Predicción y análisis de patrones de generación de energía reactiva en una planta de productos lácteos

Estudiantes:

Javier Andrés Causil Martínez

Yohiner Andrés Borja Góez

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

Asesor

Dr. Gabriel Dario Uribe Guerra

Universidad de Antioquia

Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2024

Cita	*(Causil Martínez & Borja Goez, 2024)*
Referencia	Causil Martínez, J. A., & Borja Goez, Y. A. (2024). *Predicción y análisis de patrones de generación de energía reactiva en una planta de productos lácteos*, trabajo de grado especialización. Universidad de Antioquia, Medellín, Colombia
Estilo IEEE (2024)	

Especialización en Analítica y Ciencia de Datos, Cohorte VII.
Centro de Investigación Ambientales y de Ingeniería (CIA).

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

Agradecimientos

Expresamos nuestro más sincero agradecimiento a Santiago Gómez Vélez, CEO de Leaf IoT, por su invaluable apoyo y confianza durante este proceso. Asimismo, extendemos nuestra gratitud a nuestras familias, al Dr. Gabriel Dario Uribe Guerra asesor del proyecto y a nuestras parejas, quienes con su respaldo y aliento constante hicieron posible la culminación de este trabajo.

Índice general

Índice de figuras

Índice de cuadros

Resumen

Este trabajo aborda la problemática de una empresa de productos lácteos relacionada con penalizaciones por la generación de energía reactiva. Para contribuir a su solución, se implementaron modelos de Machine Learning, incluyendo SARIMAX, LSTM, y Redes Neuronales Recurrentes (RNNs). Además, se utilizó un modelo no supervisado K-Means para identificar patrones en los días y horas según la generación de energía reactiva inductiva. Los resultados de las métricas de desempeño destacaron a la LSTM como el modelo más preciso para la predicción, mientras que el análisis con K-Means permitió identificar periodos específicos de alta generación de energía reactiva, proporcionando información valiosa para la optimización del consumo energético.

Abstract

This work addresses the issue faced by a dairy company regarding penalties for the generation of reactive energy. To contribute to its solution, Machine Learning models were implemented, including SARIMAX, LSTM, and Recurrent Neural Networks (RNNs). Additionally, an unsupervised K-Means model was employed to identify patterns in days and hours based on inductive reactive energy generation. Performance metrics highlighted LSTM as the most accurate model for prediction, while the K-Means analysis identified specific periods of high reactive energy generation, providing valuable insights for optimizing energy consumption.

1 Descripción del problema

La empresa **Leaf Iot**, proveedor de soluciones tecnológicas, colabora con la compañía de productos lácteos Auralac S.A. para abordar un problema relacionado con el alto costo de consumo energético en su planta de producción. En particular, Auralac busca reducir los costos asociados a la generación de **energía reactiva**, la cual incrementa el consumo eléctrico sin aportar valor directo al proceso productivo.

Para abordar el problema, se propone el desarrollo de **modelos de pronóstico** que permitan estimar la generación de energía reactiva, con el objetivo de identificar los días de mayor consumo y establecer estrategias para reducir la generación de energía reactiva. Para ello, se cuenta con datos históricos de consumo de energía eléctrica desde febrero de 2020 hasta septiembre de 2024.

1.1. Problema de negocio

Según **CELSIA** [5], la energía reactiva es un tipo de energía eléctrica que es absorbida o que se inyecta a la red por algunos equipos. Para su funcionamiento necesitan un campo magnético, tales como motores, transformadores, ascensores, sistemas de bombeo de agua, motores de aireación de piscinas, iluminación eficiente, y otros. La unidad de medida de este tipo de energía es $kVarh$ que se interpreta como kilo voltio amperios reactivos por hora. Por otro lado, de acuerdo con EPM [7], la energía reactiva se puede entender como una energía que ocupa espacio de las redes eléctricas, pero no es útil a la hora de hacer un trabajo. Como la energía reactiva satura las redes, es necesario que las empresas tomen medidas para reducir su uso para evitar problemas de sobrecargas, ineficiencias y de calidad de la energía, lo cual repercute en sobre costos en la prestación del servicio.

De acuerdo con lo indicado por el CEO de Leaf Iot y el personal técnico de la Planta, en los últimos años, el nivel de generación de energía reactiva en la planta ha sufrido un aumento, y aunque la información de el gasto de energía global se conoce, no es posible identificar la totalidad de las fuentes que la están generando, lo cual tiene un impacto significativo sobre los costos de producción. Por esta razón, se esta adelantando una estrategía de control por medio del uso de sensores en las maquinas y motores que intervienen en el proceso de producción, con la finalidad de poder discriminar por zonas el impacto que estas herramientas le aportan al incremento de la energía reactiva. Este proceso se encuentra en una etapa de calibración de los sensores, por lo cual aún no es posible discriminar el aporte de cada fuente a la generación de energía reactiva de una forma efectiva.

La Comisión Reguladora de Energía y Gas CREG en [5] define la variable M como lo indica el siguiente artículo

"ARTÍCULO 70. La definición de la variable M incluida en el Capítulo 12 del anexo general de la Resolución CREG 015 de 2018 quedará así:

M: Variable asociada con el periodo mensual en el que se presenta el transporte de energía reactiva sobre el límite establecido, variando entre 1 y 12.

Cuando el transporte de energía reactiva en exceso sobre el límite se presente durante cualquier período horario en diez (10) días o menos en un mismo mes calendario, la variable M será igual a 1.

Cuando el transporte de energía reactiva en exceso sobre el límite se presente durante cualquier período horario en más de diez (10) días en un mismo mes calendario, la variable M será igual a 1 durante los primeros 12 meses en los que se presente esta condición. A partir del décimo tercer mes de transporte de energía reactiva con la misma condición, esta variable se incrementará mensualmente en una unidad hasta alcanzar el valor de 6.

Si el transporte de energía reactiva en exceso sobre el límite desaparece durante más de tres meses consecutivos, la variable reiniciará a partir de 1.

Cuando el valor de $M = 6$ se haya mantenido durante 12 meses, en caso de persistir el consumo de energía reactiva en exceso sobre el límite, a partir del mes siguiente la variable continuará incrementándose mensualmente en una unidad hasta alcanzar el valor de 12."

En la figura 1.1 se relaciona con el cálculo de la variable M

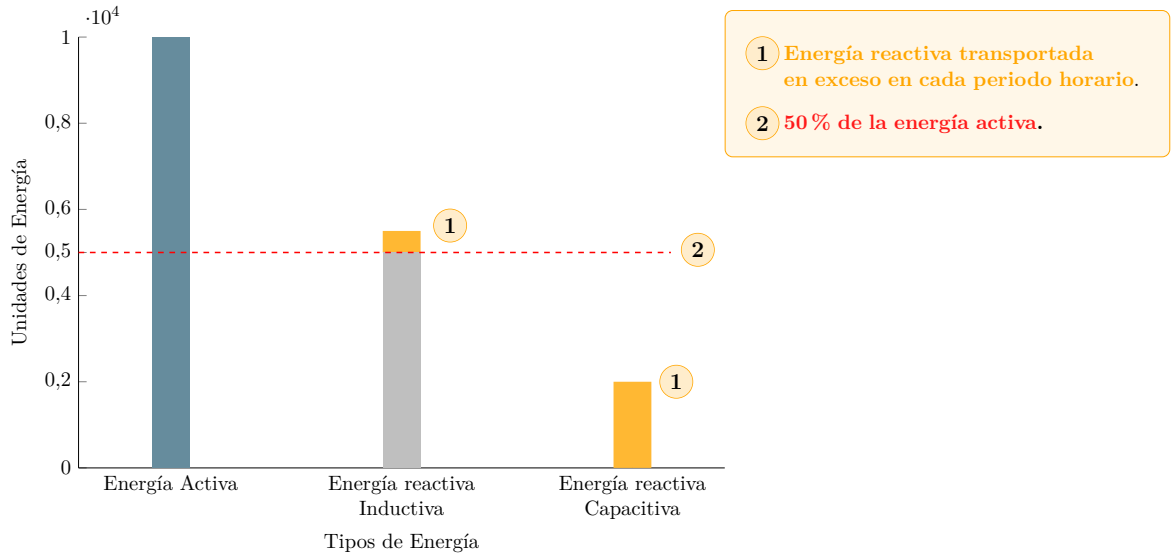

Figura 1.1: Exceso de la energía reactiva

En ella se observa que la cantidad de energía activa (barra azul), se produjó en más de 50 %, la cantidad de energía reactiva inductiva (Barra gris con naranja) y la cantidad de energía capacitiva (Barra naranja). Esto respresenta que la empresa prestadora de servicios de energía eléctrica, en este caso EPM, penalizará a la Planta Auralac S.A. por el exceso de energía reactiva generada, ya que hay un exceso (color naranja) que satura las redes eléctricas y no aporta valor al proceso productivo, afectando los transformadores y otros elementos. Por lo que un día sin exceso de energía reactiva se puede ver reflejado en la figura 1.1 cuando solo se pueden observar las barras en color azul y gris, este es ideal si solo esta la barra azul.

Entender el incremento de la variable M permite identificar que acciones se deben adoptar para reducir la generación de energía reactiva. Desde Leaf Iot se han planteado tres estrategias que son:

1. **Monitoreo y optimización del consumo eléctrico :** Identificar puntos de mejora en la planta para minimizar su impacto.
2. **Uso de equipos más eficientes:** Reducir el uso de equipos con altos consumos de energía reactiva.
3. **Instalación de condensadores o bancos capacitores:** Ayudan a compensar la energía reactiva y mejorar el factor potencia del sistema.

Se propone avanzar con la primera estrategía debido a que los numerales 2 y 3 son estrategias que se pueden implementar una vez se haya identificado la fuente de la generación de energía reactiva.

1.2. Aproximación desde la analítica de datos

Desde la analítica de datos se busca identificar los generadores de energía reactiva y activa por medio de la aplicación de modelos estadísticos y de Machine Learning (ML) por sus siglas en inglés. Y asi poder pronosticar los factores que influyen en los sobrescostos y penalizaciones de la empresa prestadores de servicio. Se espera que los resultados del modelo pueda dar a el equipo técnico de la Planta elementos para la toma decisiones que permitan controlar la energía reactiva generada y poder reducir del valor de la facturación mensual.

1.3. Origen de los datos

Los datos a utilizar en este proyecto fueron proporcionados por la empresa prestadora de servicios de energía eléctrica como soporte a los cobros realizados al consumo de energía eléctrica a la empresa Auralac.S.A.

1.4. Métricas de desempeño

Las métricas permiten observar y evaluar varios aspectos clave de los modelos, cómo la precisión, la robustez frente a errores, en está monografía se utilizan las métricas de la librería scikit-learn en [2] y un contexto de [8], ¿Cómo interpretamos las métricas MAE, RMSE y MAPE con energía reactiva y activa? Para esto los modelos que a implementar arrojan un valor de predicción, analicemos que nos dice cada métrica, primeramente con la reactiva inductiva y de manera análoga es con la energía activa:

1. **MAE:** Toma el valor absoluto de la diferencia entre cada registro o patrón de energía inductiva y la predicción del modelo, podemos apreciar en este punto, el MAE cálcula la distancia entre el registro real y la prediccón, luego promedia todas las distancias, lo

que representa una distancia promedio entre los valores de la predicción y los registros de la reactiva inductiva.

2. **RMSE:** En este caso se cálcula el cuadrado de la diferencia entre el registro de reactiva inductiva y la predicción, el RMSE al realizar el cuadrado le agrega un peso a la "distancia", debido al cuadrado, luego suma todos los valores, los promedia y saca la raíz. Cómo el cuadrado le agrega un peso a mi diferencia, cuando tengo registros que arrojan la diferencia muy grande estás se van a sumar y como la raíz no es distribuye con respecto a la suma, se tiene una contribución mayor por lo cual es sensible a valores atípicos. Esto quiere decir que si tenemos un RMSE muy grande, probablemente se tiene un dominio de está métrica por presencia de elementos atípicos.

3. **MAPE:** Es el valor absoluto de la diferencia entre el registro de reactiva inductiva y la preducción dividida entre el valor de la energía reactiva. Lo que seda a entender como una razón, esto indica que tanto es la diferencia con respecto al registro de energía reactiva, normalizando lo que le da independencia de la escala de los datos. Luego se promedian los resultados y se convierte a porcentaje. Si el MAPE es inferior al 10 %, indica un buen rendimiento del modelo. Cómo se promedian cocientes va hacer sensible si los registros de la reactiva inductiva son cercanos a cero, debido al que el cociente se vuelve grande y esto pude distorsionar el promedio.

2 Objetivos

2.1. Objetivo general

Desarrollar modelos utilizando técnicas estadísticas y de aprendizaje automático para pronosticar la generación de energía reactiva y activa, en una planta de productos lácteos, con datos del consumo energético en el periodo 2020 a 2024.

2.2. Objetivos específicos

Para desarrollar el objetivo general se plantean los siguientes objetivos específicos:

- Realizar una búsqueda de información relacionada al tema para tener un mayor contexto sobre el exceso de la energía reactiva.
- Relizar la extracción de los datos, transformación y carga de los datos que nos permita analizar y desarrollar los modelos.
- Realizar los pasos contemplados en la metodología CRISP-DM para la creación de los modelos de analítica.
- Evaluar la precisión de los modelos aplicados considerando errores MAE, MAPE y RMSE.
- Establecer, mediante el uso del modelo, los días del mes que hacen que la generación de energía reactiva sobrepase los 10 días permitidos.

3 Datos

3.1. Datos originales

El conjunto de datos original son 24 archivos Excel que describen la matriz de consumos de Energía Activa y Reactiva por hora de Auralac S.A., los archivos nos proporcionan información sobre la dirección, municipio, suscripción, servicio suscrito, nivel de tensión, periodo de consulta Día/Mes/Año, además de la matriz de los datos de la energía Activa Consumo (kWh), Energía Activa Generación (kWh), Energía Reactiva Capacitiva (kVarh) y Energía Reactiva Inductiva (kVarh) por hora. El periodo de tiempo va desde el 16/02/2020 hasta el 30/09/2024, que corresponden a un total de 1668 días de registro.

3.2. Datasets

Dado que los datos de cada año se encuentran dispersos en diferentes archivos de Excel, es necesario integrar esta información en una única tabla maestra. Para ello, se llevó a cabo el siguiente proceso:

1. Se extrajeron los datos relevantes de cada archivo, identificando las celdas que contienen la información de interés.
2. Se eliminaron las columnas que contienen únicamente valores nulos, asegurando la calidad de los datos a consolidar.
3. Se unieron los datos de todos los archivos en una sola tabla, creando una visión global de los datos a lo largo de los años.
4. Se creó una tabla que detalla la información para cada hora, manteniendo la granularidad de los datos.
5. Se creó una tabla que agrupa la información por día, sumando los valores correspondientes de cada hora para obtener los totales diarios de cada variable.

De esta manera, se obtuvieron dos conjuntos de datos, uno con un nivel de detalle por hora y otro con un resumen diario, con el fin de facilitar el análisis en dos esquemas de tiempo diferentes.

Cuadro 3.1: Muestra de datos de energía por día

date	activa_consumo	activa_generacion	reactiva_capacitiva	reactiva_inductiva
2020-02-16	4948	0	0	1382
2020-02-17	14400	0	0	7070
2020-02-18	14900	0	0	7640
2020-02-19	14980	0	0	7800
2020-02-20	16104	0	0	8376

Cuadro 3.2: Muestra de datos de energía por hora

d	1	2	3	4	5	6	7	8	9	10	11	12
25	1011	966	956	973	1018	1023	953	950	835	769	763	654
26	975	1034	1044	1024	908	870	772	733	763	597	697	790
27	1049	1060	1140	1000	918	963	1065	1060	931	721	704	627
28	973	988	949	935	925	1077	961	814	702	732	785	930
29	883	864	994	982	1002	1047	966	998	927	828	603	553

13	14	15	16	17	18	19	20	21	22	23	24
738	646	624	715	970	1009	1007	847	885	1011	970	1027
727	925	1010	932	970	1079	1062	972	968	984	1071	1095
507	861	783	736	751	753	832	975	1031	1071	1031	1110
1016	859	894	981	1089	1102	1061	957	973	814	724	838
514	588	545	620	742	906	962	954	901	902	939	1071

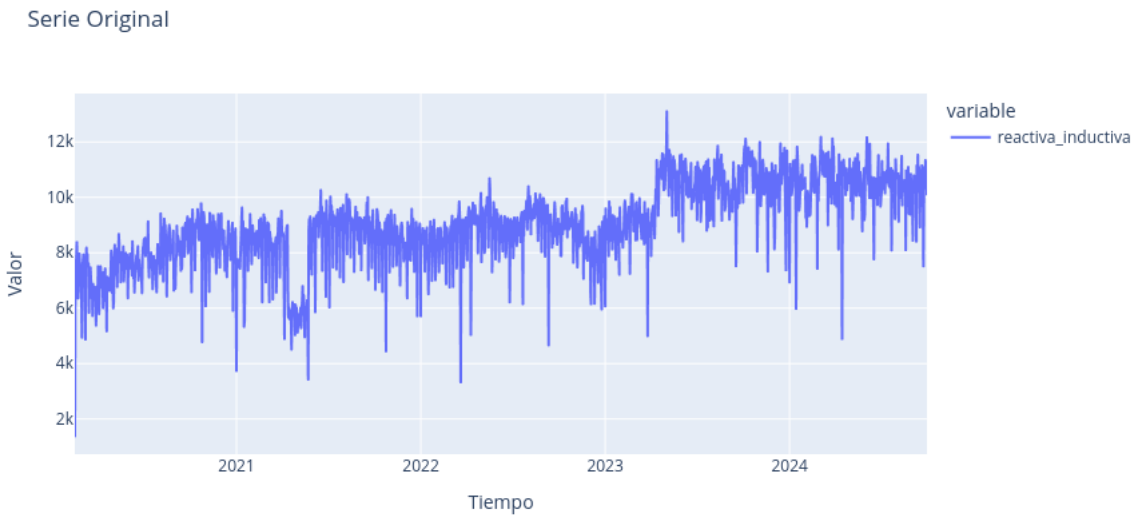

Figura 3.1: Energía Reactiva Inductiva

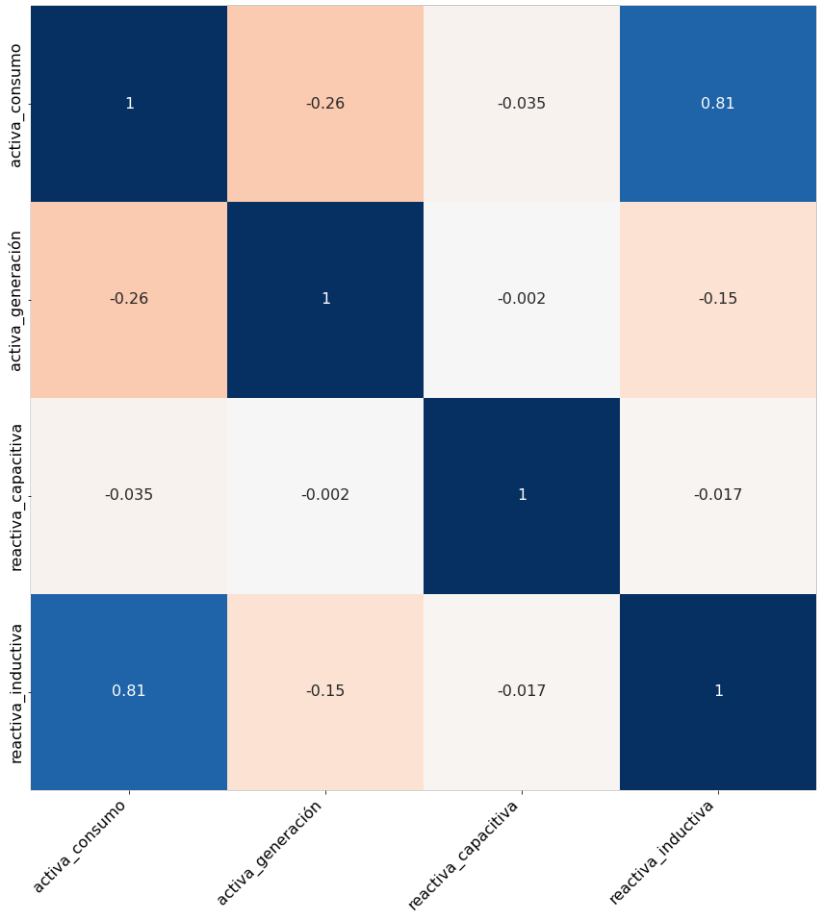

Figura 3.2: Correlación de la energía reactiva

3.3. Analítica descriptiva

Al realizar el análisis de correlación por medio del coeficiente de Pearson en la figura 3.2, se valida lo identificado gráficamente pues se obtuvo, para las variables activa_consumo y reactiva_inductiva, un coeficiente de correlación de 0.81, el cual se considera significativamente alto. Esto nos lleva a analizar la posibilidad de eliminar alguna de las dos variables. Sin embargo se hace necesario contrastarlo con el personal técnico de la empresa para conocer detalles de los expertos antes de tomar la decisión.

De acuerdo con la información disponible en la figura 3.3, se observa para cada una de las variables, lo siguiente: activa_consumo: se observa una posible distribución normal de los datos con media en 17.500. Se identifica la existencia de algunos datos por fuera de los percentiles 25 y 75. activa_generación: se observa una clara concentración de los datos alrededor de cero con presencia de posibles datos atípicos muy alejados de este valor. reactiva_capacitiva: se observa una clara concentración de los datos alrededor de cero. Se identifican 2 posibles valores atípicos en valores de 1.2 y 1.4. reactiva_inductiva: se observa que los datos posiblemente no provengan de una distribución normal, por la presencia de una cola a la izquierda. Se identifican posibles datos atípicos alejados del percentil 25.

Para la identificación de datos atípicos se tomaron aquellos valores que se encuentran un 50 % por encima y por debajo del rango intercuartil de cada una de las variables.

Agregar cuadro de identificación de valores atípicos utilizando el rango intercuartil.

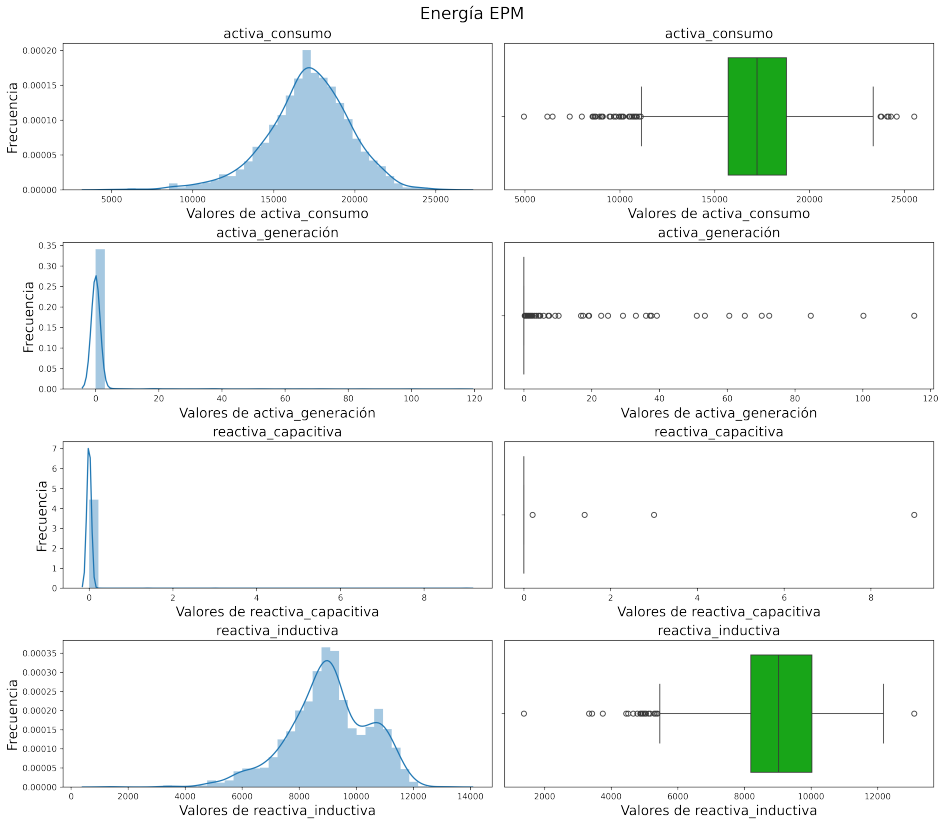

Figura 3.3: Distribución de la energía reactiva

4 Proceso de analítica

En las siguientes secciones se muestra un Pipeline principal, el preprocesamiento realizado a los datos para aplicar los diferentes modelos y el resultado de las métricas enunciadas en **??** para evaluar los modelos

4.1. Pipeline principal

Para estructurar y automatizar el flujo completo de trabajo para la preparación de datos, inicialmente se consolidó con la empresa Leaf IoT la información necesaria respecto al tipo de conexión, los datos disponibles, y el formato en que serían proporcionados. Esto permitió definir las etapas del Pipeline, asegurando que el flujo de procesamiento cumpliera con los requerimientos específicos del proyecto y que los datos estuvieran listos para el análisis y el modelado.

A continuación, se presenta una descripción detallada del proceso realizado, destacando las etapas clave del Pipeline diseñado.

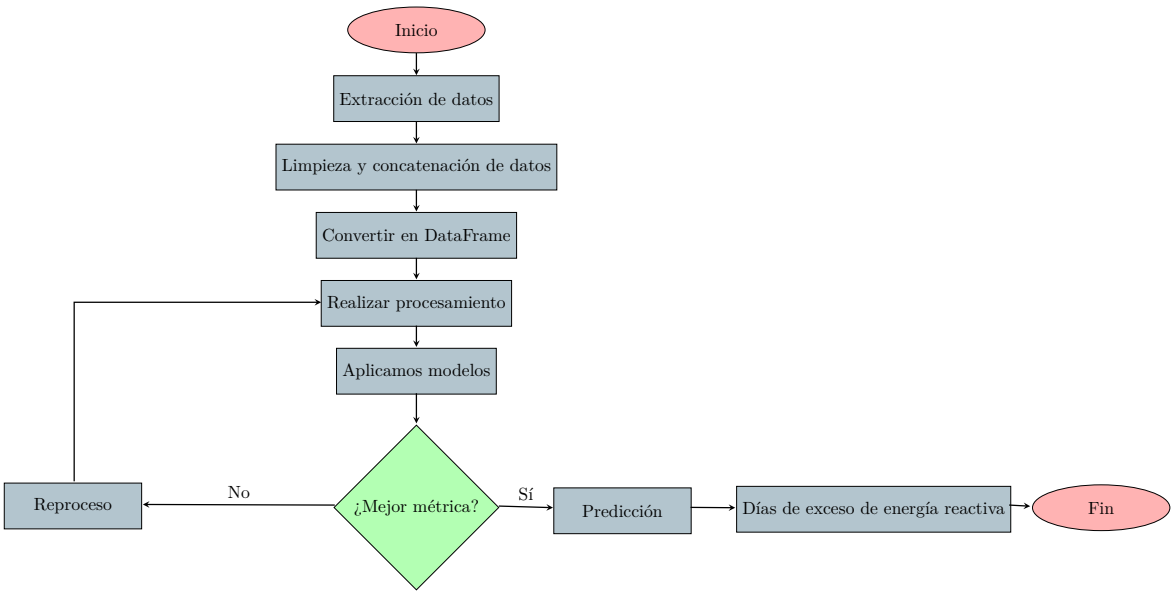

Figura 4.1: Pipeline para obtener el exceso de energía reactiva

El Pipeline mostrado en la figura 4.1 consta de tres bloques principales:

1. **Extracción y transformación de datos:** En esta etapa se recopila la información, se realiza la limpieza de datos para eliminar inconsistencias y se concatenan las distintas fuentes disponibles. Posteriormente, los datos son estructurados en formato tabular DataFrame para facilitar su procesamiento.

2. **Preprocesamiento y modelado:** En esta fase se llevan a cabo las transformaciones necesarias par ajustar los datos al modelo, incluyendo normalización, generación de características y otros pasos relacionados. A continuación, se entrenan y aplican los modelos de aprendizaje automático sobre los datos procesados.

3. **Evaluación y predicción:** Los modelos son evaluados en función de métricas previamente definidas en 6.1, eligiendo el modelo que mejor métricas arroje, para generar predicciones finales y lograr encontrar los días de exceso de energía reactiva.

4.2. Preprocesamiento

Se realizaron varias etapas clave para preparar los datos antes de alimentarlos a la red LSTM Y SimpleRNN, al SARIMAX y K-Means. A continuación, se detalla cada paso y su propósito:

4.2.1. Modelos de Deep Learning

Para los modelos LSTM y SimpleRNN se utilizo la siguiente estructura:

1. **Selección de variables:** Se selecciona una única variable (reactiva_inductiva o activa_consumo) del conjunto de datos para modelar las predicciones. En este caso se asegura que el modelo trabaje con la variable de interés, simplificando el análisis.

2. **Normalización de los datos:** Se utiliza el MinMaxScaler para esalar los valores de reactiva_inductiva al rango $[0, 1]$. Esto asegura que:

   a) Las entradas al modelo tengan valores comparables y estandarizados.

   b) Se evite que valores numéricamente grandes dominen el entrenamiento.

   c) Este paso es crucial para las redes LSTM, ya que son sensibles a la escala de los datos, en el caso de la RNN le permite una mayor convergencia.

3. **Creación de secuencias temporales:** Dado que la LSTM y la RNN requieren datos en forma de secuencias temporales, donde cada muestra contiene un número fijo de observaciones anteriores (timestep) y una etiqueta correspondiente. Esto permite crear ventanas de tiempo para capturar la dependencia temporal.

4. **División del conjunto de datos:** Se dividen el conjunto de datos escalado en:

   a) **Entrenamiento (80 %):** Datos utilizados para ajustar los pesos del modelo.

   b) **Prueba: (20 %):** Datos reservados para evaluar el modelo después del entrenamiento.

   Esta división minimiza el riesgo de sobreajuste.

5. **Reformateo de los datos:** Los datos de entrenamiento, validación y prueba se reformatean para cumplir con el formato esperado (n_samples, timesteps, n_features). Donde n_samples indica el numero de muestras, timesteps como número de pasos de tiempo en cada muestra y n_features que índica el número de características por paso de tiempo.

6. **Desescalado de predicciones:** Las predicciones realizadas por los modelos están en el rango normalizado $[0, 1]$. Para interpretar los resultados se aplica la transformación inversa del MinMaxScaler para devolver las predicciones a su escala original.

4.2.2. Modelo estadístico

Para el modelo estadístico

1. **Selección de variable relevante:** Se selecciona la variable del datasets principal, con el propósito de focalizar el análisis en la variable de interes.

2. **Descomposición de la serie temporal:** Se realiza una descomposición aditiva de la serie temporal en sus componentes:

   - **Tendencia:** Variación de largo plazo en los datos.
   - **Estacionalidad:** Patrones repetitivos en intervalos regulares, por ejemplo de 31 días en nuestras series temporales.
   - **Residuo:** Componente aleatorio o ruido.

   Está descomposición nos ayuda a identificar si el modelo debe considerar factores como estacionalidad o tendencia explícitamente.

3. **Verificación de estacionariedad:** Se utiliza la prueba de Dickey-Fuller aumentada en los datos originales y difenrenciados, con el objetivo de verificar si la serie es estacionaria, ayudandonos a determinar si es necesario aplicar transformaciones adicionales para que los modelos sean válidos.

4. **Diferenciación de la serie:** El objetivo de diferenciar la serie es transformar la serie en estacionaria, eliminando la tendencia. Nos ayuda a la identificación del parametro de diferenciación d en los modelos SARIMAX.

5. **Análisis de autocorrelación (ACF y PACF):** Este análisis nos ayuda a seleccionar los parámetros p y q del modelo SARIMAX.

   a) **ACF (Función de Autocorrelación):** Calcula la correlación entre los valores de al serie y sus rezagos en diferentes lags. Permite identificar el grado de dependecia en los datos.

   b) **PACF(Función de Autocorrelación Parcial):** Determina la correlación entre un rezago específico y los valores actuales, excluyendo la influencia de otros rezagos intermedios.

6. **División de los datos en conjuntos de entrenamiento y prueba** Se dividen los datos en:

   - **Entrenamiento:** Se seleccionan los datos hasta el 2023.

- **Prueba:** Desde el enero del año 2024 a septiembre 30 del 2024.

7. **Selección de hiperparámetros del modelo** Se utiliza pmdarima.auto_arima para determinar automáticamente los parámetros ideales (p,d,q) del modelo. Esto reduce el esfuerzo manual de ajustar hiperparámetros y nos garantiza un modelo ajustado de forma óptima.

8. **Evaluación de residuos:** Se analiza la serie de residuos (diferencia entre los valores reales y predichos) y se verifica que el modelo capture toda la estructura de la serie original.

4.2.3. K-means Clustering

Para el modelo K-means, se realizó:

- **Normalización de los datos:** Se utiliza el MinMaxScaler para escalar los datos a un rango entre 0 y 1. Esto asegura que todas la variables tengan el mismo peso durante el proceso de agrupamiento.
- **Cálculo del coeficiente de Silhouette:** Permite encontrar el número óptimo de clusters, permitiendo la calidad del agrupamiento para los diferentes valores de clusters.

4.3. Modelos

En está sección se muestra los modelos implementados, primeramente el modelo no supervisado K-means, seguido de los modelos estadísticos SARIMAX y DeepLearning SimpleRNN y LSTM.

En la aplicación de los modelos se implento un Grid Search que consiste en realizar una evaluación del modelo para cada valor de un hiperparámetro. Los hiperparámetros son parámetros que no son estimados por el modelo, sino elegidos por el modelador como se explica en [8, p. 66] y un auto arima, que es una librería de python que realiza un tipo de Grid Seach utilizando el criterio de selección AIC ver en [8].

4.3.1. K-Means

El algoritmo K-Means es un método de agrupamiento no supervisado ampliamente utilizado para dividir un conjunto de datos en k grupos o clusteres, minimizando la variabilidad interna dentro de cada grupo. Fue introducido por Lloyd en [10], en el 1982.

El algoritmo intenta minimizar la suma de las distancias al cuadrado entre los puntos de datos y sus centroides asignados. Matemáticamente es:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{4.1}$$

donde C_i representa el conjunto de puntos asignados al clúster i y μ_i es el centroide de C_i.

El algoritmo se aplico a la energía reactiva inductiva para identificar patrones en los perfiles de consumo diario, con número de clusters igual a 3, aplicando el preprocesamiento correspondiente.

4.3.2. SARIMAX

Es un modelo estadístico utilizado para series temporales, basado en el marco de modelos ARIMA y extendido con componentes adicionales. SARIMAX es un acrónimo que significa:

1. **S:** Estacional(Seasonal) permite modelar patrones estacionales en la serie temporal .
2. **AR:** Autorregresivo (Autoregressive) usa valores pasados de la serie para predecir valores futuros.
3. **I:** Integrado (Integrated) - Usa diferenciación para hacer que la serie sea estacionaria (elimina tendencias).
4. **MA:** Media móvil (Moving Average) - Usa el error pasado para ajustar predicciones.
5. **X:** Variables exógenas (Exogenous Variables) - Permite incluir variables externas que pueden influir en la serie temporal.

La formula general del modelo SARIMAX(p,d,q) $\times$ (P,D,Q,s) tiene la forma:

$$\begin{aligned} X_t &= \mu_t + \varepsilon_t + \sum_{i=1}^{p} \phi_i X_{i-1} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \\ &\text{Para las componentes estacionales:} \\ &+ \Phi_1 y_{t-s} + \cdots + \Phi_P y_{t-P_s} + \Theta_1 \varepsilon_{t-s} + \ldots + \Theta_Q \varepsilon_{t-Q_s} \end{aligned} \tag{4.2}$$

Donde:

1. X_t**:** Es el valor de la serie en el tiempo t.
2. μ**:** Es el intercepto.
3. ϕ_i**:** Coeficiente autorregresivos (AR).
4. θ_i**:** Coeficientes de promedio móvil (MA).
5. Φ_i**:** Coeficientes autorregresivos estacionales (SAR).
6. Θ_i**:** Coeficientes de promedio móvil estacionales (SMA).
7. ε_t**:** Término de error en t.
8. s**:** Periodo estacional.
9. d**:** Número de diferenciaciones necesarias para hacer estacionaria la serie.

Al aplicar el auto_arima sobre la energía reactiva y activa los modelos que mejor se adaptaron a los datos fueron: SARIMAX(6,1,1) $\times$ (1,0,1,31) y SARIMAX(2,1,3) $\times$ (0,0,1,31) respectivamente. Cabe resaltar que a la función SARIMAX no se le proporciono una variable exogena, por lo cual el comportamiento del modelo es de tipo SARIMA ver [9, p, 70].

4.3.3. RNN

Las Red Neuronal Recurrente es un tipo de red neuronal diseñada especificamente para procesar datos secuenciales, como señales, texto, audio o cualquier dato donde el orden y las dependencias temporales sean importantes.

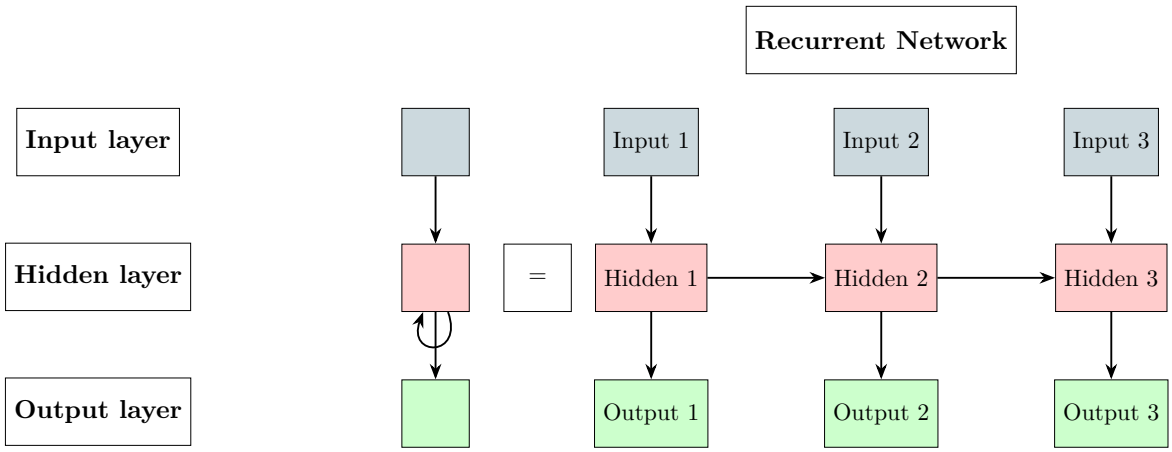

Figura 4.2: Red Recurrente

Cómo se puede ver en la figura 4.3.3 las entradas de una RNN tienen una relación de retroalimentación entre sí que permite que las salidas de pasos anteriores influyan en los actuales.

El modelo RNN$_s$ o Red Neuronal Simple se eligió ya que es adecuado para tareas donde las dependencias temporales no son extremadamente prolongadas. Por está razón se eligió como modelo inicial.

Cuadro 4.1: Arquitectura de la SimpleRNN

Proceso	**Etapas**
Entrada	Secuencias de tamaño 30
Capa SimpleRNN	número de neuronas, activación tanh
Capa de salida densa	1 neurona para regresión

En este modelo la salida y_t realiza un cáculo temporal por otros factores, es decir

$$y_t = W_y \cdot h_t + b_y \tag{4.3}$$

- W_y**:** Matriz de pesos para la salida.

- b_y**:** Vector de sesgos para la salida.

- h_t**:** Estado oculto.

El estado oculto h_t es $\tanh(W_h \cdot h_{t-1} + W_x \cdot x_t + b_h)$, es decir, se inicia la secuencia de tamaño(31) de la muestra en la SimpleRNN y se comienzan hacer los cálculos en los estados ocultos hasta producir la salida y_t.

Para la serie de energía reactiva y energía inductiva se aplica un Grid Search, los cuales tienen entre 10 a 500 neuronas, un batch_size o tamaño de lote de 16, 32, 64 y 120, con un número de epoca de 10 y 20.

Debido a que la SimpleRNN es capaz de capturar patrones básicos en series temporales, al continuar con la implementación de un modelo más avanzado como lo es LSTM.

4.3.4. LSTM

Long Short-Term Memory (LSTM), la celda LSTM introduce memoria a largo plazo de una manera más eficiente porque permite aprender más parámetros. Esto la convierte en la RNN más potente para realizar predicciones, especialmente cuando hay una tendencia a largo plazo en los datos ver [8].

El modelo LSTM pude representarse matemáticamente en términos de sus operaciones internas.

- Cálculo de las puertas:

- Puerta de olvido (Forget Gate):

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f), \tag{4.4}$$

donde

1) f_t**:** vector que indica que información debe olvidarse.
2) W_f**:** pesos de la puerta de olvido.
3) h_{t-1}**:** salida de la celda en el paso anterior.
4) x_t**:** entrada actual.
5) b_f**:** sesgo de la puerta de olvido.
6) σ**:** función sigmoide.

- Puerta de entrada (Input Gate):

$$\begin{aligned} i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ CM_t &= \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \end{aligned} \tag{4.5}$$

Donde

1) i_t**:** vector que controla qué información nueva se almacena en la memoria.
2) CM_t**:** candidata a nueva memoria.
3) W_i, W_c**:** pesos de la puerta de entrada y de la candidata.
4) b_i, b_c**:** sesgos correspondientes.

- Puerta de salida (Output Gate):

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{4.6}$$

- Actualización del estado celular:

$$C_t = f_t \cdot C_{t-1} + i_t \cdot CM_t \tag{4.7}$$

- Cálculo de salida:

$$h_t = o_t \cdot \tanh(C_t) \tag{4.8}$$

Proceso	
Entrada	($n\_samples$, número de pasos en el tiempo, característica)
Capa LSTM	número de neuronas, función de activación
Capa de salida densa	Predicción del consumo energético
Optimización	Optimizador Adam

En ambas series temporales se aplico el modelo LSTM mediante Grid Search con lo siguientes hiperparámetros:

- Número de neuronas de la capa LSTM (neurons = 10, 50, 100, 500).
- Tasa de aprendizaje (learning_rate = 0.001, 0.01).
- Tamaño de lote (batch_size = 16,32).
- Número de épocas (epochs = 10,20).

4.4. Métricas

Para evaluar los modelos empleados en la predicción de la energía reactiva e activa se utilizaron las métricas MAE, MAPE y RMSE.

La siguiente figura nos muestra las métricas utilizados en los diferentes modelos:

1. Error absoluto medio (MAE): Nos arroja la diferencia promedio absoluta entre los valores pronosticados y los valores reales. Es decir que tan distante están los valores pronosticados de los valores reales.

$$MAE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n-1} |y_i - \hat{y}_i|, \tag{4.9}$$

2. Error cuadrático medio (RMSE): Similar al MAE, pero considera los errores al cuadrado, penalizando más los errores grandes. Un RMSE bajo indica mayor precisión.

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n-1}(y_i - \hat{y}_i)^2}, \tag{4.10}$$

3. Error porcentual absoluto medio (MAPE): Expresa el error en términos porcentuales, siendo útil para comparar series de tiempo con diferentes escalas.

$$MAPE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n-1} \left|\frac{y_i - \hat{y}_i}{y_i}\right|. \tag{4.11}$$

Para el modelo K-Means ajustado a los datos de energía reactiva, se calculó el indice de Calinski-Harabasz, dio 949.342 lo que indica una buena separación entre clusters.

5 Metodología

5.1. Baseline

La metodología Baseline en el contexto del aprendizaje automático, se refiere a la creación de un modelo de referencia básico y sencillo contra el cual se pueden comparar otros modelos más complejos. El objetivo de un baseline es establecer un punto de partida o referencia para evaluar el rendimiento de otros métodos.

En la siguiente gráfica se evidencia los resultados de las metrícas aplicados a los diferentes modelos en la sección 4.3. Se toma como linea base el modelo SARIMAX, debido a que este modelo presento errores considerables con las métricas de MAE Y RMSE, se considero, redes neuronales recurrentes LSTM y SimpleRNN. las cuales presentaron un mejor comportamiento con respecto a los datos por los erros arrojados por las métricas, como se puede observar en la figura 6.1.

5.2. Validación e iteraciones

Se validaron los resultados con equipo técnico de Leaf Iot, debido a que en la ventana de tiempo de marzo a junio en el 2021, como se puede apreciar en el gráfico **??**, se nota una caída. Al realizar la limpieza los datos estos aparecían como outliers. Se realizaron reuniones y de estás se concluyo que las anomalías se debían a problemas técnicos asociados, mantenimientos, entre otras factores.

Para validar los los modelos se aplico la estrategía train-test split ver [8], donde la data de test representa un 20 % del conjunto de datos.

Comparación visual entre los datos de test y los datos de predicción.

De las figuras 5.1 y 5.2 se puede validar como se comporta la predicción con respecto al conjunto de datos real.

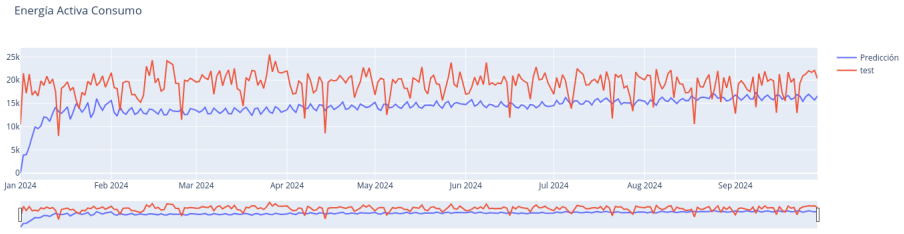

Figura 5.1: Energía Activa Consumo: test y predicción

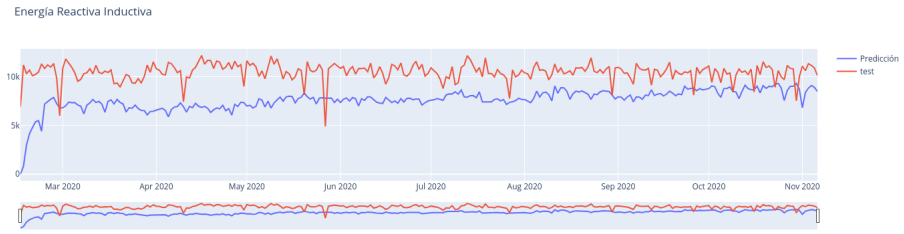

Figura 5.2: Reactiva inductiva: test y predicción

5.3. Herramientas

Las herramientas utilizadas en la monografía son:

- Overleaf.
- Excel.
- Google Colab.
- VS Code.
- Power BI.
- Copilot, ChatGPT y Gemini.
- Github como repositorio.
- Git como control de versiones.
- Python.
- LaTeX.

6 Resultado y discusión

En este capítulo mostraremos los resultados obtenidos de los diferentes modelos aplicados ver 4.3 y sobre el mejor modelo se analiza la predicción para el mes siguiente de la cantidad de días con exceso de enería reactiva.

6.1. Métricas

Cuadro 6.1: Métricas por energía

Métricas de desempeño					
Energía	**Modelo**	**Tipo**	**MAE**	**MAPE**	**RMSE**
Reactiva	SARIMAX	Train	548.48	6.92	793.57
		Test	2911	27.49	3177.17
	SimpleRNN	Train	685.33	7.88	885.58
		Test	734.93	7.82	951.17
	LSTM	Train	620.16	7.68	891.67
		Test	580.99	6.58	764.85
Activa	SARIMAX	Train	1390	9.21	1949.55
		Test	5195	26.68	5786.52
	SimpleRNN	Train	1516.33	9.32	1991.97
		Test	1636.83	9.42	2163.35
	LSTM	Train	1433.33	9.42	1999.94
		Test	1439.04	8.86	1902.93

De la tabla 6.1 se obtiene que el modelo que mejor se adapta a los datos de energía reactiva inductiva y energía activa es el modelo LSTM, pues se evidencia, que el resultado del MAPE para este modelo sobrestima o subestima el valor real en porcentajes inferiores a los demás modelos, lo que sugiere que este modelo está realizando predicciones bastante precisas, tanto para la energía activa como para la energía reactiva inductiva.

6.2. Evaluación cualitativa

Se presentan los resultados para los dos modelos seleccionados: LSTM y K-Means.

6.2.1. Clustering para perfil de generación de energía reactiva usando K-means

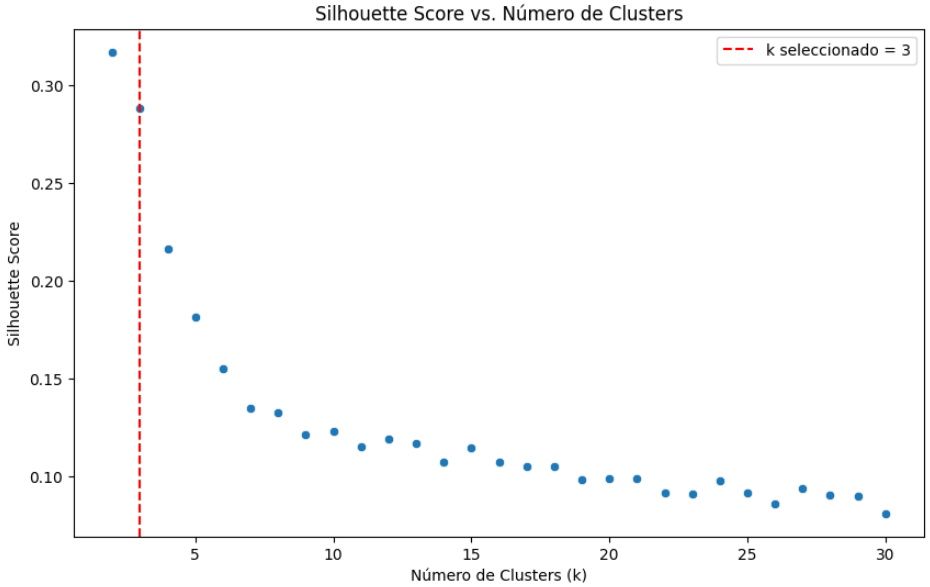

Figura 6.1: Silhouette Score vs Número de Clusters

El gráfico de 6.1 indica que $k = 2$ maximiza el coeficiente de silueta, sugiriendo la mejor segmentación de los datos en dos grupos. Sin embargo, al considerar la necesidad de identificar subgrupos más detallados y comprender mejor los patrones de generación de energía reactiva, se optó por utilizar tres clusters ($k = 3$). Esta elección permite revelar una granularidad mayor en el análisis, proporcionando insights más precisos sobre las características de cada grupo.

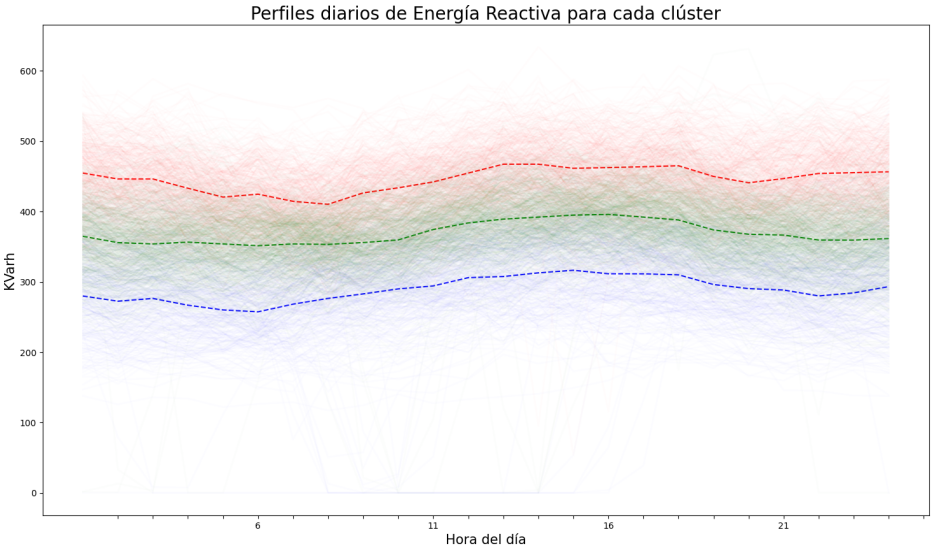

Figura 6.2: Perfil de Energía

El gráfico 6.2 muestra tres clusters diferenciados por color (rojo, azul y verde), cada uno representando un patrón de consumo de energía reactiva a lo largo de las 24 horas del día. Las líneas representan los perfiles individuales de cada día dentro de un clúster, mientras que las líneas discontinuas representan la mediana del consumo para cada hora dentro de ese

clúster.

El clúster con mayor mediana de energía reactiva a lo largo del día es el rojo, indicando valores de energía reactiva más elevados en comparación con los otros dos.

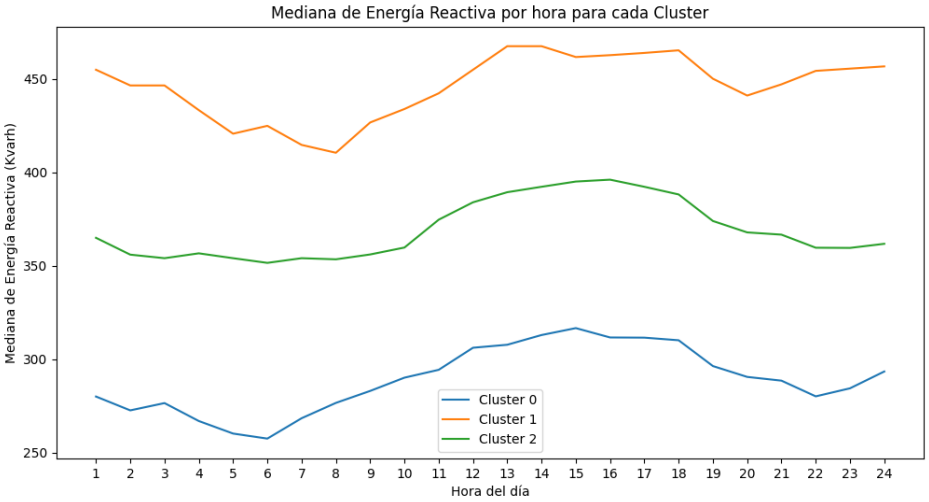

Figura 6.3: Mediana de Energía Reactiva por hora para cada Clúster

El gráfico compara los perfiles de carga horaria mediana (en kVArh) de los tres clusters, evidenciando una clara diferenciación en los patrones de generación de energía reactiva. El Clúster 1 se destaca por tener mayor valores de energía reactiva, con picos pronunciados durante las horas laborales (10:00 a 18:00) y nocturnas (21:00 a 01:00). Estos resultados sugieren la presencia de procesos productivos intensivos en energía reactiva durante estos periodos.

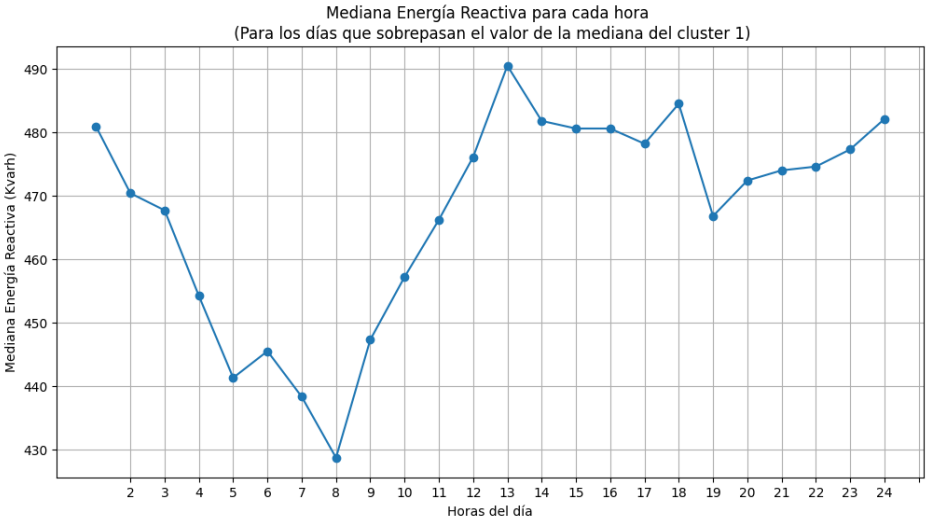

Figura 6.4: Días que superan las medianas del clúster 1 de la Energía Reactiva Inductiva

El gráfico muestra la energía reactiva mediana para cada hora del día, calculada únicamente para los días de la serie de tiempo que excedieron la mediana del clúster de mayor generación de energía reactiva (cluster 1). Con ello, se identifica un patrón idiario caracterizado por picos entre la 1 p.m. y la 1 a.m. Estos periodos sugieren la existencia de procesos o actividades específicas que requieren un mayor control.

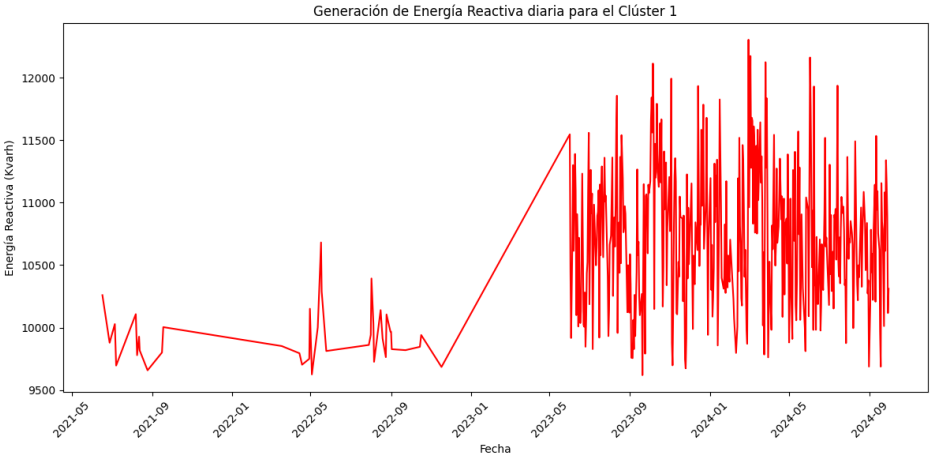

Figura 6.5: Generación de Energía Reactiva diaria para el Clúster 1

El gráfico de generación de energía reactiva del Clúster 1 evidencia una alta variabilidad a lo largo del tiempo, con picos y valles pronunciados. Particularmente se observa, un aumento significativo en la frecuencia de días clasificados en el clúster de mayor generación de energía reactiva a partir de mayo de 2023. Este incremento sugiere la posible implementación de nuevos equipos o cambios en los procesos productivos a partir de esa fecha, lo cual podría estar influyendo directamente en la generación de energía reactiva.

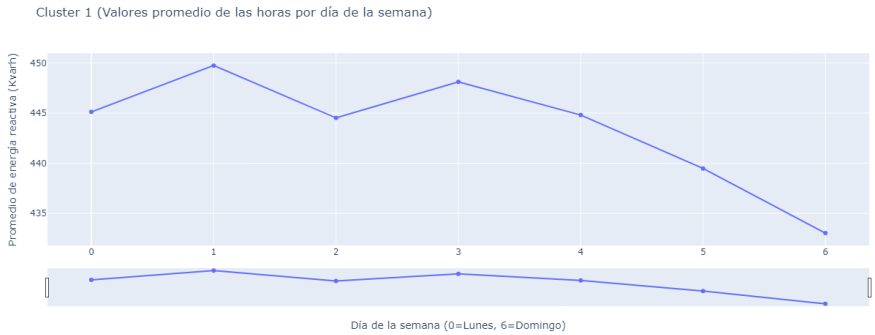

Figura 6.6: Valores promedio de las horas de la por día de la semana del Clúster 1

El gráfico 6.6 muestra el promedio de energía reactiva consumida a lo largo de cada día de la semana para el clúster 1. El eje X representa los días de la semana (0 para lunes, 6 para domingo), y el eje Y muestra el promedio del consumo de energía reactiva en Kvarh. De aquí se identifica que los martes y jueves, son los días de mayor generación de energía reactiva.

6.2.2. Exceso de Energía Reactiva usando LSTM

Se sabe del planteamiento del problema ver en 1.1, que los días de exceso de generación de energía reactiva debe cumplir dos condiciones principales:

$$\begin{aligned} &\text{-Energía reactiva inductiva } > 50\,\% \text{ de Energía activa.} \\ &\text{-Energía Reactiva Capacitiva} > 0 \end{aligned} \tag{6.1}$$

Con relación a la energía reactiva capacitiva se tiene el siguiente gráfico

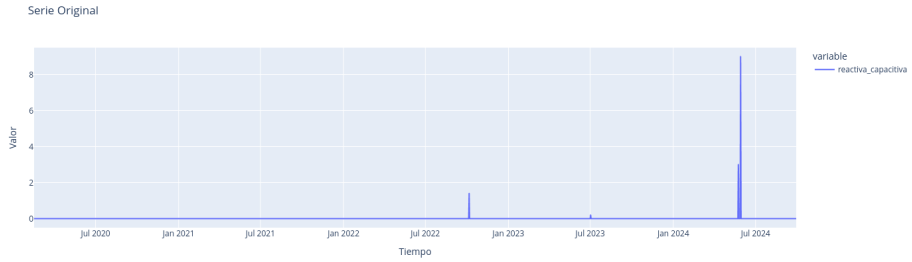

Figura 6.7: Energía Capacitiva

Se puede apreciar que el valor de la energía capacitiva, en su mayoría es cero, lo que dificulta la predicción de la energía reactiva capacitiva. Por lo anterior las predicciones se realizaron sobre la energía reactiva inductiva y la energía activa.

Al aplicar el modelo LSTM para pronosticar los valores de la energía activa y reactiva inductiva se obtienen los siguientes gráficos para 30 días.

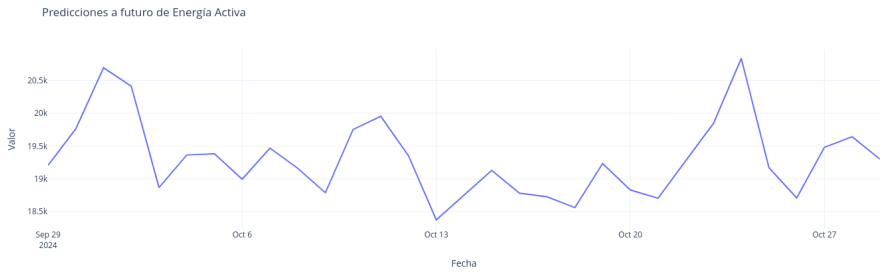

Figura 6.8: Proyección a 30 días de la Energía Activa

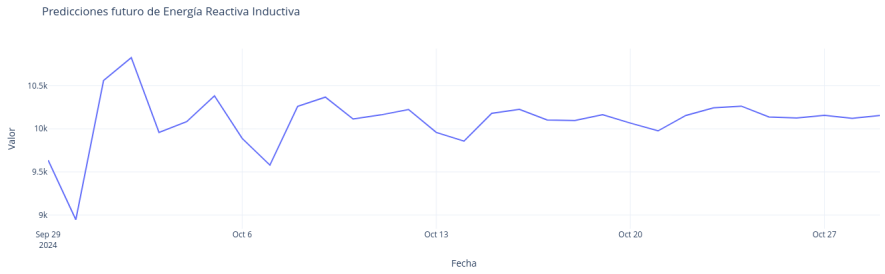

Figura 6.9: Proyección a 30 días de la Energía Reactiva Inductiva

Para identificar el exceso de energía reactiva para el mes de octubre de 2024 se realizó el cálculo de la condición 1 de 6.1, encontrando que en 28 de los 31 días, la energía reactiva inductiva supera al 50 % de la energía activa, lo que implica que hay exceso de energía reactiva en más de 11 días contemplados en la normativa de la CREG [5] y posterior mente una sanción para dicho mes.

6.3. Consideraciones de producción

Para la producción con el equipo Leaf IoT, se encuentra en proceso el despliegue de un modelo para predecir el exceso de energía reactiva. La plataforma seleccionada para este propósito es AWS.

Actualmente, se están resolviendo dudas relacionadas con los costos de producción y estrategía a elegir, se propuso usar servicios como Lambda, S3, Glue y Athena. La propuesta consiste en orquestar la ejecución del modelo mediante Lambda, almacenar las predicciones en S3 y utilizar Glue para estructurar los datos en tablas que puedan ser consumidas directamente en Power BI, conectándose a Athena.

7 Conclusiones

- El modelo seleccionado LSTM tiene un mejor desempeño general en comparación con SARIMAX y RNN, mostrando para la predicción de la energía reactiva y energía activa es LSTM, pues en ambos tipos de energía ofrece la mejor precisión y capacidad de generalización.
- Si bien el número de días clasificados en el clúster de mayor generación de energía reactiva ha aumentado considerablemente desde mayo de 2023, se requieren análisis adicionales para determinar si este incremento se debe a factores estacionales, cambios en la demanda energética o a la introducción de nuevos equipos en la planta.
- La aplicación de modelos como LSTM para predecir la cantidad de días de exceso de energía reactiva, puede contribuir a mejorar la planeación del proceso productivo en planta. Lo anterior, balanceando las horas y los días identificados como potenciales para exceder los valores permitidos.

8 Recomendaciones

Debido a que actualmente se adelanta un proyecto de calibración de sensores para medición de energía reactiva, se recomienda realizar nuevos modelos con los datos arrojados por los equipos y comparar con los datos suministrados por la empresa prestadora de servicios de energía. Lo anterior con el fin de reducir los errores en los modelos actuales y mejorar las predicciones de las energías reactiva inductiva y activa.

Bibliografía

[1] D. Peña Sánchez de Riviera, *Análisis de series temporales*. 2ª ed. Madrid, España: Alianza Editorial, 2010.

[2] "Metrics and scoring", scikit-learn: Machine Learning in Python, 1.5 documentation. [Online]. Disponible: https://scikit-learn.org/1.5/modules/model_evaluation.html[Fecha de acceso: 13-Nov-2024].

[3] Celsia, *"Conoce todo lo que debes saber sobre energía reactiva: ¿Por qué se refleja en tu factura la energía reactiva?,"* [En línea]. Disponible: https://www.celsia.com/en/\blog-celsia/conoce-todo-lo-que-debes-saber-sobre-energia-reactiva-por-que-se-refleja\-en-tu-factura-la-energia-reactiva/. [Accedido: 19-abr-2024].

[4] Comisión de Regulación de Energía y Gas (CREG), *"Resolución CREG 15 de 2018," 03-feb-2018.* [En línea]. Disponible: https://gestornormativo.creg.gov.co/gestor/entorno/docs/resolucion_creg_0015_2018.htm. [Accedido: 19-abr-2024].

[5] Comisión de Regulación de Energía y Gas (CREG), *"Resolución CREG 199 de 2019," 20-ene-2020.* [En línea]. Disponible: https://gestornormativo.creg.gov.co/gestor/entorno/docs/resolucion_creg_0199_2019.htm. [Accedido: 19-abr-2024].

[6] Comisión de Regulación de Energía y Gas (CREG), *"Resolución CREG 195 de 2020," 22-oct-2022.* [En línea]. Disponible: https://gestornormativo.creg.gov.co/gestor/entorno/docs/resolucion_creg_0195_2020.htm. [Accedido: 19-abr-2024].

[7] Empresas Públicas de Medellín (EPM), *"Energía reactiva," 2023.* [En línea]. Disponible: https://www.epm.com.co/clientesyusuarios/energia/tarifas-energia/energia-reactiva.html. [Accedido: 19-abr-2024].

[8] Korstanje, J.(2021). Ethe SARIMAX Model. In: Advanced Forecasting with Python.Apress, Berkley, CA. https://doi.org/10.1007/978-1-4842-7150-6_8

[9] Chatfield, C. (2024). *The Analysis of Time Series: An introduction, 6th ed.,* Chapman & Hall/CRC, Boca Raton, Fla.

[10] Llyod, S. P. (1982). *Least squares quantization in PCM. IEEE Transactions on Information Theory***28**(2), 129-137.