Primera entrega trabajo de grado

**Proyecto analítico de predicción de costo unitario de energía en Colombia en el sector residencial**

Daniel Felipe Rodríguez, Fabian Ricardo Martínez, Diana Katherin Ibáñez, Paola Andrea Castro

Juan Pablo Pájaro Hernández (Tutor técnico)

Katherine Cabrera Rodríguez (Tutora Negocio)



Maestría en analítica para la inteligencia de Negocios

*Pontificia Universidad Javeriana, Bogotá, Colombia*

**Tabla de contenido**

[1. Descripción del Mercado Energético 3](#_Toc174302922)

[2. Estado del arte 5](#_Toc174302923)

[3. Determinar los Objetivos del Negocio 7](#_Toc174302924)

[**3.1.** Objetivo General 7](#_Toc174302925)

[3.2. Objetivos específicos 7](#_Toc174302926)

[4. Contexto 8](#_Toc174302927)

[4.1. Modelo analítico 8](#_Toc174302928)

[4.2. Tablero de visualización 9](#_Toc174302929)

[5. Problema de Negocio 10](#_Toc174302930)

[6. Tarea analítica 10](#_Toc174302931)

[7. Flujo de datos 11](#_Toc174302932)

[7.1. Consideraciones para la construcción del modelo 11](#_Toc174302933)

[7.2. Determinar y obtener los datos necesarios 12](#_Toc174302934)

[7.2.1. Componentes 12](#_Toc174302935)

[7.2.2. Variables Exógenas 12](#_Toc174302936)

[7.3. Preprocesamiento de datos 13](#_Toc174302937)

[7.3.1. Selección de datos 13](#_Toc174302938)

[**7.3.2.** Integración y limpieza de datos 13](#_Toc174302939)

[**7.3.3.** Validación de datos 13](#_Toc174302940)

[**7.3.4.** Formato de datos 13](#_Toc174302941)

[**7.4.** Entendimiento de los datos 14](#_Toc174302942)

[**7.4.1.** Análisis Exploratorio de los datos 14](#_Toc174302943)

[**7.4.2.** Análisis de Correlación 14](#_Toc174302944)

[**7.4.3.** Pruebas estadísticas en modelos de series temporales 14](#_Toc174302945)

[**7.5.** Selección del modelo 14](#_Toc174302946)

[7.5.1. ARIMA: 14](#_Toc174302947)

[7.5.2. Long Short-Term Memory (LTSM): 14](#_Toc174302948)

[7.5.3. XGBoost: 15](#_Toc174302949)

[7.5.4. Temporal Convolutional Networks (TCN): 15](#_Toc174302950)

[7.5.5. Nonlinear Autoregressive with Exogenous Inputs (NARX): 15](#_Toc174302951)

[**7.6.** División de datos en entrenamiento y prueba. 15](#_Toc174302952)

[**7.7.** Evaluación del modelo 16](#_Toc174302953)

[7.7.1. Métrica de evaluación 16](#_Toc174302954)

[Referencias 19](#_Toc174302955)

# Descripción del Mercado Energético

Las principales fuentes de energía en Colombia son la energía hidráulica, el petróleo, el carbón y el gas natural. De igual forma, las actividades energéticas fundamentales incluyen la generación, transmisión, distribución y comercialización. Según la regulación de la Comisión de Regulación de Energía y Gas (CREG), se establece la medición del costo unitario de energía, clasificada según los niveles de tensión, los cuales se determinan por la potencia y el tipo de consumo mensual. En el mercado no regulado (consumo > 0.1 MW), los precios son fijados por el comprador y el vendedor. En el mercado regulado (consumo >50 MW), que representa el 65% de la demanda total de Colombia, los precios son fijados por la regulación.

Según informes recientes del Administrador del Mercado de Energía Mayorista (Corficolombiana, 2021), la generación de energía sigue siendo predominantemente hidráulica, con un 70%, mientras que el 27% proviene de fuentes térmicas, incluyendo gas y carbón. Colombia cuenta con 44 agentes generadores, de los cuales seis (ENDESA - Emgesa, EPM, ISAGEN, GECELCA, CELSIA y AES) son los principales. ISAGEN posee aproximadamente el 80% del sistema de transmisión nacional, que incluye redes de 550 kV y 220 kV de libre acceso.

Todas las empresas distribuidoras también actúan como comercializadoras en el mercado regulado dentro de sus respectivas zonas. Actualmente, existen cuatro zonas geográficas para el control, seguimiento y distribución de la energía en el país.

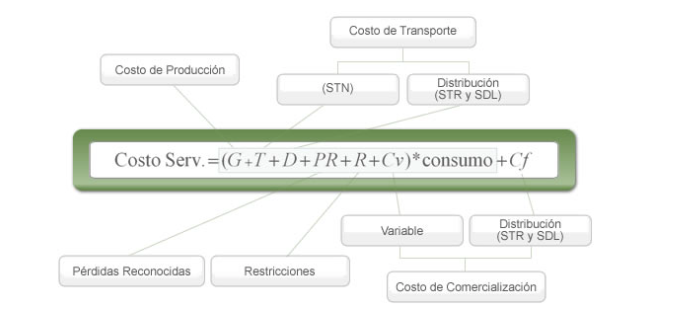
Colombia es un país con una rica y diversa variedad de recursos naturales, lo que representa un gran potencial para la generación y suministro de energía eléctrica en todo el territorio. Durante los últimos años, el país ha enfrentado el reto de revisar el costo unitario de la energía con el objetivo de lograr eficiencia en el consumo y atraer la inversión extranjera. Desde la década de los 90, se han creado organismos de administración y control del sector energético, como la Comisión de Regulación de Energía y Gas (CREG), la Unidad de Planeación Minero-Energética (UPME) y la Unidad de Información Minero Energética (UIME), que actúan como las principales entidades de vigilancia en el sector.

El costo unitario de la energía eléctrica en Colombia es crucial porque impacta directamente a los consumidores, tanto en los hogares como en las empresas que dependen de este servicio en su vida diaria. Afecta el presupuesto familiar y la calidad de vida. Además, el precio de la energía es un componente importante en la inflación del país, por lo que un incremento en su costo puede tener impactos significativos en la estabilidad económica.

Para las empresas, la energía es un recurso vital para su operación y desarrollo. Las variaciones en los costos energéticos pueden representar ventajas o desventajas en la competitividad nacional e internacional, particularmente en lo que afecta la exportación y la atracción de inversión extranjera.

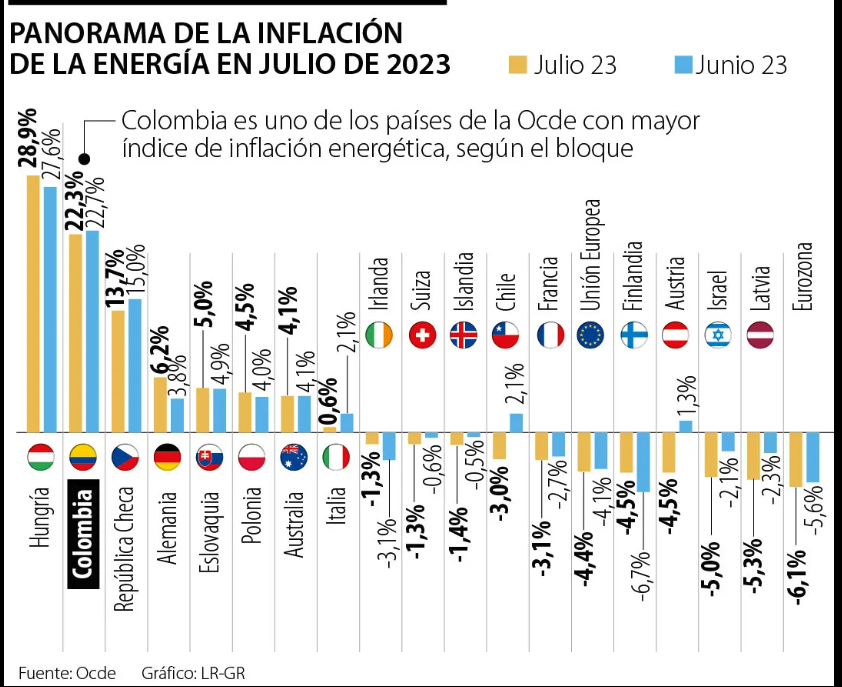
El costo unitario de la energía también tiene un gran impacto en el ámbito de la sostenibilidad. Por lo tanto, el uso eficiente de los recursos es primordial para el desarrollo de proyectos que busquen fuentes de energía más económicas, que beneficien financieramente al país y promuevan mejores prácticas de cuidado del medio ambiente.

**Componentes Res. CREG 119 de 2007**



**Panorama Internacional**

Figura 1 Panorama de la inflación de la energía 2023



*Fuente: OCDE*

Colombia ocupa el segundo lugar en la OCDE en inflación energética, con un 22,3%, debido principalmente al fenómeno de El Niño, que reduce la generación de energía hidroeléctrica y aumenta la dependencia de fuentes térmicas más costosas. Hungría lidera con una inflación del 28,9%, mientras que países como España tienen la menor inflación energética debido a las matrices de energía más diversificadas y menos dependientes de combustibles fósiles. (Daza, 2023)

# Estado del arte

De acuerdo con el artículo (Hernández N., 2005), de los modelos más utilizados para predicciones de costo de energía, se encuentra el modelo ARIMA (Autoregressive Integrated Moving Average), que debido a su capacidad de capturar tanto la tendencia como la estacionalidad de los datos se utiliza ampliamente para el análisis y la predicción de series temporales. A través de la combinación de autorregresión (AR), media móvil (MA) e integración (I), el modelo ARIMA puede ajustarse a una gran variedad de patrones en los datos temporales, lo que se convierte en una opción flexible y robusta para la predicción en campos como la economía, la meteorología y el control de calidad, entre otros.

El modelo ARIMA, aunque es robusto y ampliamente utilizado en series de tiempo, tiende a mostrar una disminución en su poder predictivo a medida que se extiende el horizonte de predicción. Así como lo expresan en diversos artículos, textos, trabajos de grado, entre otros, donde al realizar comparaciones entre los diferentes modelos en métricas como el MAPE, RMSE, entre otros, se evidencia como otros modelos obtienen mejores resultados. Como por ejemplo (Alberto Muñoz, 2017) en su artículo lo expresan.

Por otro lado, (Chaparro, Martínez, & Trujillo, 2012) emplearon una metodología rigurosa para la predicción de la demanda de energía en el Sistema Interconectado Nacional (SIN) de Colombia. La serie temporal fue preprocesada utilizando la Transformada Discreta Wavelet (DWT) para eliminar tendencias y estacionalidades. Posteriormente, se implementó un modelo neuronal autorregresivo no lineal (NAR) con los siguientes hiperparámetros: 3 capas ocultas, 4 neuronas por capa, una tasa de aprendizaje de 0.01, y un número máximo de 500 épocas de entrenamiento, utilizando el algoritmo de Levenberg–Marquardt.

La serie temporal se dividió en conjuntos in-sample (entrenamiento, validación) y out-of-sample (prueba), donde la selección de entradas pasadas se realizó mediante un autocorrelograma. Los autores eligieron una función de activación sigmoidal logística para las capas ocultas y una función lineal para la capa de salida, optimizando la red para minimizar el error de predicción. Las métricas de evaluación utilizadas fueron MAD, RMSE y el error porcentual. Concluyen así que, el uso de MRA Wavelet mejora sustancialmente los resultados en las métricas de los modelos.

Adicionalmente, en el artículo de (Adriana P. Agudelo, 2015) (Adriana P. Agudelo, 2015)se desarrolló un modelo NARX para predecir el precio de la electricidad en la bolsa de Colombia. El modelo utilizó datos históricos del precio de la electricidad y consideró 4 variables exógenas: 3 que dependen de las condiciones hidrologías (la relación entre generación hidráulica-térmica, probabilidad del Fenómeno de El Niño y volumen útil diario energía), y la demanda de energía eléctrica junto con 4 retardos de la variable de interés. Se llevó a cabo un proceso de preprocesamiento de datos para limpiar y preparar las series temporales, y se seleccionaron 2 capas (una oculta con función de activación Tansig y una de salida con función de activación lineal), 12 neuronas y 4 rezagos para el modelo final. Los hiperparámetros fueron optimizados minimizando el RMSE y maximizando el R2, logrando un RMSE de 12,93 y un R2 del 96% en el conjunto de prueba, lo que demuestra un alto nivel de precisión en la predicción.

Por otro lado, (Cárdenas, 2021), establece que, en el desarrollo de modelos predictivos para estimar el precio de la energía en Colombia, la selección adecuada de variables y la sintonización de hiperparámetros son esenciales para lograr pronósticos precisos. En el trabajo se resalta que variables exógenas como el precio de oferta promedio, la demanda del Sistema Interconectado Nacional (SIN), las reservas hídricas y la tasa de cambio representativa del mercado (TRM) se han identificado como altamente correlacionadas con el comportamiento del precio de la energía, justificando su inclusión en los modelos predictivos.

Para mejorar la precisión, Cárdenas optó por una combinación de técnicas autoregresivas como ARIMA y redes neuronales autorregresivas (NNAR), integrando regresores externos para potenciar la exactitud del pronóstico. En particular, el uso de ARIMA para predecir los regresores externos y NNAR para modelar el precio de la energía ha sido una estrategia efectiva. En la sintonización de hiperparámetros, en el modelo de redes neuronales se probaron 280 combinaciones variando los retrasos entre 1 y 20 y el número de nodos entre 1 y 14, seleccionándose finalmente una red con 10 retrasos y 6 nodos en la capa oculta. En el modelo ARIMA, se evaluaron 150 escenarios variando los parámetros 𝑝p entre 1 y 10, 𝑑d entre 0 y 2, y 𝑞 q entre 1 y 5, seleccionándose el modelo ARIMA (5,1,2). La optimización de estos parámetros, junto con el uso de métricas como RMSE y MAPE, permitió ajustar los modelos para capturar mejor la dinámica del precio de la energía. Los resultados muestran que, aunque los modelos son capaces de identificar tendencias y ciclos de precios, existen desafíos en la precisión de las estimaciones puntuales.

Según el estudio realizado por la universidad de los Andes. (Luis Fernando González Pérez, 2022) respecto al modelo implementado para la predicción de la demandada eléctrica en Colombia se consideran utilizar tres técnicas analíticas ARIMA, SARIMAX ,PROHET y LMST, dentro de la limpieza de datos que contemplan los autores del estudio, se encuentran pruebas iterativas de las variables con y sin transformación Log, BoxCox,normalización, eliminación de outliers e imputación de datos faltantes, así mismo, el estudio contempla el análisis de la matriz de correlaciones sin embargo se concluye que todas las variables contemplados en el modelo son relevantes para su desarrollo, por tanto continúan las 11 variables mencionadas en el documento, estas son demanda, vertimentos, aporte Gwh, reservas, generación Fncer M, generación Fncer FM, generación hidro M, fidro FM, generación termo M, generación termo FM, precio en la bolsa. Dentro de los análisis mencionados se realizaron análisis de componentes principales, enmarcando grupos de clústeres en periodos de tiempo de Lunes a Viernes, Sábados, Domingos y festivos. Como conclusión SARIMAX tuvo un mejor desempeño con variables exógenas. ARIMA con variables exógenas tiene una precisión de 81.54%, adicionalmente, LSTM mostró mejores resultados en MSE y MaxError.

De acuerdo al documento de investigación realizado en la unión europea (Laura Melgar García, s.f.), respecto a técnicas Big Data para la Predicción de la Demanda y Precio Eléctrico en España 2023, se ha considerado evaluar los componentes del precio eléctrico en Europa de manera individual y explorar modelos analíticos tanto de Deep Learning y Modelos de Machine tradicionales, los principales ajustes que realizaron en el tratamiento de datos es los cambios de horarios en las series temporales los cuales se van agregando incrementalmente al transcurrir los periodos de estudio. Las métricas que se utilizaron para evaluar el desempeño fueron el MAPE y MAE de los cuales se consideró que de los modelos aplico el mejor desempeño para predicción de precio fue el StreamWNN (Stream Weighted Nearest Neighbors) sin embargo de acuerdo con lo investigado éste, indican que funciona mejor para datos seleccionados en tiempo real, cuando los modelos no usan información en tiempo real se consideran mejor los modelos LSTM (Memoria a Corto Plazo).

El algoritmo XGBoost también es un candidato para la exploración de los modelos a implementar, el cual consiste en entrenar iterativamente una serie de predicciones secuenciales, donde cada árbol intenta corregir los errores cometidos por los árboles anteriores. El proceso se basa en el concepto de boosting, donde los árboles se construyen uno tras otro, y cada nuevo árbol se enfoca en los errores residuales del conjunto de árboles anteriores. XGBoost utiliza un enfoque de gradient boosting, que optimiza la función de pérdida mediante el ajuste de los árboles en función del gradiente de esa función. Según (Chuanbin Li, 2018), sugiere realizar una serie de pruebas que incluyen el coeficiente de correlación de Pearson y el análisis de la importancia de las características antes de decidir usar finalmente el modelo XGBoost.

Por su parte, los autores (Haoran Zhang, 2024) utilizaron un modelo TCN (Temporal Convolutional Network) para la predicción del precio de la electricidad, comparando el rendimiento con otros modelos tradicionales como el ARIMA y otros modelos de redes neuronales como el BPNN, LSTM, EAL y EAT, siendo el TCN el de mejor rendimiento tanto en métricas de error como en tiempo de procesamiento. Los parámetros finales fueron 5 capas de 128 neuronas, un kernel de tamaño 5, learning rate de 0.001 y un optimizador tipo Adam, dando como resultado un MAPE de 0.13, lo cual refleja que es eficiente y preciso para la predicción de series temporales.

# Determinar los Objetivos del Negocio

## Objetivo General

Desarrollar un modelo de machine learning que, mediante el análisis de datos históricos, identifique patrones y tendencias para estimar el costo unitario de energía eléctrica en el segmento residencial de Colombia a seis meses, con el propósito de mejorar la precisión del modelo actualmente utilizado por el ministerio entregando una mejor proyección del costo unitario de energía en Colombia, para soportar la toma de decisiones de negocio que realiza el ministerio de Energía.

## Objetivos específicos

• Comprender el contexto del negocio y los datos históricos del costo unitario de energía eléctrica residencial en Colombia, mediante la exploración y análisis del modelo de predicción actual del ministerio y los datos disponibles.

• Identificar modelos de Machine Learning que permitan predecir patrones y tendencias en los datos históricos del costo unitario de energía eléctrica residencial en Colombia y evalúen el impacto de las variables externas en el pronóstico del costo.

• Diseñar, entrenar y validar modelos de machine learning de series temporales utilizando datos históricos del costo unitario de energía eléctrica para el segmento residencial en Colombia, y evaluar su precisión y rendimiento para seleccionar el modelo más adecuado en comparación con el modelo actual del ministerio.

# Contexto

El Ministerio de Minas y Energía ha desarrollado un proyecto de integración y análisis de datos enfocado en la creación de la plataforma Intégrame, una herramienta que consolida, procesa y analiza datos provenientes de 17 fuentes de información relacionadas con los sectores de minería, hidrocarburos y energía (Minenergía, 2023). Este proyecto inició en 2018 con una hoja de ruta que incluyó las siguientes fases: definición de la arquitectura empresarial, identificación de necesidades de información, priorización de los datos requeridos, desarrollo de un modelo de Gobierno de Datos, puesta en marcha, fortalecimiento y apropiación.

Esta plataforma centraliza y gestiona la información sectorial, reduciendo el tiempo necesario para la generación de informes, todo basado en tecnologías de la Cuarta Revolución Industrial (4RI), como soluciones de analítica avanzada, Machine Learning e inteligencia de negocios aplicada en el sector público. El Ministerio busca ofrecer acceso a más de 70 tableros, permitir la descarga de datos depurados a través de un catálogo de datos y proporcionar una conexión directa a la información disponible.

Figura 2 Arquitectura de datos MME

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

*Fuente: Ministerio de Minas y Energía, Intégrame*

La arquitectura implementada para este proyecto se basa en un ecosistema de Microsoft, utilizando Azure como gestor de datos en la nube, con herramientas integradas en cada etapa de almacenamiento, procesamiento, análisis y visualización de datos. Esta arquitectura es crucial para el desarrollo del reto analítico, ya que la implementación del modelo final, junto con su despliegue, debe estar respaldada por las herramientas de la institución, aprovechando al máximo los recursos disponibles.

## Modelo analítico

En anteriores proyectos, se adelantó la implementación de un modelo analítico enfocado a predecir el costo de la tarifa unitaria de energía en el nivel de tensión 1, haciendo una aproximación con series de tiempo multivariadas. El modelo resultante de este desarrollo fue un Vector de Correlación del Error (VEC), esta herramienta se caracteriza por contener variables cointegradas, es decir, variables que guardan una relación de equilibrio de largo plazo entre ellas (Sánchez, 2014). En esta aproximación se realizó la predicción de cada uno de los componentes de la tarifa total incluyendo variables exógenas como el precio de bolsa, aportes hídricos, capacidad de plantas, TRM, IPC e IPP.

Tabla 1 Métricas de error del modelo final VEC



*Fuente: Ministerio de Minas y Energía, Intégrame*

El modelo genera las predicciones finales de cada componente, analizando las correlaciones de las variables exógenas frente a cada componente, para saber que el indicador tiene mayor peso dentro de las fluctuaciones del precio final. Por último, y aunque se contrastan varias métricas de error usadas en un modelo de regresión, el MAPE es la principal para elegir el modelo final.

## Tablero de visualización

Como componente final, el proyecto anteriormente desarrollado generó un tablero de seguimiento a las proyecciones resultantes del modelo explicado en el punto anterior, junto con la visualización del comportamiento histórico de cada uno de los componentes del costo unitario de energía. Además, enfocaron la construcción del tablero al seguimiento eficiente y entendible de las correlaciones de las variables exógenas frente a la variable principal, mostrando gráficamente el impacto de cada uno en los diferentes escenarios propuestos.

Figura 3 Tablero con proyecciones finales

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

*Fuente: Ministerio de Minas y Energía, Intégrame*

Los elementos desarrollados en tiempos pasados son útiles para el avance y construcción del proyecto actual, ya que dan unas bases no solo del tratamiento y modelamiento de los datos, sino un punto de partida y comparación frente al modelo que se quiere construir enfocado en mejorar la predicción en el Costo Unitario de Energía. Además, muestra el foco de interés propio del negocio y cuáles son las problemáticas a resolver, que deben generar valor en las actividades diarias del Ministerio.

1. Problema de Negocio

El Ministerio de Energía requiere conocer la proyección del costo unitario de energía en Colombia para poder proporcionar recomendaciones basadas en datos y así optimizar la estructura tarifaria, balanceando eficiencia, sostenibilidad, transparencia y equidad, así como, proporcionar alertas tempranas para identificar posibles anomalías o cambios en los costos futuros. También permite disponibilizar la información para los usuarios y stakeholders del sector, lo cual es fundamental para la planificación presupuestaria de entidades públicas y privadas, anticipando y gestionando los gastos energéticos futuros de manera informada.

El Ministerio realizó un primer acercamiento en el ámbito analítico para la construcción de un modelo de predicción de los componentes del costo unitario de energía eléctrica en Colombia; sin embargo, este no se encuentra actualizado ni ofrece la precisión o la granularidad necesarias para evaluar adecuadamente el impacto de variables exógenas. Por lo tanto, es esencial desarrollar y explorar modelos analíticos más avanzados que superen estas limitaciones y proporcionen predicciones más detalladas y fiables.

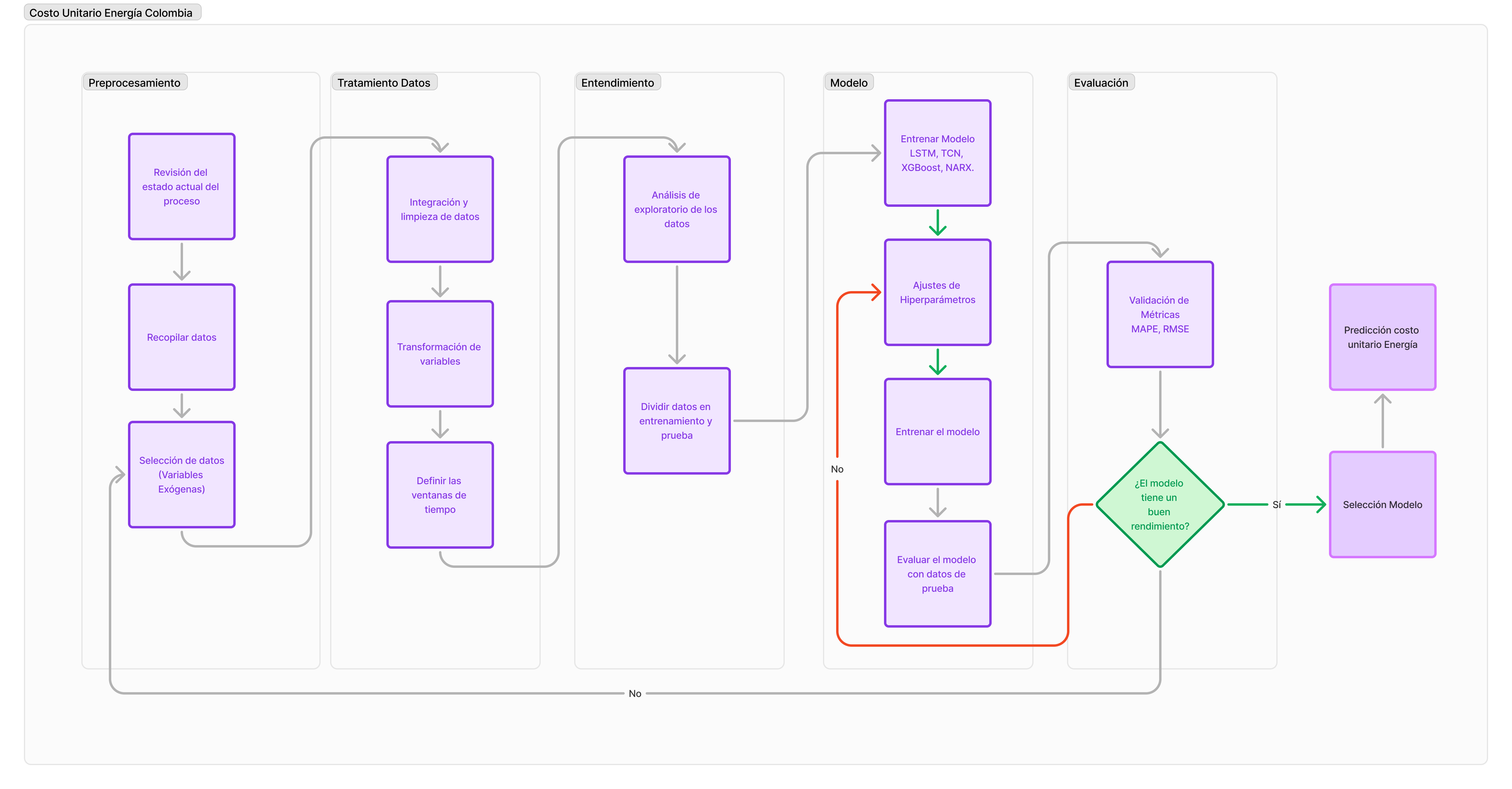
# Tarea analítica

Desarrollar y validar un modelo de regresión para series de tiempo que, utilizando técnicas de machine learning, analice datos históricos y patrones de variables externas para predecir con precisión el costo unitario de la energía eléctrica en el segmento residencial de Colombia a seis meses, mejorando así las proyecciones actuales del Ministerio de Minas y Energía.

# Flujo de datos

A continuación, se presenta un flujo de datos propuesto para desarrollar el proyecto de grado que consiste en pronosticar el costo unitario del costo de energía en Colombia.

Figura 4. Flujo de Datos Costo Unitario de Energía en Colombia



*Fuente: Elaboración Propia*

## Consideraciones para la construcción del modelo

El ministerio actualmente cuenta con un modelo de predicción de costo unitario (VEC), el cual será considerado como el modelo de referencia para ser comparado con el modelo propuesto en el presente trabajo.

Las siguientes son las consideraciones que fueron determinadas por parte del ministerio de energía para la construcción del modelo (VEC), las cuales serán consideradas en nuestro equipo de trabajo para que una comparación entre el modelo actual y el modelo propuesto sea posible:

* Se determina que los cargos horarios a tener en cuenta de las tarifas de energía son los cargo horario monomio.
* Se garantiza que el nivel de tensión corresponde a la línea residencial, es decir, Nivel de tensión 1-100 NT1 Propiedad OR.
* Se valida que todos los mercados contengan un prestador que sea incumbente.
* Métricas horarias para Precio de Bolsa.
* Aporte hídrico métrica periodicidad diaria.
* Capacidad efectiva neta periodicidad diaria.

## Determinar y obtener los datos necesarios

Obtener datos históricos de los componentes que determinan el costo unitario de energía. Así como las variables exógenas que impactan en indirectamente el precio de este costo. Para posteriormente proceder a cargar las bases en el entorno de análisis y el repositorio Github. Los siguientes son los componentes que determinan el costo unitario de energía en Colombia:

### Componentes

Se recogen datos de los costos componentes (comercialización, distribución, generación, pérdidas, restricción y transmisión) desde las empresas del sector.

* **Costo de Comercialización**: Son los costos asociados a la compra y venta de energía por parte de las empresas comercializadoras. Incluye gastos administrativos, de marketing, y gestión de clientes.
* **Costo de Distribución**: Costos asociados a la entrega de energía desde las subestaciones hasta los usuarios finales. Incluye el mantenimiento de las redes de distribución.
* **Costo de Generación**: Costos relacionados con la producción de energía eléctrica. Varía según el tipo de generación (hidráulica, térmica, eólica, etc.).
* **Costo de Pérdidas**: Pérdidas técnicas y no técnicas durante el transporte y distribución de la energía. Las pérdidas técnicas son inherentes al proceso físico de transmisión y distribución, mientras que las no técnicas suelen estar relacionadas con fraudes y errores de medición.
* **Costo de Restricción**: Costos incurridos cuando hay limitaciones en la red de transmisión que impiden el transporte eficiente de energía desde los generadores hasta los consumidores.
* **Costo de Transmisión**: Costos relacionados con el transporte de energía desde las plantas generadoras hasta las subestaciones de distribución. Incluye el mantenimiento de las líneas de alta tensión.

Se obtienen datos de las variables exógenas (aportes hídricos, precio de bolsa, capacidad efectiva neta, IPP, IPC y TRM) de fuentes como el IDEAM, XM, DANE y el Banco de la República.

### Variables Exógenas

* **Aportes Hídricos**: Cantidad de agua disponible en los embalses y ríos, que afecta la capacidad de generación hidroeléctrica. **Importancia:** Los niveles de agua influyen directamente en la capacidad de generación hidroeléctrica, que es una de las fuentes principales de energía en Colombia. (IDEAM, 2006)
* **Precio de Bolsa**: Precio al cual se transa la energía en el mercado mayorista. Se determina por la oferta y demanda de energía en un período específico. **Importancia**: Refleja las condiciones del mercado y puede ser un indicador de los costos futuros.
* **Capacidad Efectiva Neta**: Capacidad máxima de generación de energía que puede ser sostenida durante un periodo prolongado bajo condiciones normales de operación. **Importancia**: Indica la disponibilidad real de la infraestructura de generación de energía.
* **IPP (Índice de Precios al Productor)**: Indicador económico que mide la variación de los precios de venta de los productos a nivel mayorista. **Importancia**: Afecta los costos de los insumos y servicios necesarios para la generación y distribución de energía.
* **IPC (Índice de Precios al Consumidor)**: Indicador económico que mide la variación de los precios de bienes y servicios que consume la población. **Importancia**: Influye en los costos operativos y administrativos de las empresas del sector energético.
* **TRM (Tasa Representativa del Mercado)**: Tasa de cambio entre el peso colombiano y el dólar estadounidense, fijada diariamente por el Banco de la República. **Importancia**: Impacta los costos de importación de insumos y tecnologías necesarias para la generación y distribución de energía.

## Preprocesamiento de datos

Se realiza la carga de los datos en el repositorio de Github.

### Selección de datos

En primer lugar, se procede a cargar las diferentes bases de datos que contienen las variables exógenas obtenidas de cada una de sus fuentes. Este proceso incluye la validación de que todas las bases de datos tengan el formato de fecha correcto, lo cual es crucial para los análisis de series temporales. Además, se verifica que los montos estén en el formato adecuado para garantizar la coherencia y comparabilidad de los datos.

### Integración y limpieza de datos

Posteriormente, se realiza la integración de las diferentes bases de datos utilizando la variable BK\_FECHA como llave primaria, la cual corresponde al primer día de cada mes. Este paso es fundamental para combinar los datos de manera precisa. Durante esta etapa, también se eliminan las columnas duplicadas relacionadas con las fechas para evitar redundancias y posibles errores en los análisis posteriores. Se lleva a cabo una validación de los valores, asegurando que no existan valores nulos o NA que puedan afectar la calidad de los datos.

### Validación de datos

Una vez integrados y limpiados los datos, se valida la cantidad de información disponible por año y por mes. Este paso es importante para garantizar la integridad y completitud de los datos a lo largo del tiempo. Luego, se agregan los datos relevantes para los parámetros de interés, facilitando así el análisis específico que se desea realizar.

### Formato de datos

Se crea un dataframe denominado df\_graph, que será utilizado para graficar las series de tiempo y llevar a cabo procesos de imputación de datos cuando sea necesario. Este dataframe es una herramienta clave para visualizar tendencias y patrones en los datos a lo largo del tiempo. Sobre este dataframe, se realiza un agrupamiento por la variable BK\_FECHA (desde 1-Ene-2015 hasta 01-Ene-2023), y se resume la información utilizando el promedio de los valores, lo que permite obtener una visión más clara y concisa de las tendencias subyacentes.

## Entendimiento de los datos

### Análisis Exploratorio de los datos

Ya con la base consolidada, en esta etapa se pretende realizar una revisión general de la estructura (filas, columnas, entre otros), para posteriormente realizar el análisis estadístico descriptivo como la media, mediana, moda, desviación estándar, valores mínimos y máximos. Y como último paso poder visualizar la distribución de los datos utilizando histogramas, diagramas de Boxplot, diagrama de líneas, entre otros.

### Análisis de Correlación

El análisis de correlación ayuda a identificar la relación entre variables numéricas. Esto es clave para entender cómo las variables están relacionadas entre sí y cuál podría ser su impacto en la variable de interés en este caso del costo unitario de energía. Para este caso se explorarán algunos coeficientes de correlación como lo es el caso del coeficiente de correlación de Pearson, el cual mide la fuerza y a dirección de la relación lineal entre dos variables numéricas. Los valores oscilan entre -1 (correlación negativa perfecta) y 1 (correlación positiva perfecta), o también la correlación de spearman, el cual Mide la fuerza y la dirección de la relación monotónica entre dos variables ordinales o numéricas.

### Pruebas estadísticas en modelos de series temporales

Las pruebas estadísticas para series de tiempo son herramientas útiles para analizar y modelar datos temporales. Estas pruebas ayudan a identificar características importantes de las series, como la estacionalidad y la tendencia, y son esenciales para seleccionar y validar modelos de pronóstico adecuados. Tales como la prueba de estacionariedad, la cual es clave en series de tiempo porque modelos como el ARIMA asumen estacionariedad y para esto se aplicará el test Dickey-Fuller que evalúa la hipótesis nula de una unidad de raíz está presente en una serie temporal y eso nos ayuda a determinar si una serie es estacionaria.

## Selección del modelo

Utilizar varios modelos enfocados en la utilización de series de tiempo, los cuales ayuden a determinar el objetivo propuesto en este proyecto. Dentro de los modelos de series de tiempo se encuentran algunas de las siguientes opciones:

* + 1. ARIMA: Es un modelo estadístico que combina autoregresión, integración (diferenciación) y media móvil. **SARIMA**: (Seasonal ARIMA) extiende ARIMA para manejar la estacionalidad. Adecuado para series de tiempo con tendencia y estacionalidad. Ampliamente utilizado en análisis económicos y financieros.

* + 1. Long Short-Term Memory (LTSM): Es un tipo de red neuronal recurrente diseñada para manejar problemas de dependencia a largo plazo, lo que la hace adecuada para series de tiempo. Puede recordar valores durante intervalos de tiempo largos y cortos. Ideal para series de tiempo con patrones complejos y no lineales, como datos financieros, meteorológicos o de ventas.

* + 1. XGBoost: Es una implementación optimizada de los árboles de decisión en forma de boosting. Para series de tiempo, se puede usar para crear modelos basados en características creadas a partir de los datos temporales. Utilizado en competiciones de machine learning y en aplicaciones donde la precisión es crucial, como predicciones de demanda y precios.

* + 1. Temporal Convolutional Networks (TCN): Es una arquitectura de red neuronal convolucional diseñada específicamente para series de tiempo. Las TCN utilizan convoluciones dilatadas para capturar dependencias a largo plazo y son capaces de aprender representaciones jerárquicas de los datos.
    2. Nonlinear Autoregressive with Exogenous Inputs (NARX): tipo de modelo de redes neuronales recurrentes (RNN) utilizado para predecir series de tiempo. Es especialmente útil cuando la serie temporal que se quiere predecir depende no solo de sus valores pasados (autoregresión) sino también de otras variables externas (exógenas).

Estos modelos llegan a ser útiles para predecir precios en una serie de tiempo, por lo que resultan efectivos para intentar resolver nuestro problema de predicción del precio unitario del costo de la energía en Colombia.

## División de datos en entrenamiento y prueba.

La división de los datos para el modelo de series de tiempo tiene una alta relevancia la naturaleza secuencial de los datos. Las series de tiempo requieren mantener el orden temporal para no introducir sesgos y asegurar una evaluación correcta del modelo. Es necesario tener en cuenta los siguientes pasos:

* Mantener el orden temporal de los datos es fundamental para poder realizar la división de los datos de manera que no afecte el modelo.
* Se dividen los datos en Train y test, aunque la recomendación es 90% - 10% se parametriza con 87% - 13% la partición con el fin de tener 12 meses para testear. Se utilizan los datos con este periodo, que es igual a los datos del modelo ya desarrollado por el ministerio, para poder realizar la comparación y evaluación de la métrica de desempeño bajo los mismos criterios.
* Realizar validación cruzada para la selección de los hiperparámetros, lo que implica entrenar el modelo en un subconjunto de datos y luego probarlo en el siguiente periodo. Este proceso se repite deslizando la ventana de entrenamiento hacia adelante. Por ejemplo, se puede entrenar el modelo con los datos de enero de 2015 a diciembre de 2016 y probarlo con datos de enero de 2017 a diciembre de 2017, luego entrenar con datos de enero de 2015 a diciembre de 2017 y probar con datos de enero de 2018 a diciembre de 2018. Repitiendo este proceso hasta cubrir todo el periodo de prueba.

## Evaluación del modelo

### Métrica de evaluación

La utilización de modelos de series de tiempo en el pronóstico de costos unitarios es crucial debido a la capacidad de estos modelos para capturar patrones temporales como tendencias, estacionalidades y ciclos, lo que permite una proyección más precisa de valores futuros. Estudios como "Evaluating time series forecasting models: an empirical study on performance estimation methods" de (Cerqueira, Torgo, & Mozetič, 2020) enfatizan que el uso de series de tiempo puede mejorar la precisión de los pronósticos al considerar la dependencia temporal de los datos, lo cual es esencial en contextos donde los costos unitarios fluctúan a lo largo del tiempo​.

Para evaluar el desempeño de los modelos de series de tiempo, la literatura especializada sugiere utilizar diversas métricas de error que facilitan la valoración de la precisión en las predicciones. Estas métricas permiten determinar la eficacia del modelo en la captura de tendencias y patrones de los datos, comparar diferentes modelos para identificar el que mejor se ajusta a los datos, evaluar el desempeño bajo diversas combinaciones de hiperparámetros y proporcionar una medida de confianza que respalde las decisiones basadas en los resultados obtenidos.

Entre las métricas sugeridas de acuerdo con (Andrés, 2023), se encuentran el Error Absoluto Medio (MAE) y el Error Cuadrático Medio (MSE) son ampliamente utilizados debido a su simplicidad y facilidad de interpretación. Estos errores cuantifican las desviaciones promedio entre los valores predichos y los observados, proporcionando una medida clara de la precisión del modelo. Además, métricas como el Error Porcentual Absoluto Medio (MAPE) y el Error Absoluto Escalado (MASE) ofrecen perspectivas adicionales, siendo útiles para comparar el rendimiento de modelos en diferentes contextos y escalas.

Por otro lado, (Hyndman & Athanasopoulos, 2021) destacan la importancia de elegir métricas de error adecuadas para evaluar la precisión de los modelos de series de tiempo. El libro enfatiza que no existe una única métrica que sea ideal en todas las situaciones, y la selección debe depender del contexto y de los objetivos específicos del análisis. Las métricas más comunes, como el Error Absoluto Medio (MAE), el Error Cuadrático Medio (MSE) y la Raíz del Error Cuadrático Medio (RMSE), proporcionan diferentes perspectivas sobre el desempeño del modelo, desde una evaluación directa del error promedio hasta una penalización de errores grandes. Además, el libro aborda métricas porcentuales como el Error Absoluto Porcentual Medio (MAPE), que facilitan la interpretación en términos relativos que ofrece robustez frente a valores extremos y fluctuaciones en los datos. Los autores subrayan que la comprensión de las ventajas y limitaciones de cada métrica es crucial para una evaluación completa y precisa del rendimiento del modelo.

Tras evaluar diversas métricas para la comparación de modelos de series temporales, y considerando que uno de los objetivos de este trabajo es mejorar el desempeño del modelo existente para la predicción del costo unitario de energía, se adoptará el Mean Absolute Percentage Error (MAPE) como la medida principal de desempeño. Esta decisión se fundamenta en el uso previo del MAPE como referencia principal en el modelo actual y en la necesidad de seleccionar métricas adecuadas según el contexto específico del pronóstico. El MAPE se calcula como la media de los errores porcentuales absolutos, obtenidos dividiendo la diferencia absoluta entre los valores observados y los predichos por los valores observados, y luego multiplicando por 100. Esta métrica es altamente interpretable, ya que expresa el error en términos de proporciones relativas, lo cual es particularmente útil en el caso de estudio. Al predecir varios costos que componen el costo unitario total, los cuales pueden encontrarse en distintas magnitudes, el MAPE permite una evaluación clara y directa de la precisión del modelo.

El MAPE matemáticamente se expresa de la siguiente manera:

Ecuación 1. Ecuación MAPE

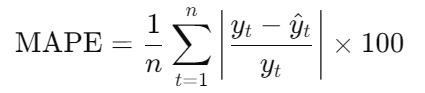


Tabla 2. Interpretación Métrica MAPE

|  |  |
| --- | --- |
| **Rango del resultado** | **Interpretación** |
| **0%** | Indica una predicción perfecta, donde no hay errores entre los valores observados y los valores predichos. |
| **< 10%** | Se considera generalmente una predicción excelente. |
| **>=10% < 20%** | Se considera una predicción buena. |
| **>=20% < 50%** | Se considera una predicción razonable. |
| **> 50%** | Indica una predicción deficiente. |

Según las referencias encontradas, evaluaremos la raíz del Error Cuadrático Medio (RMSE) El RMSE será comparado junto con el MAPE para ofrecer una visión completa del desempeño del modelo. El RMSE se calcula tomando la raíz cuadrada de la media de los cuadrados de las diferencias entre los valores predichos y los valores reales. Esta métrica proporciona una medida de la magnitud del error en las mismas unidades que la variable de salida, permitiendo una interpretación directa de la precisión del modelo.

En la siguiente tabla, se proporciona una visión general de los hiperparámetros clave que se pueden ajustar para optimizar los modelos de series de tiempo seleccionados en el presente trabajo:

Tabla 3. Descripción Hiperparámetros Modelos

|  |  |  |
| --- | --- | --- |
| **Modelo** | **Hiperparámetro** | **Descripción** |
| **ARIMA** | p (AR order) | Número de rezagos en la parte autorregresiva del modelo. |
| d (Differencing order) | Número de veces que se necesita diferenciar la serie para hacerla estacionaria. |
| q (MA order) | Número de rezagos en la parte de media móvil del modelo. |
| **RNN / LSTM** | Number of units | Cantidad de neuronas en cada capa LSTM/RNN. |
| Number of hidden layers | Profundidad de la red. |
| Learning rate | Tasa de aprendizaje para la actualización de los pesos del modelo. |
| Dropout rate | Porcentaje de neuronas descartadas durante el entrenamiento para prevenir sobreajuste. |
| Batch size | Número de muestras procesadas antes de actualizar el modelo. |
| **XGBoost** | n\_estimators | Número de árboles en el modelo. |
| max\_depth | Profundidad máxima de los árboles. |
| learning\_rate | Tasa de aprendizaje. |
| subsample | Fracción de muestras utilizadas para entrenar cada árbol. |
| colsample\_bytree | Fracción de características utilizadas por cada árbol. |
| **Temporal Convolutional Networks (TCN)** | Number of convolutional layers | Profundidad de la red. |
| Kernel size | Determina el rango temporal que cada filtro puede capturar |
| Number of filters per layer | Controla la cantidad de características extraídas en cada capa convolucional. |
| Dilation strategy | Establece el factor de dilatación en las convoluciones, lo que afecta la resolución temporal y la capacidad para captar dependencias a largo plazo. |
| Activation function | La función utilizada para introducir no linealidades en el modelo |
| **Nonlinear Autoregressive Exogenous (NARX)** | Number of input lags | Define cuántos valores anteriores del input se utilizan para la predicción futura. |
| Number of output lags | Establece cuántos valores anteriores de la salida se consideran en la predicción. |
| Type of activation function | La función utilizada para introducir no linealidades en el modelo |
| Number of neurons in the hidden layer | Controla la capacidad de la red para aprender representaciones complejas de los datos. |

*Fuente: Elaboración Propia*

Estos hiperparámetros pueden ser sintonizados por medio de técnicas utilizadas para encontrar la combinación óptima que maximice el rendimiento de un modelo predictivo. Estos métodos exploran el espacio de posibles valores de hiperparámetros de manera sistemática o aleatoria. Entre los más comunes se encuentran Grid Search, Random Search y Bayesian Optimization. Grid Search evalúa todas las combinaciones posibles dentro de un rango predefinido, garantizando una cobertura exhaustiva del espacio de búsqueda, aunque es computacionalmente intensivo. Random Search, por otro lado, selecciona combinaciones aleatorias de hiperparámetros, ofreciendo una alternativa menos costosa en términos de recursos y a menudo más efectiva en espacios grandes. Bayesian Optimization utiliza modelos probabilísticos para dirigir la búsqueda hacia áreas prometedoras del espacio de hiperparámetros, mejorando la eficiencia y reduciendo el número de evaluaciones necesarias.

# Referencias

Chaparro, C. M., Martínez, L. F., & Trujillo, E. R. (2012). Predicción de la demanda de energía eléctrica basado en análisis Wavelet y un modelo neuronal auto-regresivo no lineal NAR. *Tecnura, Vol. 16*, pp 86 - 99. Obtenido de https://livejaverianaedu-my.sharepoint.com/personal/pa\_castro\_javeriana\_edu\_co/\_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fpa%5Fcastro%5Fjaveriana%5Fedu%5Fco%2FDocuments%2FTrabajo%20de%20grado%2FArticulo%20Prediccion%20energia%2Epdf&parent=%2Fpersonal%2Fpa

Cerqueira, V., Torgo, L., & Mozetič, I. (2020). *Evaluating time series forecasting models: an empirical study on performance estimation methods.* (S. Nature, Ed.) Obtenido de https://doi.org/10.1007/s10994-020-05910-7

Andrés, D. (22 de junio de 2023). *Error Metrics for Time Series Forecasting*. Obtenido de https://mlpills.dev/time-series/error-metrics-for-time-series-forecasting/

Hyndman, R., & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice.* Melbourne, Australia. Obtenido de OTexts.com/fpp3

Cárdenas, J. E. (2021). Empleo de modelos predictivos en el precio de bolsa de la energía en Colombia. Bogotá D.C., Colombia. Obtenido de chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://repositorio.uniandes.edu.co/server/api/core/bitstreams/cba24c35-9e02-4c2f-af11-662b9cd44ca5/content

Hernández N., O. V. (1 de Diciembre de 2005). *Modelos ARIMA y estructural de la serie de precios promedio de los contratos en el Mercado Mayorista de Energía Eléctrica en Colombia.* Obtenido de Redalyc: https://www.redalyc.org/comocitar.oa?id=147019373001

Lombana, A. M. (31 de octubre de 2017). *Pronóstico del precio de la energía en Colombia utilizando modelos ARIMA con IGARCH*. Obtenido de Revistas Urosario: https://revistas.urosario.edu.co/xml/5095/509554391005/html/index.html

Alberto Muñoz, J. U. (31 de octubre de 2017). *Pronóstico del precio de la energía en Colombia utilizando modelos ARIMA con IGARCH*. Obtenido de Revistas Urosario: https://revistas.urosario.edu.co/xml/5095/509554391005/html/index.html

Chuanbin Li, X. Z. (1 de May de 2018). *Wiley Oline Library*. Obtenido de Predicting Short-Term Electricity Demand by Combining the Advantages of ARMA and XGBoost in Fog Computing Environment: https://onlinelibrary.wiley.com/doi/full/10.1155/2018/5018053

Minenergía. (2 de Agosto de 2023). *minenergia.gov.* Obtenido de https://www.minenergia.gov.co/es/sala-de-prensa/noticias-index/minenerg%C3%ADa-presenta-int%C3%A9grame-la-herramienta-que-facilita-el-acceso-a-datos-del-sector-minero-energ%C3%A9tico-en-colombia/

Adriana P. Agudelo, J. M.-L. (2015). *Predicción del Precio de la Electricidad en la Bolsa mediante un Modelo Neuronal No-Lineal Autorregresivo con Entradas Exógenas.* Medellín: Grupo de Investigación GIMEL, Facultad de Ingeniería, Universidad de Antioquia.

Sánchez, J. J. (2014). *MODELO VEC PARA LA ESTIMACIÓN DE INFLACIÓN BURSÁTIL: EVIDENCIA EMPIRICA EN MERCADOS NORTEAMERICANOS.* Medellín: Doctorado en Economía y Administración de Empresas, Universidad Privada Boliviana.

Haoran Zhang, W. H. (2024). A Temporal Convolutional Network Based Hybrid Model for Short-term Electricity Price Forecasting. *CSEE JOURNAL OF POWER AND ENERGY SYSTEMS*, VOL. 10.

(24 de Agosto de 2023). (Xm, Productor) Recuperado el Julio de 2024, de Variables del Mercado de Energía En Julio de 2023

Corficolombiana. (24 de Mayo de 2021). Obtenido de Investigaciones Corficolombiana: https://investigaciones.corficolombiana.com/documents/38211/0/Generaci%C3%B3n%20el%C3%A9ctrica%20en%20Colombia%20y%20su%20transici%C3%B3n%20hacia%20Fuentes%20Renovables%20No%20Convencionales.pdf/5ffcba57-f7b8-f4b6-35c0-ae9302bd1a0a

Luis Fernando González Pérez, S. N. (2022). *Universidad de los Andes*. Obtenido de Repositorio: https://repositorio.uniandes.edu.co/server/api/core/bitstreams/ec097e02-76ff-41ca-b603-ce82eb8a8636/content

Laura Melgar García, J. F. (s.f.). Obtenido de https://www.mintur.gob.es/Publicaciones/Publicacionesperiodicas/EconomiaIndustrial/RevistaEconomiaIndustrial/431/12\_MELGAR.pdf

Daza, J. M. (27 de Septiembre de 2023). Obtenido de Diario la República: https://www.larepublica.co/globoeconomia/hungria-y-colombia-son-los-paises-de-la-ocde-con-los-precios-mas-elevados-en-energia-3714233