**Logotipo

Descripción generada automáticamente**

**Predicción de la demanda de gas natural de las Centrales Térmicas en Colombia para el periodo 2024-2030**

Jesús Alberto Flórez Lizarazo

Lucero Zárate Herrera

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

Asesor  
Nombres completos, Título académico más alto

Universidad de Antioquia  
Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2024

| **Cita** | (Flórez Lizarazo & Zárate Herrera, 2024) |
| --- | --- |
| **Referencia**  **Estilo APA 7 (2020)** | Flórez Lizarazo , J. A., & Zárate Herrera, L. (2024). *Predicción de la demanda de gas natural de las centrales térmicas en Colombia para el periodo 2024-2023*. Universidad de Antioquia, Medellín, Colombia. |

**** 

Especialización en Analítica y Ciencia de Datos, CohorteVII.

Centro de Investigación Ambientales y de Ingeniería (CIA).

|  | Diagrama  Descripción generada automáticamente con confianza media |
| --- | --- |

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.

**Dedicatoria**

Texto de dedicatoria centrado.

**Agradecimientos**

Texto de agradecimientos centrado.

**Tabla de contenido**

[Resumen 8](#_heading=h.3rdcrjn)

[Abstract 9](#_heading=h.26in1rg)

[1. Descripción del problema 10](#_heading=h.35nkun2)

[1.1. Problema de negocio 10](#_heading=h.1ksv4uv)

[1.2. Aproximación desde la analítica de datos 10](#_heading=h.44sinio)

[1.3. Origen de los datos 10](#_heading=h.z337ya)

[1.4. Métricas de desempeño 10](#_heading=h.3j2qqm3)

[2. Objetivos 11](#_heading=h.4i7ojhp)

[2.1. Objetivo general 11](#_heading=h.2xcytpi)

[2.2. Objetivos específicos 11](#_heading=h.3whwml4)

[3. Datos 12](#_heading=h.2bn6wsx)

[3.1. Datos originales 12](#_heading=h.qsh70q)

[3.2. Datsets 12](#_heading=h.3as4poj)

[3.3. Analítica descriptiva 13](#_heading=h.1pxezwc)

[4. Conclusiones 17](#_heading=h.kgcv8k)

[Referencias 17](#_heading=h.1jlao46)

[Anexos 17](#_heading=h.43ky6rz)

[Anexo 1. XXXX 18](#_heading=h.xvir7l)

[Anexo 2. 18](#_heading=h.1baon6m)

[Anexo 3. 18](#_heading=h.3vac5uf)

**Lista de tablas**

[**Tabla 1**](#_heading=h.2u6wntf)  [21](#_heading=h.3tbugp1)

**Lista de figuras**

[**Figura 1** 22](#_heading=h.2r0uhxc)

[**Figura 2** 22](#_heading=h.1664s55)

**Siglas, acrónimos y abreviaturas**

**Esp.** Especialista

**LSTM**  Long Short-Term Memory

**MAPE** Mean Absolute Percentage Error / Porcentaje de Error Absoluto Medio

**MBTU** Millón de British Thermal Units

**MSE** Mean Squared Error / Error Cuadrático Medio

**RNN** Recurrent Neural Network /Serie neuronal Recurrente

**SARIMAX** Seasonal AutoRegressive Integrated Moving Average with eXogenous variables

**TS** Time Series /Serie de Tiempo

**UdeA** Universidad de Antioquia

# Resumen

Una correcta planeación de la demanda de gas natural en Colombia, es crucial para la toma de decisiones a nivel político-administrativo, sin embargo, a diferencia de otros sectores, el mercado termoeléctrico se ve afectado por cantidad de variables exógenas que convierten en todo un reto, realizar una adecuada predicción de la demanda de gas del sector.

El presente proyecto tiene como objetivo desarrollar una predicción de la demanda de gas natural de las centrales térmicas que sea coherente con la realidad del mercado; a través de la implementación y comparación de diferentes modelos predictivos como técnicas avanzadas de series temporales, redes neuronales recurrentes y SARIMAX. Para este análisis se hará uso de diferentes fuentes informativas oficiales como el administrador del mercado de energía mayorista en Colombia (XM).

Los datos y modelos aplicados pueden consultarse en los siguientes repositorios de GitHub [1]:

[**https://github.com/LuceroZaHe**](https://github.com/LuceroZaHe), [**https://github.com/Jflorez05**](https://github.com/Jflorez05)

*Palabras clave*: gas natural, termoeléctrica, demanda, XM, Colombia, .

# Abstract

A correct planning of natural gas demand in Colombia is crucial for political-administrative decision-making. However, unlike other sectors, the thermoelectric market is affected by many exogenous variables, making it a significant challenge to accurately predict the sector's gas demand.

This project aims to achieve a prediction of natural gas demand for thermal power plants that is consistent with the market reality. This will be done through the implementation and comparison of different predictive models such as advanced time series techniques, recurrent neural networks, and SARIMAX. The analysis will be based on data from official sources, such as the Wholesale Energy Market Administrator in Colombia (XM).

The data and models applied can be consulted in the following GitHub repositories: [**https://github.com/LuceroZaHe**](https://github.com/LuceroZaHe), [**https://github.com/Jflorez05**](https://github.com/Jflorez05)

*Keywords***:** natural gas, thermoelectric, demand, XM, Colombia

# Descripción del problema

## Problema de negocio

Actualmente no se cuenta con una herramienta que permita predecir con un alto nivel de certeza, la demanda de gas natural por parte del sector térmico en el país. El pronóstico de la demanda de gas natural es fundamental para la toma de decisiones operativas y estratégicas en el mercado energético del país.

Las predicciones permiten determinar de forma previa la carencia del energético en el sistema y ayudan a impedir el racionamientos de energía, desde otro punto de vista permite comprometer a plantas de generación a futuro con suministro de energía para cubrir la demanda de energía en las zonas necesarias del país. [2]

Se generará un modelo de series temporales para el análisis de datos históricos de demanda de gas en centrales térmicas en Colombia. Los datos provienen del operador del mercado mayorista de energía XM. Las métricas del modelo serán su precisión y la criticidad de sus errores.

## Aproximación desde la analítica de datos

Para predecir la demanda de gas por parte de las termoeléctricas, se desarrollarán modelos de series temporales. Dado que la información de demanda de gas se estima diariamente, los modelos elegidos para su análisis son SARIMAX y Redes Neuronales Recurrentes (RNN). La naturaleza secuencial de los datos de demanda de gas, con observaciones registradas en intervalos de tiempo regulares, se presta bien al análisis de series temporales. SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous factors) es particularmente útil para capturar patrones estacionales y tendencias a largo plazo, además de incorporar variables exógenas que influyen en la demanda [3]. Por otro lado, las RNN, y específicamente las variantes como LSTM (Long Short-Term Memory), son capaces de manejar dependencias a largo plazo y no linealidades en los datos temporales, lo que las hace ideales para capturar dinámicas complejas [4]. Ambas metodologías permiten modelar y prever de manera efectiva la evolución de la demanda de gas, aprovechando las capacidades analíticas avanzadas y técnicas de aprendizaje automático para proporcionar predicciones precisas y confiables.

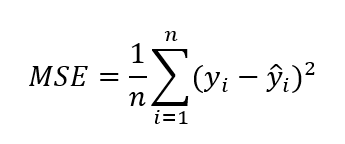
## Origen de los datos

Los datos usados en este proyecto se extraen de la API oficial de XM [6], repositorio público y de libre acceso que contiene información transaccional del mercado mayorista de energía. El dataset inicial contiene información asociada al tipo de combustible, agente térmico y consumo horario de combustible para la generación de energía eléctrica. El periodo de extracción inicial fue enero de 2018 a diciembre de 2023..

## Métricas de desempeño

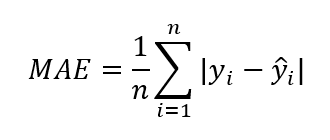
La comparación y evaluación de la precisión y rendimiento global de los modelos predictivos ejecutados se realizará por medio de métricas como:

*Error Cuadrático Medio (MSE):* Promedio del cuadrado de los errores entre los valores predichos y los valores reales.



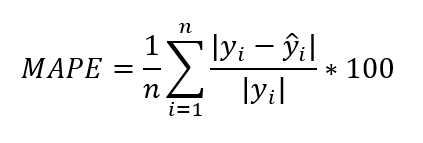
Esta métrica, penaliza severamente los errores grandes, una característica útil cuando se requiere minimizar grandes desviaciones en el resultado, puesto que, un error elevado en la predicción de la demanda puede tener un impacto significativo en la toma de decisiones que se deriva de ello.

*Error Absoluto Medio (MAE)*: Promedio de los errores absolutos.



Este indicador ofrece una medida intuitiva y lineal del error. Al ser menos sensible a outliers que el MSE, esta métrica es ideal para comprender los errores promedio del modelo sin la influencia desproporcionada de grandes desviaciones.

*Porcentaje de Error Absoluto Medio (MAPE)*: Calcula el promedio de los errores porcentuales absolutos.



Este parámetro ayuda a entender el rendimiento del modelo en términos porcentuales, facilitando la comparación de las precisiones de ambos modelos.

Los anteriores parámetros son la base para entender a profundidad las limitaciones y capacidades de cada modelo, pues mientras mientras MSE ayuda a controlar grandes errores, MAE ofrece una visión clara del error promedio y MAPE permite comparaciones relativas en diferentes escalas.

Por otra parte, se pretende evaluar el nivel esperado de satisfacción de la demanda de gas acorde con las cantidades de gas declaradas para los años venideros; la reducción del costo de exceso y defecto, esto es, el costo de adquirir mayores o menores cantidades que las realmente demandadas en determinado periodo; y el posible ahorro en costos de adquisición del gas al ejecutar compras en momentos óptimos, es decir, en condiciones de normalidad con precios más bajos que en un contexto de probable de escasez.

# Objetivos

## Objetivo general

Predecir las cantidades de gas natural demandado por las centrales termoeléctricas del país en el corto y mediano plazo haciendo uso de herramientas de análisis de datos y modelos predictivos

## Objetivos específicos

* Realizar un análisis exploratorio de datos para identificar tendencias y patrones en el consumo de gas natural.
* Seleccionar y aplicar modelos de series temporales adecuados para predecir la demanda futura de gas natural y validar sus resultados utilizando técnicas de validación cruzada.
* Evaluar la precisión de los pronósticos y proponer estrategias para mejorar la precisión de las predicciones.
* Generar pronósticos de demanda de gas natural para el período 2024-2030, considerando diferentes escenarios.

# Datos

## Datos originales

Los datos usados en este proyecto se extraen de la API oficial de XM [4], que contiene información transaccional del mercado mayorista de energía. El dataset inicial contiene información asociada al tipo de combustible, agente térmico y consumo horario de combustible para la generación.

Los datos que se extraen con la API son valores que los generadores térmicos deben reportar diariamente según resolución CREG 084 de 2005 [2] el consumo de combustible en MBTU, el periodo de extracción es del 2018 a 2023, los campos de la fuente de datos son:

* **Id**: Recurso
* **Values\_Name**: Indica el tipo de recurso con el que se produce la energía.
* **Values\_code:** Identificador del agente térmico productor de energía.
* **Values\_Hour0(n)\*:** Valor de consumo por un agente en una hora del día.
* **Date:** Fecha en la que el agente consume el energético para producir energía.

\* Existen 24 columnas para el campo Values\_Hour, cada una para cada hora de un día particular.

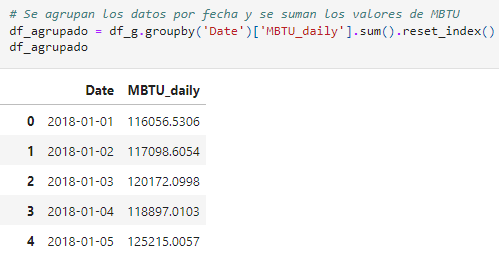
En el rango de fechas descargado para el estudio se obtiene un total de 59344 registros. Para acceder a la información lo único que se debe realizar es la instalación e importación de la librería “ pydataxm” los datos son públicos y regulados. [6]

## Datasets

A partir de la fuente de datos originales se consolidan los datos agrupados de la siguiente forma:

* Suma del valor de consumo total de MBTU de GAS consumido diario para cada fecha.
* Agrupar en cada una de las fechas el valor agregado de todos los agentes que producen energía con GAS. La figura 1 se muestra la última agrupación y formato de datos.

**Figura 1.** Dataset Agrupado

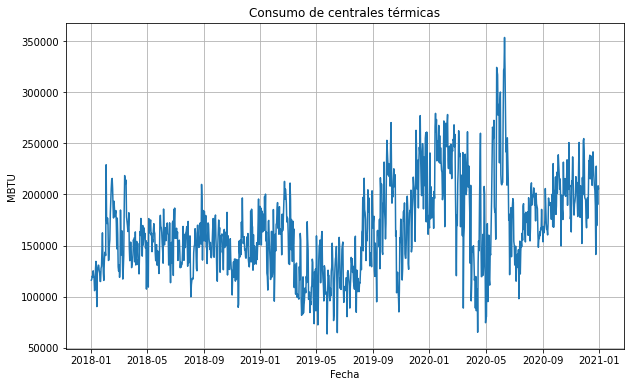


* Los modelos que se van a implementar con los modelos SARIMAX y redes neuronales recurrentes (LSTM), para el modelo SARIMAX se realiza una división de datos en proporción 90% entrenamiento y 10% de validación para predicción. En el caso de LSTM se realizan 2 secuencias de 2 datos consecutivos.

## Analítica descriptiva

Las series de tiempo pueden clasificarse en estocásticas y determinísticas y su análisis se basa en encontrar patrones en el comportamiento de los datos de acuerdo a un orden particular (tiempo). Como primer acercamiento para el análisis de comportamiento de los datos, se genera una visualización de la serie con información correspondiente al periodo 2018-2020.

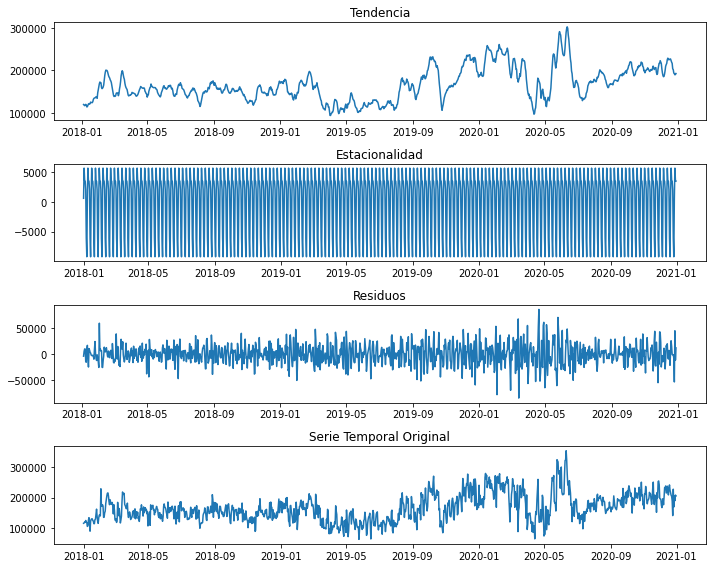
**Figura 2.** Gráfico de consumos 2018-2020



En la figura 2 se muestra la serie de tiempo resultante, no obstante, desde esta perspectiva, el comportamiento inicial de los datos no brinda mayor información sobre la serie temporal, lo que dificulta llegar a una conclusión clara al respecto.

Para el análisis exploratorio de los datos, el primer paso es realizar una descomposición de la tendencia de los datos, su estacionalidad y error. En la figura 3 se muestran los resultados de esta descomposición.

**Figura 3.** Gráficos de tendencia, estacionalidad y error de los datos



De acuerdo con la figura anterior se nota que la tendencia de la serie temporal no tiene una tendencia lineal, ni sigue un patrón particular que se pueda evidenciar en el periodo considerado. En cuanto a la estacionalidad de la serie de tiempo, se nota que existe un patrón estacional marcado en la serie temporal, esta es una buena señal para análisis posteriores. La última gráfica muestra los residuos o la componente estocástica temporal que tiene una variación irregular a lo largo del periodo elegido.

Con el fin de definir la estacionariedad de la serie de tiempo se realiza la prueba de Dickey&Fuller ampliada, esta es una prueba estadística utilizada para determinar si una serie temporal tiene una raíz unitaria, lo que implica que la serie es no estacionaria.

Hipotesis del test:

**Hipótesis nula (H0)**: La serie temporal tiene una raíz unitaria (es decir, no es estacionaria).

**Hipótesis alternativa (H1)**: La serie temporal no tiene una raíz unitaria (es decir, es estacionaria)

El test Dickey-Fuller aumentado se basa en la estimación del siguiente modelo de regresión:

Δ𝑦𝑡 = 𝛼 + 𝛽𝑡 + 𝛾𝑦𝑡−1 + 𝛿1Δ𝑦𝑡−1 + 𝛿2Δ𝑦𝑡−2 + … + 𝛿𝑝Δ𝑦𝑡−𝑝 + 𝜖𝑡

Donde:

Δ: denota la diferencia de primer orden.

𝑦𝑡: es la serie temporal en el tiempo

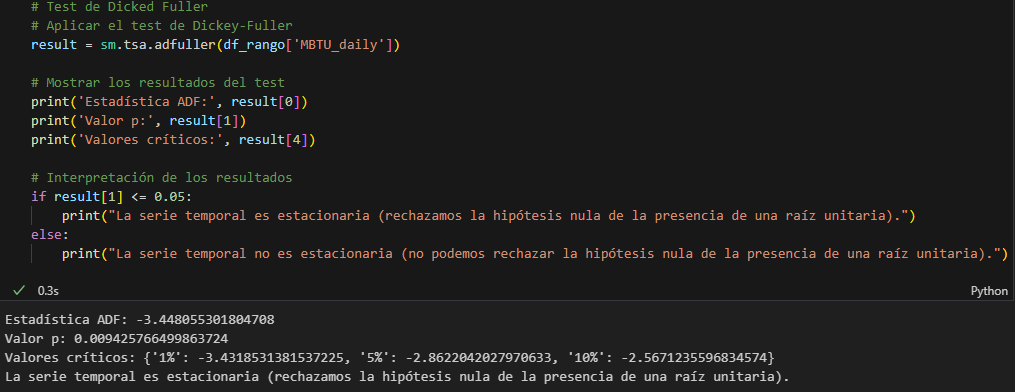
𝑡: es una tendencia temporal determinista (opcional).

𝜖𝑡: es un término de error blanco.

La estacionariedad es una propiedad crucial en el análisis de series temporales, ya que muchas técnicas de modelado asumen que las series son estacionarias. En caso de no contar con una serie estacionaria se deben aplicar técnicas de transformación (generalmente suavizados) para obtener una serie temporal estacionaria.

En el siguiente código se muestra el test aplicado a los datos considerados.

**Figura 4.** Aplicación del método Dickey&Fuller



En los resultados se muestra que se rechaza la hipótesis nula de la presencia de una raíz unitaria, con lo cual se puede concluir que la serie de tiempo cumple con un nivel de confianza del 95% , lo que significa que es una serie estacionaria.

# Conclusiones

# Referencias

[1] Zárate Herrera, L, & Flórez Lizarazo , J. A. especialización-analitica-monografia, [**https://github.com/LuceroZaHe**](https://github.com/LuceroZaHe), [**https://github.com/Jflorez05**](https://github.com/Jflorez05)

[2] D. Garcia Torres, 2014, "Dinámica de la generación termoeléctrica en Colombia: pronóstico a través de series con memoria a largo plazo", disponible en línea:<https://manglar.uninorte.edu.co/bitstream/handle/10584/11045/73207919.pdf?sequence=1&isAllowed=y> [Acceso: abril 25, 2024].

[3] Goodfellow, I., Bengio, Y., & Courville, A. (2016). "Deep Learning". MIT Press. This book covers the fundamentals of deep learning, including RNNs and their applications in time series forecasting

[4] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). "Time Series Analysis: Forecasting and Control". This book is a seminal reference in time series analysis, detailing ARIMA models and their extensions, including SARIMAX, and discussing their applications in various fields.

.[5] Atwan, T. (2022) Time Series Analysis with Python Cookbook. 1st edn. Packt Publishing

[6] API para acceder a base de datos de mercado de energía mayorista, "API\_XM," GitHub, 2022. [Accedido el 25 de abril de 2024]. Disponible en:<https://github.com/EquipoAnaliticaXM/API_XM>

# Anexos

**Anexo 1**.

## Anexo 2.

## 

## Anexo 3.