.7x10^4**	**5.2x10^4**	**0.0345**
- Random Forest Autoregressive con variable exógena del tipo de semana	1.9x10^5	2.2x10^5	0.1369
- RNN LSTM Simple	1.2x10^4	1.5x10^4	0.0091
- RNN LSTM con variable exógena de temperatura promedio	1.7x10^4	2.1x10^4	0.0128
- RNN LSTM con variable exógena del tipo de día	1.7x10^4	2.0x10^4	0.0126
- RNN LSTM con variable exógena del tipo de semana	**1.0x10^4**	**1.2x10^4**	**0.0076**

Al comparar los resultados de la Tabla 1 con la Tabla 3, y la Tabla 2 con la Tabla 4 se puede notar el ajuste de parámetros mejoró las métricas de desempeño, en especial en los modelos SARIMAX y ARIMA, en el caso de los RNN LSTM las mejoras en las métricas de desempeño no es tan considerable y en el caso de los modelos Random Forest autoregressive, se presenta una ligera disminución en las métricas.

De todos los tipos de modelo presentados con ajuste de hiperparámetros los que mejores métricas presentan en general para la predicción de la demanda de energía eléctrica a corto plazo con un horizonte, ARIMA y por último SARIMAX. De estos estos dieciséis modelos con ajuste de hiperparámetros el que mejores métricas presenta es el RNN LSTM con variables exógenas del tipo de semana teniendo un error de 0.7%.

A continuación, se presentan los gráficos de los resultados de los mejores modelos de las predicciones de la demanda de energía eléctrica para cada tipo de modelo con ajuste de hiperparámetros.

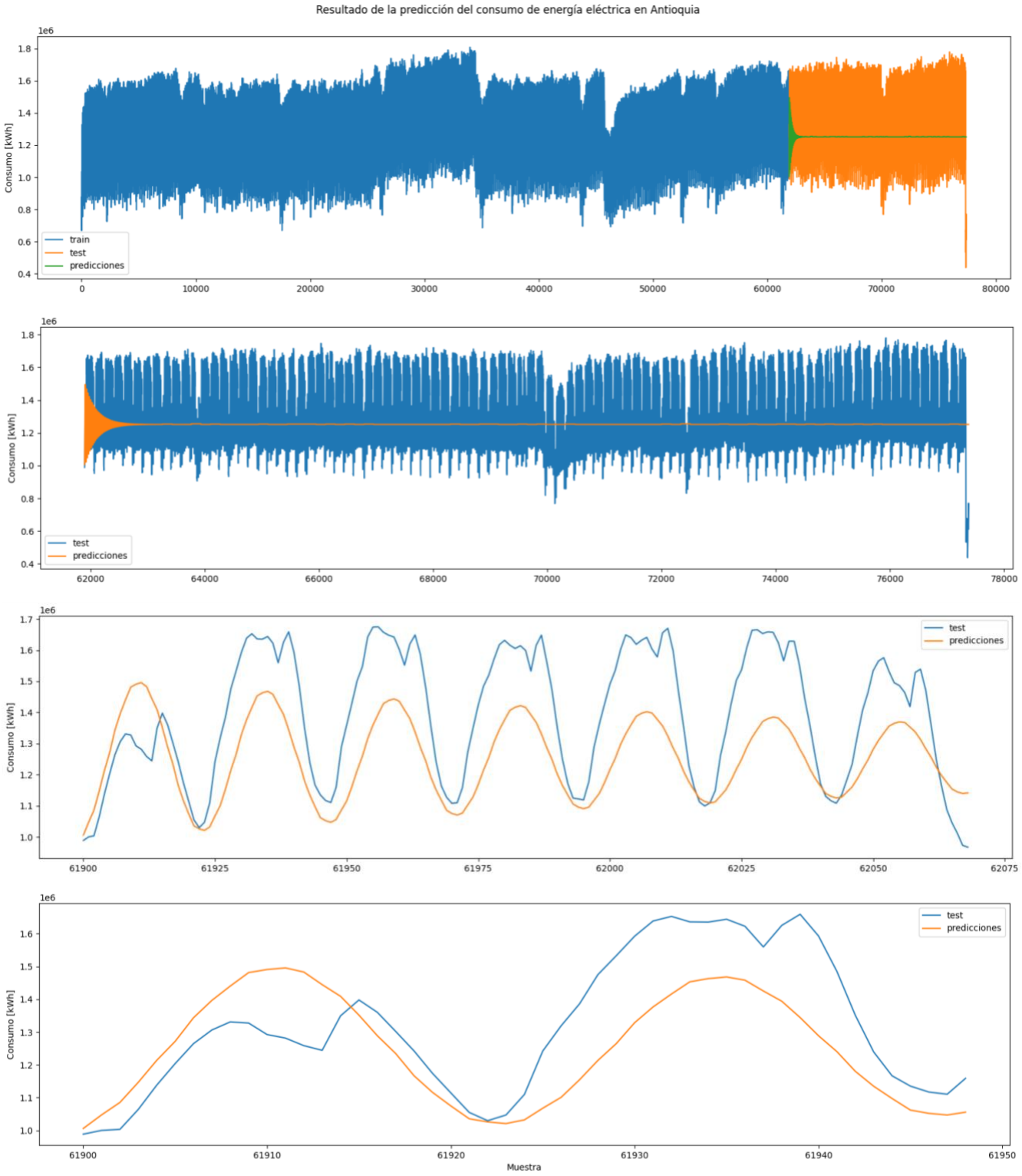


Figura 25: Resultado de la predicción de la demanda de energía eléctrica en Antioquia del modelo ARIMA con ajuste de hiperparámetros con variable exógena del tipo de semana.

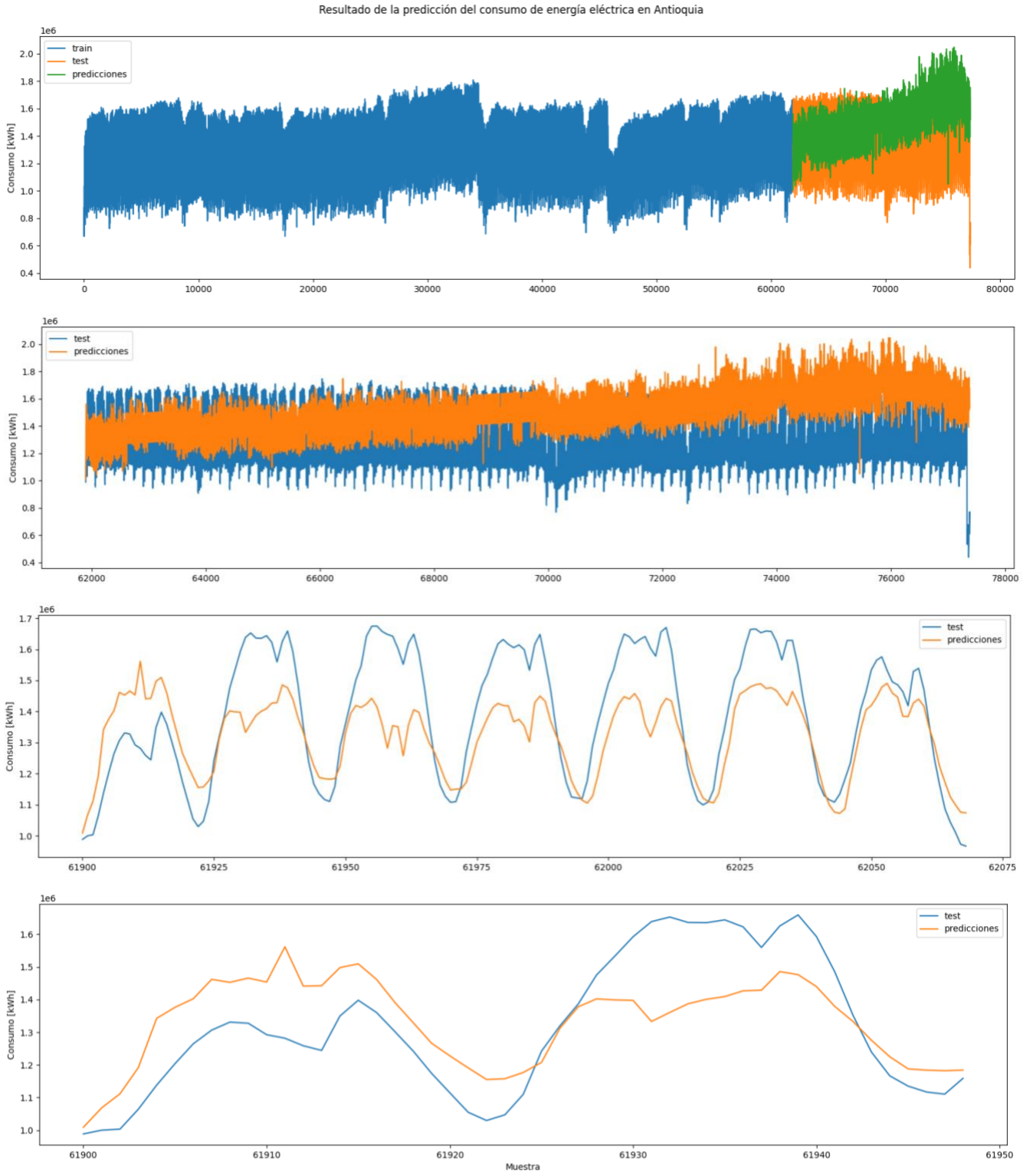


Figura 26: Resultado de la predicción de la demanda de energía eléctrica en Antioquia del modelo SARIMAX con ajuste de hiperparámetros con variable exógena de temperatura promedio.

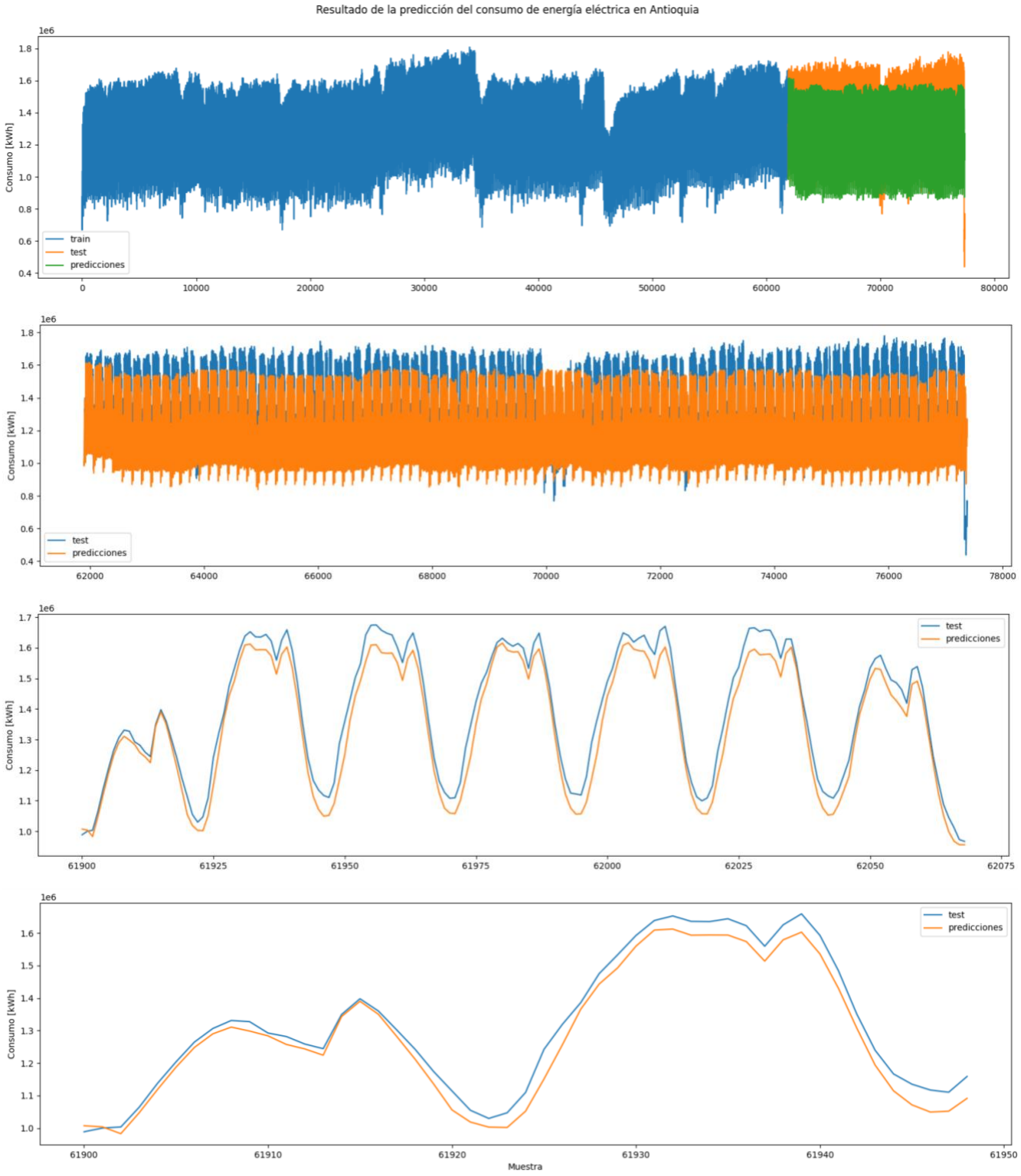


Figura 27: Resultado de la predicción de la demanda de energía eléctrica en Antioquia del modelo Random Forest Autoregressive con ajuste de hiperparámetros con variable exógena del tipo de día.

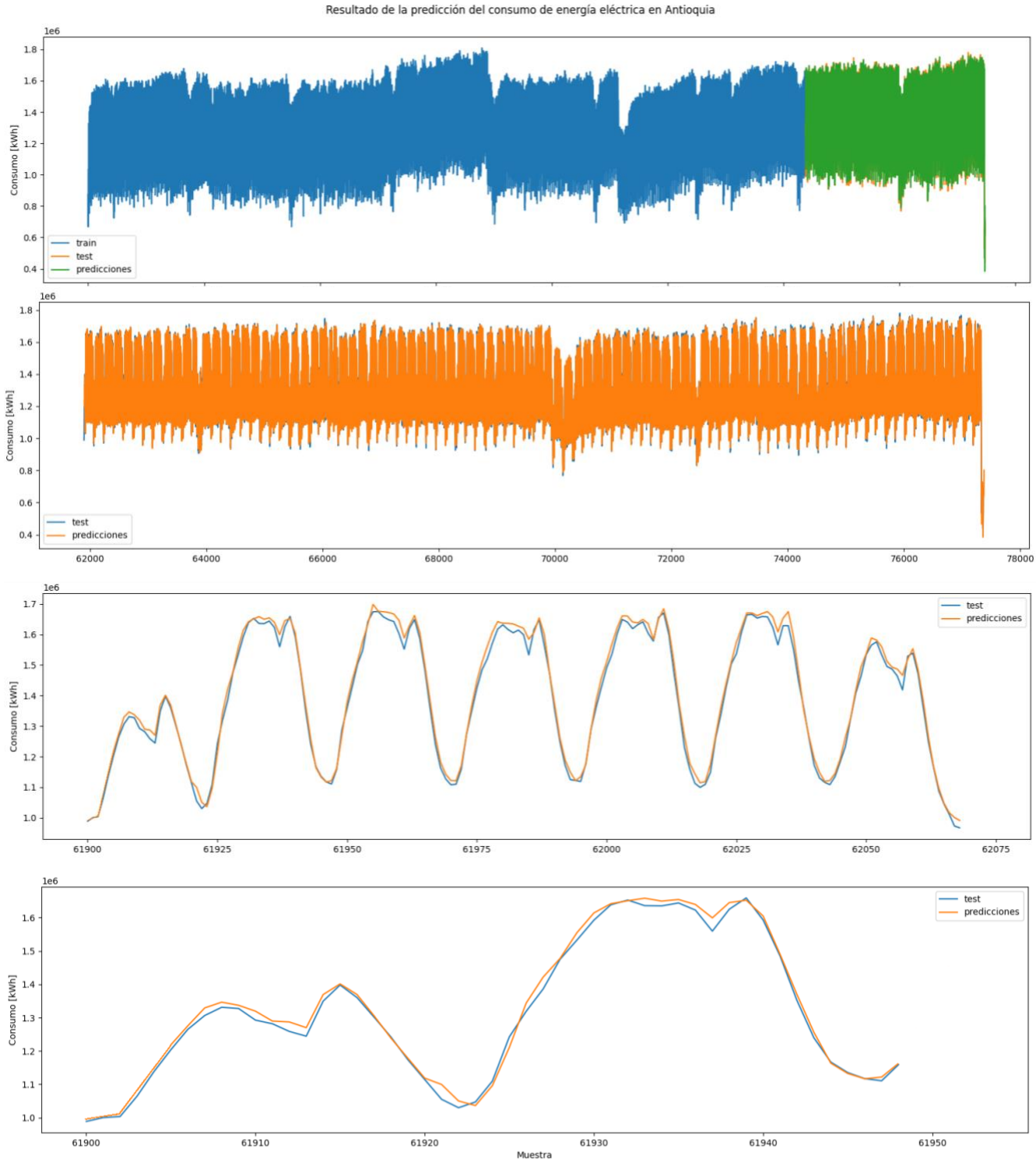


Figura 28: Resultado de la predicción de la demanda de energía eléctrica en Antioquia del modelo RNN LSTM con ajuste de hiperparámetros con variable exógena del tipo de día.

En las Figura 25, Figura 26, Figura 27 y Figura 28, se presentan 4 sub gráficos, los cuales, representan lo siguiente, el primer sub grafico es la visualización del resultado del modelos para todo el conjunto de datos de test, incluyendo el conjunto de datos de train; El segundo sub grafico es la visualización del resultado del modelo para el conjunto de datos de test, haciendo énfasis solo en el conjunto de test; El tercer sub grafico es la visualización del resultado del modelo para el conjunto de datos de test, haciendo énfasis en la primer semana de la predicción; El cuarto sub grafico es la visualización del resultado del modelo para el conjunto de datos de test, haciendo énfasis en los primer dos días de la predicción.

En la Figura 25 donde se muestra la predicción de la demanda de energía eléctrica con un modelo ARIMA con ajuste de hiperparámetros, en este caso usando el tipo de semana como variable exógena, se puede apreciar que la predicción al inicio oscila intentando mantener la estacionalidad de los datos, pero luego de un cierto periodo de tiempo, la predicción se estabiliza en el valor medio de los datos, al centrarse en el horizonte de la predicción, que se muestra en el tercer y cuarto sub grafico se puede notar que existe una mejoría en las predicciones usando el modelo ARIMA respecto a lo que se puede apreciar en la Figura 20, que el modelo con el ajuste de hiperparámetros se ajuste de mejor manera a la estacionalidad de los datos, sin embargo, sigue sin poder ajustarse correctamente a la tendencia de los mismos.

En la Figura 26 donde se muestra la predicción de la demanda de energía eléctrica con un modelo SARIMAX con ajuste de hiperparámetros, en este caso usando la temperatura promedio como variable exógena, se puede apreciar que sucede algo similar a como sucede en la Figura 21, que luego de un cierto periodo de tiempo, el valor medio de la predicción comienza a aumentar en este caso a su valor medio, sin embargo a diferencia de este, el aumento que se presenta en el modelo SARIMAX con ajuste de hiperparámetros no sucede tan rápido y con tanta desviación como el modelo SARIMAX sin ajuste hiperparámetros, por lo cual, el ajuste de hiperparámetros ayuda a la estabilidad de las predicciones. Por otro lado, al centrarse en el tercer y cuarto sub grafico se puede notar que en este modelo la estacionalidad y la tendencia de los datos se ajustan mejor.

En la Figura 27 donde se muestra la predicción de la demanda de energía eléctrica con un modelo Random Forest Autorregresive con ajuste de hiperparámetros, en este caso usando el tipo de día como variable exógena, se puede notar que al igual a como sucede en la Figura 23, luego de cierto periodo de tiempo la predicción comienza a oscilar intentando mantener la estacionalidad de los datos, en este caso, no se puede establecer cosas entre estas, además de que como explicó anteriormente la predicción aparte de ajustarse a la estacionalidad de los datos, también logra ajustarse correctamente a los máximos y mínimos de la demanda de energía durante el día. El comportamiento del modelo el Random Forest Autorregresive con ajuste de hiperparámetros no presenta mucha mejoría gráficamente respecto al modelo Random Forest Autorregresive sin ajuste de hiperparámetros, inclusive en los valores expuestos en la Tabla 4 en comparación con los expuestos en la Tabla 2 desmejoran las métricas de desempeño.

En la Figura 28 donde se muestra la predicción de la demanda de energía eléctrica con un modelo RNN LSTM con ajuste de hiperparámetros, en este caso usando el tipo de semana como variable exógena, se puede notar que al igual a como sucede en la Figura 24, este tipo de modelo es sigue siendo el que mejores predicciones realiza de los datos, logrando mantener la estacionalidad y tendencia de los datos, logrando predicciones que se ajustan con exactitud tanto en el corto, mediano y largo plazo. En general todos los modelos RNN LSTM, con o sin ajuste de

hiperparámetros son los que mejores predicciones realizan de todos los modelos sus respectivas iteraciones.

7.3. Consideraciones de producción

Como bien se mencionó en la Sección 7.1 y la Sección 7.2, los mejores modelos para realizar la predicción del consumo de energía eléctrica en Antioquia, son los modelos de redes neuronales recurrentes con LSTM, debido a que fueron los modelos que mejores métricas de desempeño presentaron, ajustándose a los parámetros y valores definido para la selección del modelo, pero también es necesario tener en cuenta otros factores cuando estos son puestos en producción, tal como el tiempo de ejecución de entrenamiento y predicción, ya que, si un modelo tiene buenas métricas de desempeño, pero tardar mucho tiempo en realizar el proceso de entrenamiento, hará que el proceso para la predicción de los valores futuros sea demorado y costoso a nivel económico, por esto, se decidió tomar los tiempos de ejecución de los modelos, obteniendo los siguientes resultados:

- ARIMA: Tiempo de ejecución promedio 24s.
- SARIMAX: Tiempo de ejecución promedio 235s.
- Linear Regressor Autoregressive: Tiempo de ejecución promedio 97s.
- Random Forest Autoregressive: Tiempo de ejecución promedio 3500s.
- RNN LSTM: Tiempo de ejecución promedio 222s.

* Nota: Los tiempos promedios sacados tomados los tiempos de las diferentes iteraciones de cada tipo de modelo, estos modelos se ejecutaron en el Google Colab, en una maquina con un procesador Intel ® Xeon (R) CPU @ 2.20GHz, memoria RAM 12GB y disco duro de 107GB.

Según los resultados expuestos en la Sección 7.1 y la Sección 7.2, los mejores modelos fueron las RNN LSTM, seguido de Random Forest Autoregressive, Linear Regressor Autoregressive, SARIMAX y ARIMA, se puede notar que modelos como los presentados con Random Forest Autoregressive presentan tiempos excesivos de ejecución y si bien presentan métricas que compiten fuertemente con RNN LSTM, al tener tiempos tan altos de ejecución para el entrenamiento y predicción, son modelos que no son viables en producción. Ahora modelos como los SARIMAX, donde presentan tiempos similares a los modelos RNN LSTM, pero con métricas de desempeños bajas tampoco son viables en producción y si bien los modelos ARIMA presentan los tiempos más bajos de ejecución, las métricas de desempeño no están dentro del rango esperado. Por estos motivos, si se deseara utilizar algunos de estos modelos en producción los mejores candidatos son RNN LSTM y Linear Regressor Autoregressive, siendo las RNN LSTM los mejores modelos según las métricas de desempeño y tiempos de ejecución.

8. Conclusiones

Todos los modelos de pronóstico de la demanda usando redes neuronales recurrentes del tipo LSTM, fueron siempre lo que obtuvieron mejores métricas de desempeño en el pronóstico de la demanda de energía eléctrica, sin embargo es necesario realizar más pruebas probando otras arquitecturas de redes o con más variables exógenas dentro de los modelos para obtener mejores resultados si así se requiera.

Los modelos de estadística clásica como ARIMA y SARIMAX, fueron utilizados como fuente de partida para poder comprobar modelos de machine learning como son los modelos autorregresivos y redes neuronales recurrentes. En dichos modelos de estadística clásica se puede evidenciar, que si bien son modelos de fácil implementación y tiempo de ejecución bajos, se requiere de forma obligatoria de un ajuste de hiperparámetros, debido a que los modelos sin esta optimización presentan errores altos, lo cual se evidencia en las tablas de resultados y las figuras que presentan los pronósticos realizados, nos obstante, estos modelos pueden ser de utilidad para una primera iteración de problemas similares, ya que, estos modelos ayudan al entendimiento y descubrimiento de los comportamientos de tendencia y estacionalidad de los datos.

Es de vital importancia realizar un preprocesamiento de datos y un buen análisis de datos antes de realizar los modelos, en especial con series de tiempo, ya que este tipo de datos, presentan comportamientos especiales que solo pueden evidenciarse con un análisis especializado en estos, tal como se hizo con la descomposición estacional, autocorrelación, gráficos de estacionalidad, entre otros análisis exploratorio de los datos.

Referencias

[1] A. S. Torres Sánchez, especialización-analitica-monografia, https://github.com/astorress/especializacion-analitica-monografia

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

[3] DANE, "Resultado Censo Nacional De Población y Vivienda 2018", 2018, pág. 52.

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

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

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

[7] J. A Pérez Arenas, "Pronóstico de Demanda de Energía Eléctrica de Colombia Usando un Modelo Estadístico a Partir de la Metodología de Box - Jenkins", 2020, pág. 12.

[8] Sodapy. Sodapy. https://github.com/xmunoz/sodapy

[9] Datos Abiertos. https://www.datos.gov.co/

[10] R. H. Shumway, D.S. Stoffer, "Time Series Analysis and Its Applications With R Examples", Ed Blue Printing, 2016.

[11] L. Breiman, "Random Forests in Machine Learning", Kluwer Academic Publishers, 2001, pág. 5-32.

[12] M. Rivera, "Introducción a Redes Neuronales Recurrentes (RNN)". http://personal.cimat.mx:8181/~mrivera/cursos/aprendizaje_profundo/RNN_LTSM/introduccion_rnn.html

[12] Digital Guide Ionos, "ISO 8601 expresiones temporales inequívocas". https://www.ionos.es/digitalguide/paginas-web/desarrollo-web/iso-8601/

[13] SQLShack, "Revisión del operador relacional y descripción general de pivot y unpivot estático y dinámico de SQL". https://www.sqlshack.com/es/revision-del-operador-relacional-y-descripcion-general-de-pivot-y-unpivot-estatico-y-dinamico-de-sql/

[14] P. Newbold, W. L. Carlson, B. Thorne, "Estadística para administración y economía", Ed Pearson Prentice Hall, 2008, pág. 777.

[15] J. A. Rodrigo, J. Escobar, "Skforescat: forecasting series temporales con Python y Scikit-learn", 2021. https://cienciadedatos.net/documentos/py27-forecasting-series-temporales-python-scikitlearn