

**Universidad Nacional De La Matanza**  
**Especialización Ciencia de Datos**

**Minería de Datos**  
**Profesora: Lorena Matteo**

# **Análisis y Predicción de la Demanda Energética del SADI**

## **Integrantes Grupo A:**

Barvagelata, Julián Mariano  
Fica Millán, Yesica Verónica  
González De Rose, Franco Ezequiel  
Gotte, Joaquín Ezequiel  
Miranda Quisbert, Brian Alex  
Petraroia, Franco Albano

## CONTENIDO

Introducción al Caso de Estudio .....	4
Dataset asociado .....	9
Objetivo de negocio e hipótesis .....	9
Parte 1 - Selección, Preprocesamiento y Transformación de Datos .....	11
1.1. Descripción del dataset .....	12
1.2. Carga del Dataset.....	12
1.3. Preprocesamiento y limpieza de datos .....	12
1.4. Análisis exploratorio y visualizaciones .....	13
1.5. Transformación de datos.....	17
1.6. Técnicas gráficas utilizadas.....	20
Parte 2 - Modelos Predictivos y Técnicas de Evaluación.....	24
2.1. Regresión Lineal.....	25
2.1.1. Conclusiones preliminares.....	28
2.2. Flujo de Regresión con Estructura de Metanodos .....	29
2.2.1. Regresión polinómica .....	31
2.2.2. Árbol de regresión .....	35
2.2.2.1. Conclusiones preliminares.....	37
2.2.3. Gradiente Boosted Trees.....	38
2.2.3.1. Conclusiones preliminares.....	40
2.3. Comparación de modelos de predicción .....	41
2.4. Clasificación Multiclasa: Comparación Naive Bayes - Random Forest .....	43
2.4.1. Conclusiones preliminares.....	48

2.5. Conclusiones Modelos Predictivos .....	49
Parte 3 - Modelos Descriptivos y Técnicas de Evaluación .....	51
3.1. Clustering.....	52
3.1.1. Análisis de la técnica del codo .....	52
3.1.2. K-means, Silhouette y Entropía .....	53
3.1.3. Conclusiones preliminares.....	58
3.2. PCA.....	59
3.2.1. Conclusiones preliminares.....	61
3.3. Reglas de Asociación .....	63
3.3.1. Conclusiones preliminares.....	65
3.4. Conclusiones Modelos Descriptivos .....	67

## INTRODUCCIÓN AL CASO DE ESTUDIO

CAMMESA (Compañía Administradora del Mercado Mayorista Eléctrico S.A.) es una empresa argentina de gestión privada, con propósito público y sin fines de lucro, encargada de administrar y coordinar el Mercado Eléctrico Mayorista (MEM).

Fue creada en 1992 como parte de la reforma del sector eléctrico en Argentina. El 80 % del paquete accionario de CAMMESA es propiedad, por partes iguales, de las asociaciones que agrupan a los distintos Agentes del Mercado Mayorista Eléctrico (Agentes Generadores, Transportistas, Distribuidores y Grandes Usuarios). El 20% restante está en poder del ministerio público (la Secretaría de Energía) que asume la representación del interés general y de los consumidores que son abastecidos por los Agentes Distribuidores.

### Objetivos Generales:

- Maximizar la seguridad del Sistema y la calidad de los suministros y minimizar los precios mayoristas en el mercado horario de energía.
- Prever y programar eficientemente el funcionamiento del MEM y del SADI (Sistema Argentino de Interconexión).
- Operar el SADI y administrar el MEM con objetividad y máxima transparencia dentro del marco de las reglamentaciones del MEM.
- Mantener un proceso de mejora continua.

### Servicios Brindados:

- Despacho técnico-económico del SADI.
- Supervisión de la Seguridad y Calidad de funcionamiento del SADI.
- Valorización de las transacciones económicas en el Mercado SPOT y en el Mercado a Término.
- Gestión de Facturación, Cobranza, Pagos y Operación Financiera de los Fondos del Mercado.
- Servicios Adicionales: información, Administración de Contratos, Prospectiva, Gestión de ingreso de nuevos agentes, etc.

En nuestro caso de estudio, nos centraremos en determinar el despacho de generación de energía a partir de predicciones de potencia pico y demanda energética, basándonos en datos históricos reales.

### *¿Por qué es importante conocer y/o predecir la demanda y, en función de esta, determinar la generación de energía?*

Para contar con un suministro de energía constante y de calidad, se deberá mantener en todo momento el balance entre generación y demanda, así como también la calidad de los parámetros eléctricos involucrados (frecuencia y tensión constante, pureza de la forma de onda, equilibrio entre las fases).

Para lograr dicho balance, el sistema eléctrico utiliza distintos mecanismos de control. A continuación, se describen dos de los más importantes:

- Control de frecuencia:

La energía eléctrica a gran escala no puede ser almacenada; debe ser generada en el mismo instante en que es demandada por los usuarios. La frecuencia puede considerarse un indicador del seguimiento de la generación con respecto a la demanda, a partir de la siguiente ecuación:

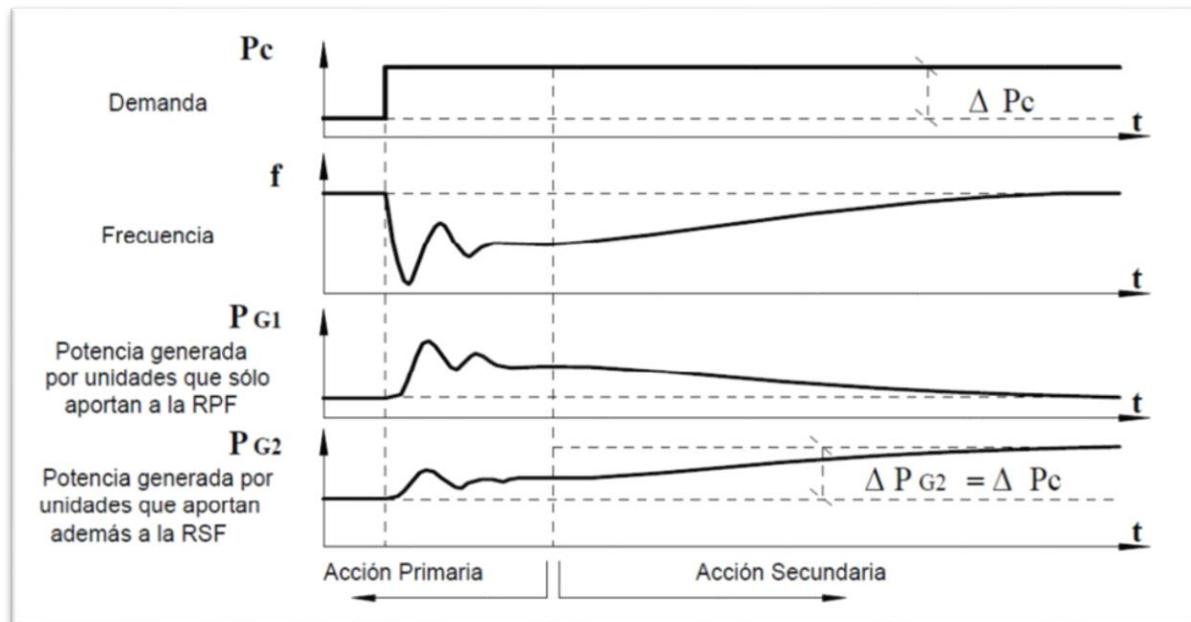
$$P_m' - P_{cons}' = P_a' = T_a \frac{\partial f}{\partial t}$$

La **frecuencia** del sistema eléctrico funciona como un **indicador del equilibrio** entre la generación y la demanda:

- Si la generación es mayor que la demanda, la frecuencia **aumenta**.
- Si la demanda es mayor que la generación, la frecuencia **disminuye**.

Por eso, monitorear la frecuencia permite comprobar si el sistema está generando la cantidad correcta de energía en tiempo real.

La regulación de frecuencia tiene las siguientes etapas:



Como se puede observar en este gráfico, ante un aumento abrupto de la demanda ( $P_c$ ), la frecuencia disminuye significativamente y se requerirá generación ( $P_g$ ) para compensarla. Para equilibrar los requerimientos variables del consumo, los generadores cuentan con mecanismos de control llamados Regulación Primaria de Frecuencia (RPF) y Regulación Secundaria de Frecuencia (RSF).

- Control de tensión:

Los procedimientos establecen para los transportistas la obligatoriedad de mantener la tensión de su red lo más próxima posible a los valores nominales y dentro de los siguientes rangos en condiciones de operación normal:

- 500 kV:  $\pm 3\%$  → entre 485 y 515 kV
- 220 kV:  $\pm 5\%$  → entre 209 y 231 kV
- 132 kV:  $\pm 5\%$  → entre 125,4 y 138,6 kV
- 66 kV:  $\pm 7\%$  → entre 61,38 y 70,62 kV
- Para otras tensiones de la red del SADI, se admite una variación de hasta  $\pm 10\%$ .

Existen muchos elementos para el control de tensión, pero en nuestro caso de estudio **nos centraremos en el análisis desde el punto de vista de la generación.**

La relación entre la caída de tensión en un nodo y su tensión nominal se aproxima a la relación entre la potencia reactiva entregada por el generador y su potencia aparente de cortocircuito.

Si no se tienen en cuenta estos factores mencionados, se podría comprometer la estabilidad del SADI y llevarlo al colapso, averiando los equipos asociados.

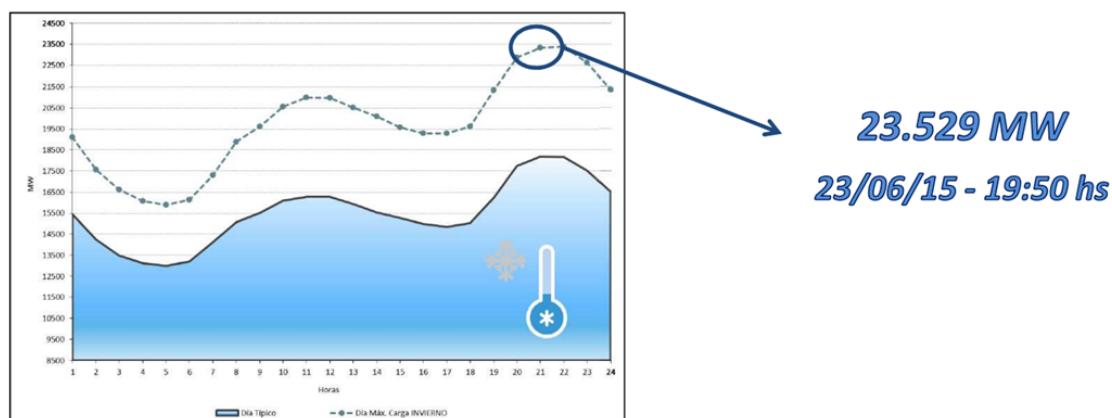
### **¿Cómo podemos predecir la demanda?**

La demanda depende principalmente de los factores climáticos. Teniendo en cuenta que dos tercios de la demanda del SADI se concentran en el área de Gran Buenos Aires, el área Buenos Aires y el área Litoral, se tomará como referencia la temperatura media diaria del área del GBA.

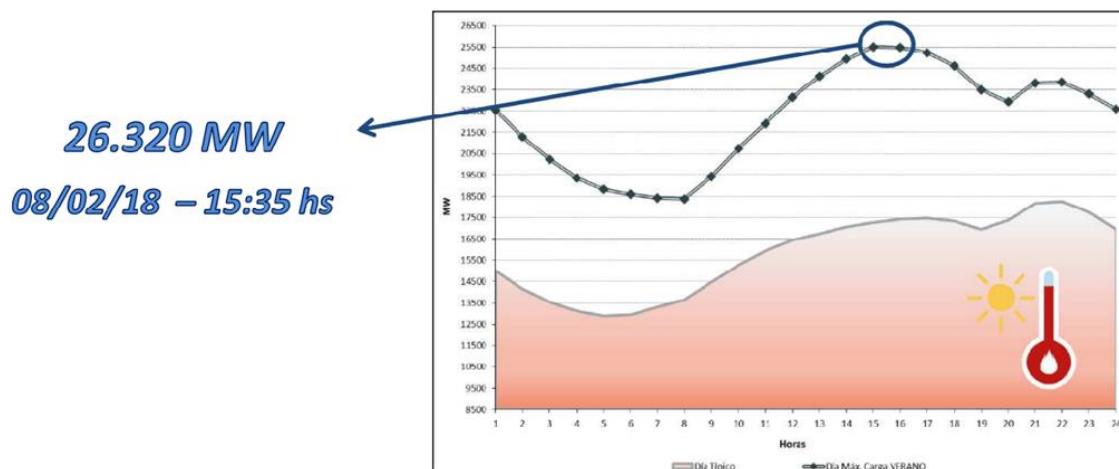
Otros factores a tener en cuenta:

- Estación del año:

En el análisis se destacarán únicamente los días de verano e invierno. En invierno los picos de potencia suelen aparecer en el horario nocturno (entre las 18:00 y las 23:00 horas), y los máximos se registran en los días de menor temperatura.



En verano, los picos de potencia ocurren generalmente al mediodía, y los valores máximos se dan durante los días de mayor temperatura.

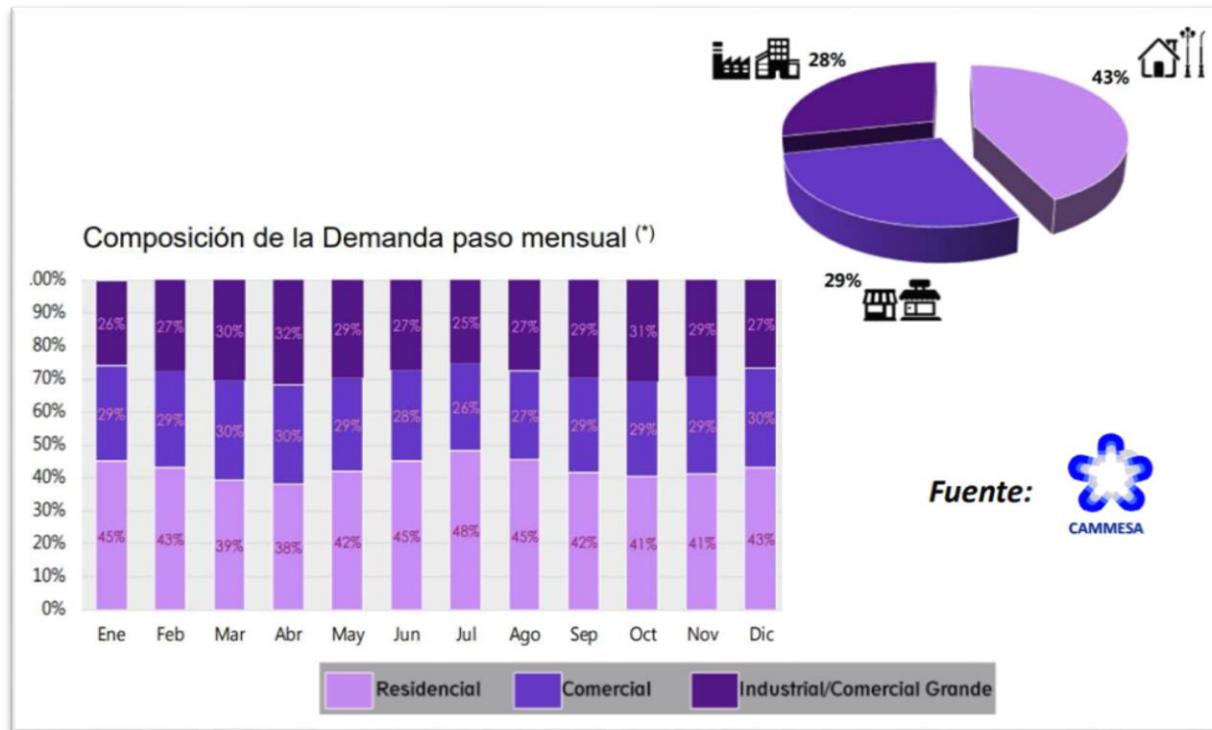


- Tipo de día y horario:

La demanda está compuesta principalmente por los sectores residencial, comercial e industrial, y varía según el tipo de día y el horario. En la siguiente tabla se resumen los patrones de consumo de acuerdo con el tipo de usuario y la condición del día (hábil o no hábil).

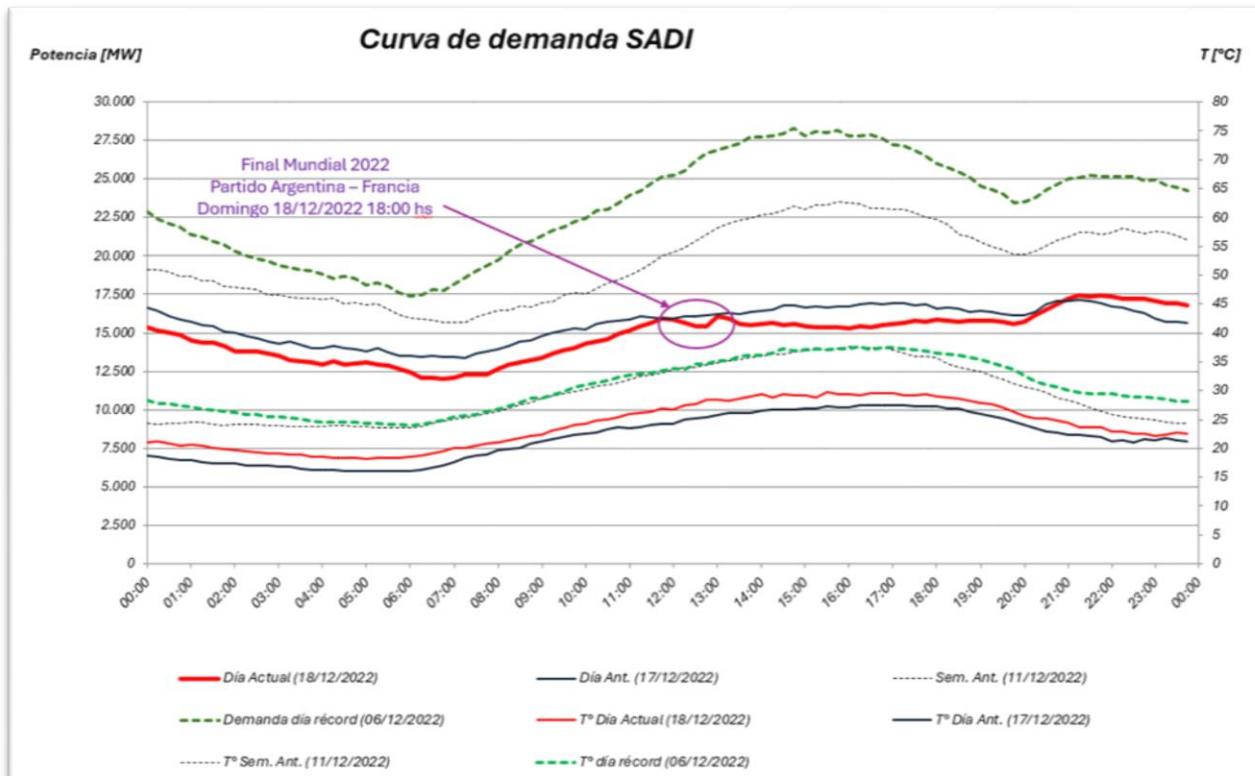
TIPO DE CONSUMO	DÍAS HÁBILES	DÍAS NO HÁBILES
RESIDENCIAL	Pico de demanda entre <b>18:00 y 23:00 hs.</b> Uso de iluminación, electrodomésticos y aire acondicionado.	Pico sostenido durante todo el día. Mayor uso de electrodomésticos, climatización y ocio electrónico.
COMERCIAL	Picos de demanda entre <b>9:00 y 20:00 hs.</b> Coincidirán con apertura y cierre de locales, oficinas, centros comerciales, etc.	Demandas reducidas. Solo se mantiene en supermercados y servicios esenciales.
INDUSTRIAL	Picos de demanda entre <b>8:00 y 18:00 hs.</b> Asociado a actividades productivas.	Demandas muy reducidas. Excepción: plantas con procesos continuos.

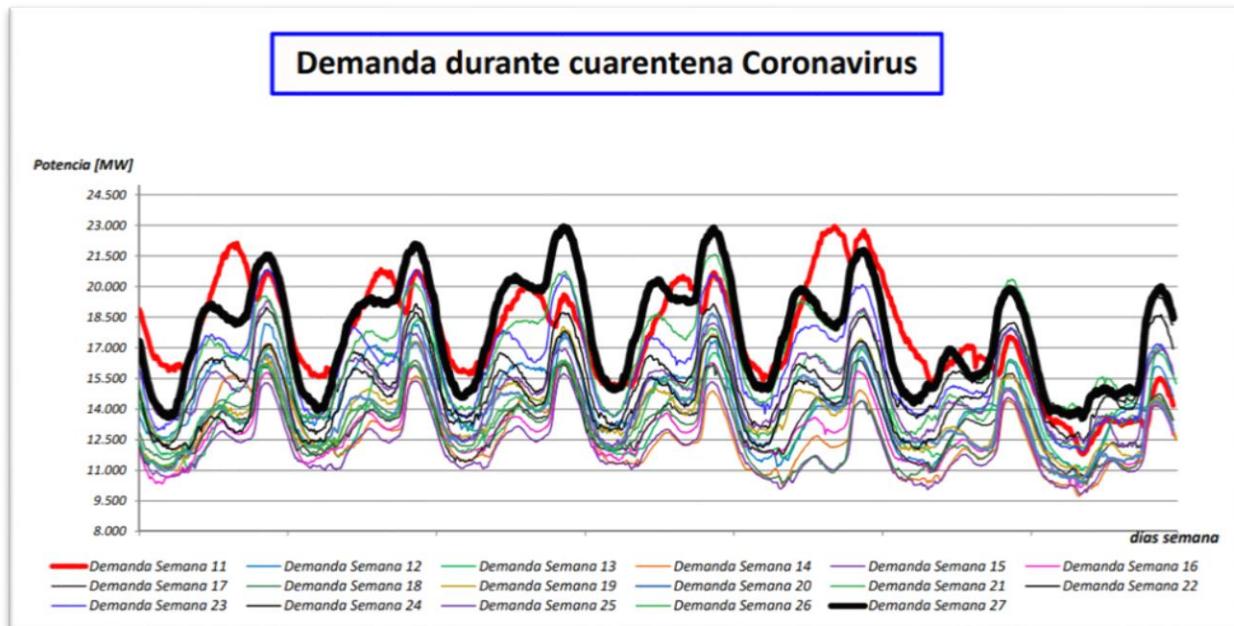
También se debe considerar que la demanda eléctrica varía según el tipo de usuario y la época del año. En la siguiente imagen se puede observar la distribución porcentual mensual de la demanda entre los diferentes sectores antes mencionados, y su composición anual global.



- Anomalías en la demanda:

Dependiendo el día y el horario, pueden registrarse ciertas anomalías en la medición de energía y potencia. Por ejemplo:





## Dataset asociado

El dataset utilizado fue descargado desde la web de CAMMESA, en la sección *Máximos Históricos de Energía y Potencia Estacionales del SADI* (<https://cammesaweb.cammesa.com/2023/03/14/maximos-historicos-de-energia-y-potencia-estacionales/>).



Históricos valores de energía y Potencia.xls:

## OBJETIVO DE NEGOCIO E HIPÓTESIS

Se busca predecir el pico máximo de potencia y la energía diaria demandada, considerando factores como condiciones meteorológicas, el tipo de día (hábil/no hábil) y los patrones de consumo históricos. **Esta proyección permitirá determinar la generación necesaria para despachar, con el fin de garantizar el abastecimiento del sistema eléctrico.**

Hipótesis:

- La temperatura afecta al consumo de energía eléctrica en el SADI.
- En verano, el aumento de temperatura tiene mayor impacto en el consumo que en invierno.
- La relación temperatura-consumo es más fuerte en días hábiles que en fines de semana.
- Los días con temperaturas extremas (altas o bajas) presentan valores anómalos de potencia pico.

Se explorarán diversos modelos de predicción, tanto de regresión como de clasificación, a fin de responder las siguientes preguntas:

- ¿Es posible evitar costos innecesarios de generación o importación de energía alcanzando un punto de equilibrio entre generación y demanda?
- ¿Con qué error se logra estimar la potencia pico en el mejor de los casos a fin de no saturar el sistema?
- ¿La complejidad del caso de estudio requiere el uso de modelos complejos para la predicción?
- ¿El problema a resolver se ajusta mejor a un enfoque de regresión o de clasificación?
- ¿Podemos predecir si la demanda será baja, media o alta en base a variables temporales y estacionales (tipo de día, temperatura media, estado del tiempo, entre otras)?
- ¿Qué variables son más influyentes para anticipar un nivel alto de demanda energética?
- ¿Qué tan bien puede predecirse la categoría de "Nivel Demanda Potencia Pico" utilizando diferentes modelos de clasificación (Bayes y Random Forest)?
- ¿El balanceo de clases con SMOTE mejora la precisión del modelo Random Forest?

En la exploración de los Modelos Descriptivos, se buscará responder:

- ¿Cómo se agrupan los días según consumo de energía (GWh) y potencia pico (MW)?
- ¿Qué factores están asociados a los picos de potencia máxima?
- ¿Hay tendencias de crecimiento anual en la demanda?

## **PARTE 1 - SELECCIÓN, PREPROCESAMIENTO Y TRANSFORMACIÓN DE DATOS**

## 1.1. Descripción del dataset

Los datos que contiene el dataset, junto con sus respectivos tipos, se detallan a continuación:

DATO	TIPO DE DATO	DESCRIPCIÓN
AÑO	int	Año del registro, desde el año 2007 al año 2025
MES	date	Indica los diferentes meses en formato “mm-yy”
Nº MES	int	Índice numérico por mes
VERANO/INVIERNO	str	Agrupa la temporada de primavera y verano como “verano” y otoño e invierno como “invierno”
SEMANA	int	Indica número de semana
FECHA	date	Fecha específica del registro en formato “yyyy/mm/dd”
TIPO DIA	str	Indica si el día es hábil, sábado, domingo o feriado
DIA	str	Indica día de la semana (lunes, martes, etc.)
Nº DIA	int	Indica el número de día dentro del mes (del 1 al 31)
ENERGIA SADI	float	Consumo de energía total del sistema SADI en GWh
POTENCIA PICO SADI	float	Potencia máxima registrada en el día para SADI en MW
HORA POTENCIA PICO	float	Hora exacta en la que se alcanzó la potencia pico
TEMPERATURA MEDIA DIARIA GBA (°C)	float	Temperatura promedio diaria del Gran Buenos Aires en grados centígrados
ESTADO DEL TIEMPO	str	Estado del clima: Claro (C), Nublado (N) o Seminublado (SN)

## 1.2. Carga del Dataset

 <b>Excel Reader</b>	Se importó el dataset "Históricos valores de energía y potencia" desde un archivo Excel.
--	--

## 1.3. Preprocesamiento y limpieza de datos

 <b>Row Filter</b>	Elimina las dos primeras filas que contienen datos vacíos o irrelevantres.
 <b>Row to Column Names</b>	Convierte una fila del dataset en los nombres de las columnas.
 <b>Column Filter</b>	Elimina columnas que no aportan valor en esta etapa del análisis.
 <b>String to Date&amp;Time</b>	Convierte la columna “Fecha” de texto a formato de fecha local.
 <b>Double to Integer</b>	Transforma la columna “Potencia Pico SADI” de número decimal a número entero.
 <b>String to Number</b>	Convierte la columna “Temperatura Media Diaria GBA” de texto a número decimal.

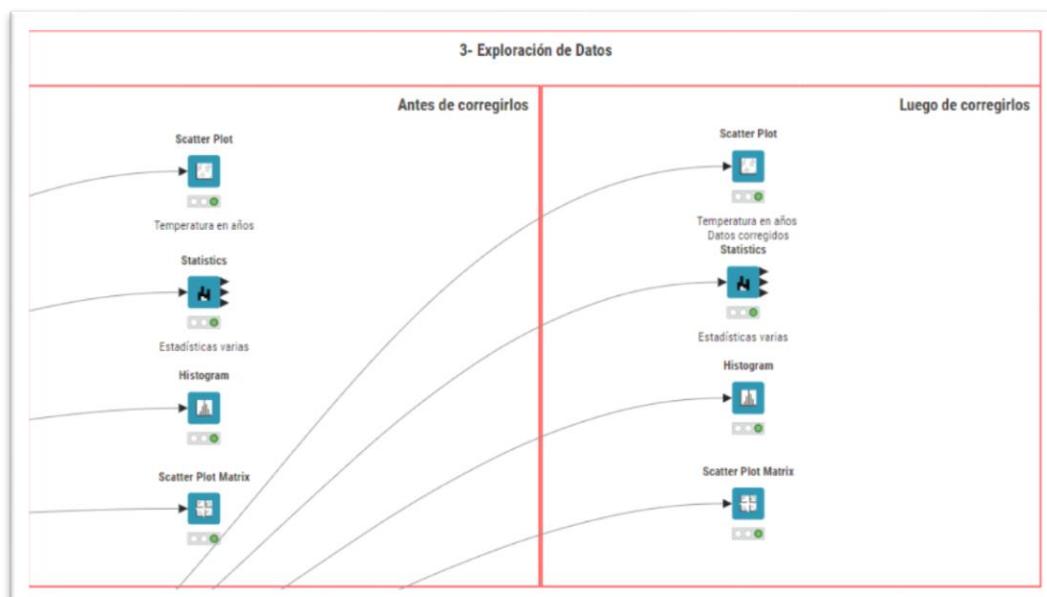
En el siguiente grafico se puede observar el inicio el workflow en KNIME, en este caso podemos ver la carga del dataset y cuáles fueron los primeros pasos para la limpieza de la base de datos, en concordancia con lo descripto anteriormente.



#### 1.4. Análisis exploratorio y visualizaciones

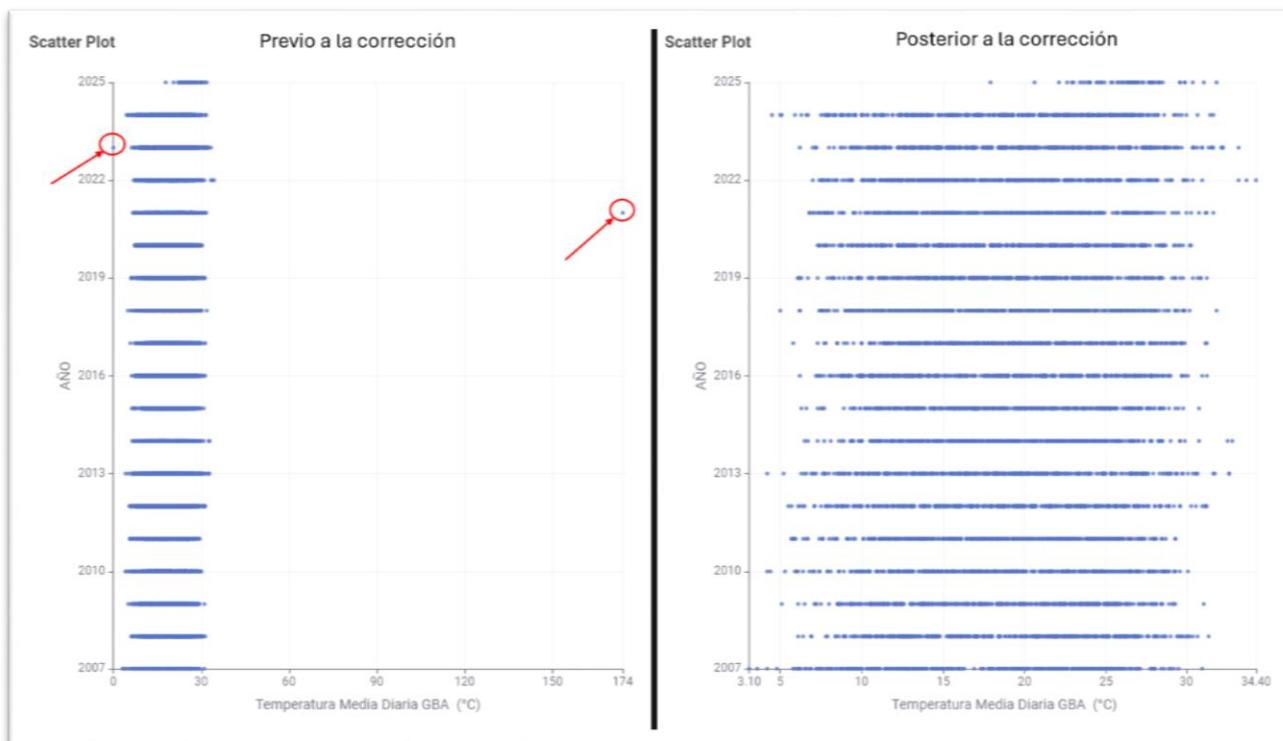
Scatter Plot	Visualiza la relación entre “Año” y “Temperatura Media Diaria”, detectando valores anómalos.
Statistics	Calcula estadísticas básicas (mínimo, máximo, media, desviación estándar, entre otras) de las variables.
Histogram	Muestra la distribución de la variable “Potencia Pico SADI”.
Scatter Plot Matrix	Explora las correlaciones entre variables, destacando la relación entre “Potencia Pico SADI” y “Energía SADI”.

En esta parte del workflow se llevó a cabo el análisis exploratorio de los datos, buscando identificar patrones y posibles relaciones entre variables. Los nodos utilizados se aplicaron tanto sobre los datos originales como sobre los datos corregidos, permitiendo observar cómo fueron modificándose a lo largo del proceso de limpieza.



A continuación, se detallan los gráficos obtenidos durante el análisis exploratorio, destacando los principales hallazgos visuales.

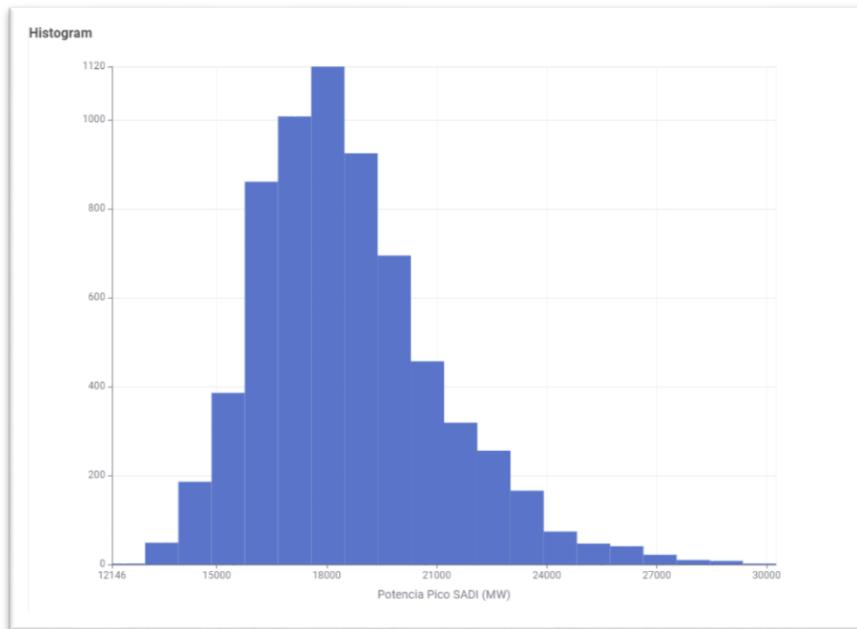
- Scatter Plot AÑO en función a la Temperatura Media Diaria GBA (°C)



Al graficar la temperatura a lo largo de los años, notamos ciertas anomalías en las mediciones del dataset: se observan temperaturas de 0°C (inconsistente con los días previos y posteriores a esa fecha) y de 174°C. Para resolver estos casos, se definió aplicar una técnica de imputación y corregir el dato con el valor promedio de la temperatura del día anterior y el día posterior. Los pasos realizados se detallan en el apartado "[Transformación de datos](#)".

- Histogram

El gráfico muestra que la potencia pico del sistema SADI se mantuvo, en la mayoría de los días, entre los 17.000 MW y 20.000 MW. Sin embargo, también se registraron valores significativamente más altos, superando los 28.000 MW. Si bien estos casos son poco frecuentes y podrían considerarse valores atípicos, no se eliminan del análisis ya que el objetivo del negocio es predecir el pico máximo de potencia y la energía diaria demandada. Por lo tanto, estos valores extremos aportan información relevante para el modelo.



- Estadísticas varias

Para tener una mejor comprensión y corroboración de los datos, se graficaron todas las variables en Statistics, donde se pueden visualizar las siguientes: Año, N° Mes, Semana, N° Dia, Potencia Pico SADI (MW), Energía SADI (GWh), hora potencia pico, temperatura media diaria GBA (°C).

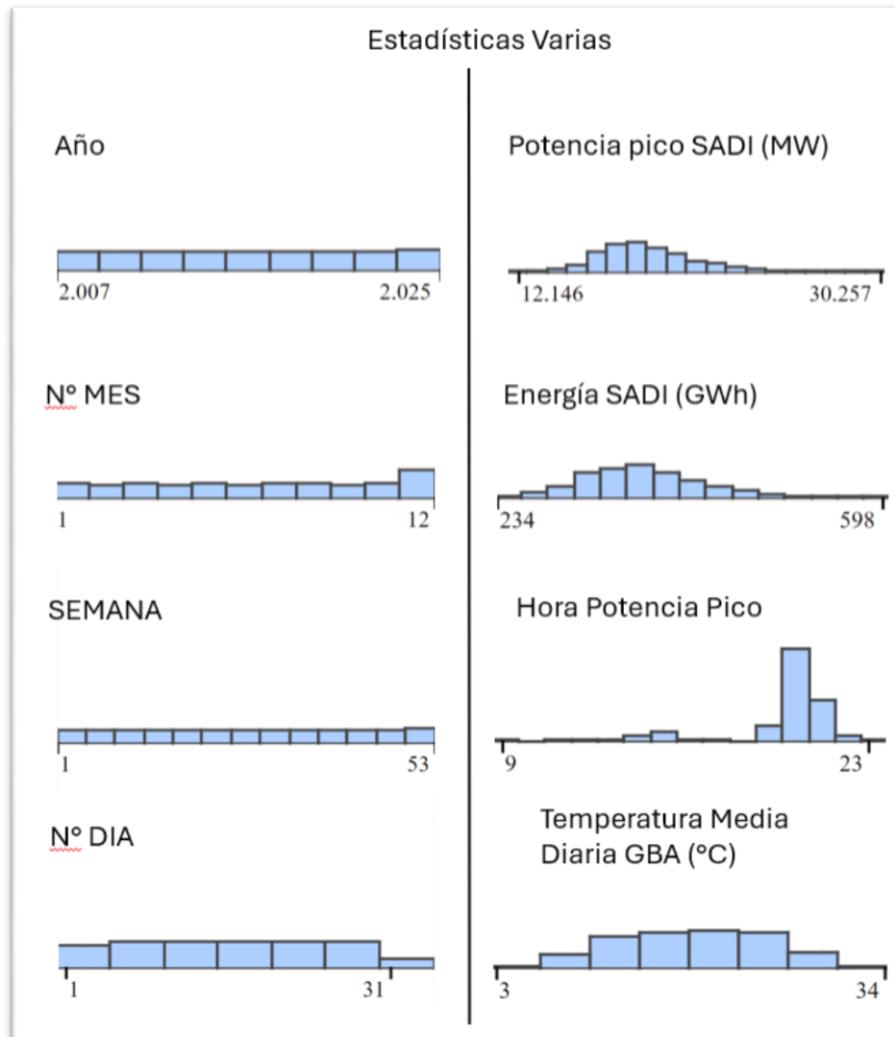
De ellas, podemos destacar:

- Existe cierta correlación entre las variables Energía SADI y Potencia Pico SADI (MW), pero no necesariamente representan lo mismo. Por ejemplo, el pico de potencia SADI máximo histórico sucedió el día 10/02/2025, mientras que la energía máxima histórica se registró el 01/02/2024.

#### Máximos Históricos de Energía y Potencia

Día	Hábil		Sábado		Domingo	
	POT MW	ENE GWh	POT MW	ENE GWh	POT MW	ENE GWh
Máxima	30257	597,7	27203	559,8	25739	543,6
Fecha	10/02/25	01/02/24	11/03/23	11/03/23	12/02/23	12/02/23
Hora	14:47	-	14:35	-	16:16	-
T° Med Bs.As.	31,1 °C	31,5 °C	32,2 °C	32,2 °C	33,3 °C	33,3 °C

- La hora potencia pico suele darse mayoritariamente entre las 19 y las 23 horas.
- La temperatura media diaria de GBA oscila entre valores de 3,1 °C y 34,4 °C.
- El récord histórico de potencia pico SADI es de 30257 MW y su valor mínimo es de 12146 MW.
- El récord histórico de energía SADI es de 597,664 GWh y su valor mínimo es de 233,679 GWh.

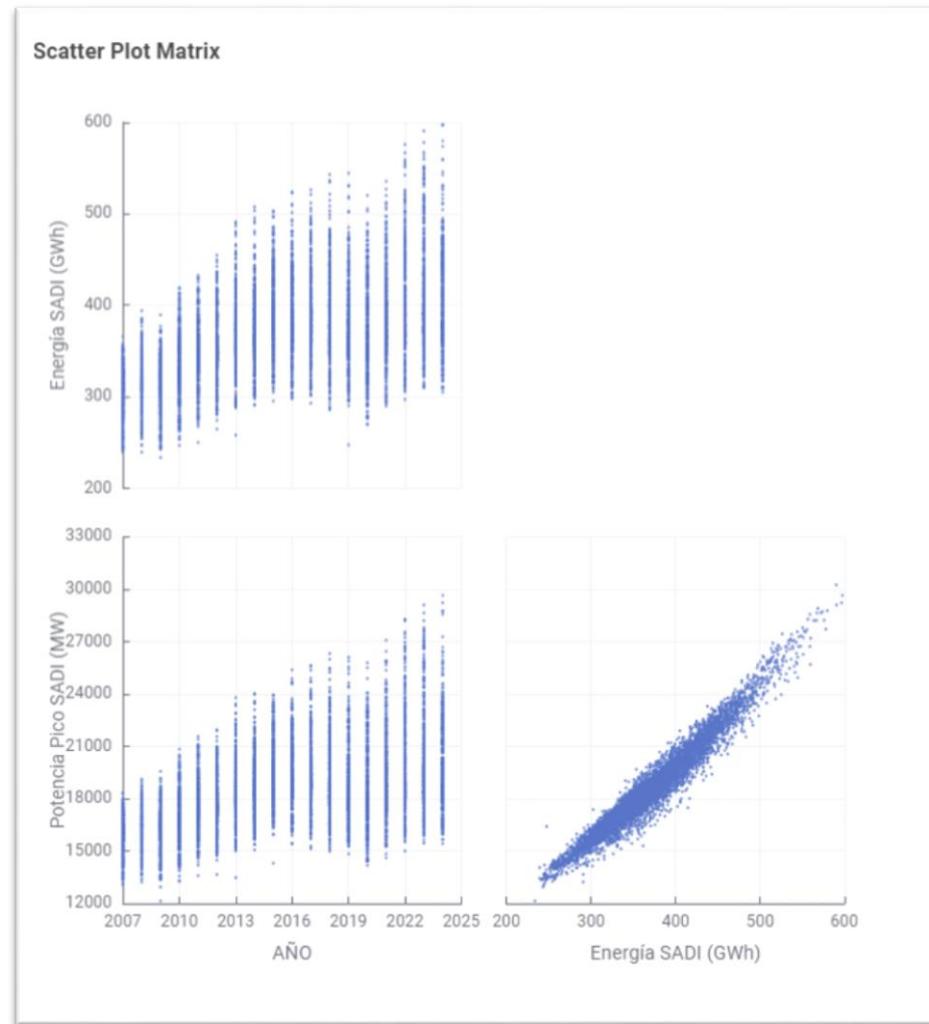


- Scatter Plot Matrix

Esta gráfica se utilizó para encontrar la relación entre las variables Potencia Pico SADI (MW) y Energía SADI (GWh), mencionada anteriormente.

Adicionalmente, podemos ver gráficamente el incremento anual de la demanda, debido a varios factores como el incremento poblacional, la industrialización, el uso de nuevas tecnologías y aparatos electrónicos, entre otros.

También podemos observar el impacto que generó la pandemia en la demanda, reduciendo significativamente este crecimiento.



## 1.5. Transformación de datos

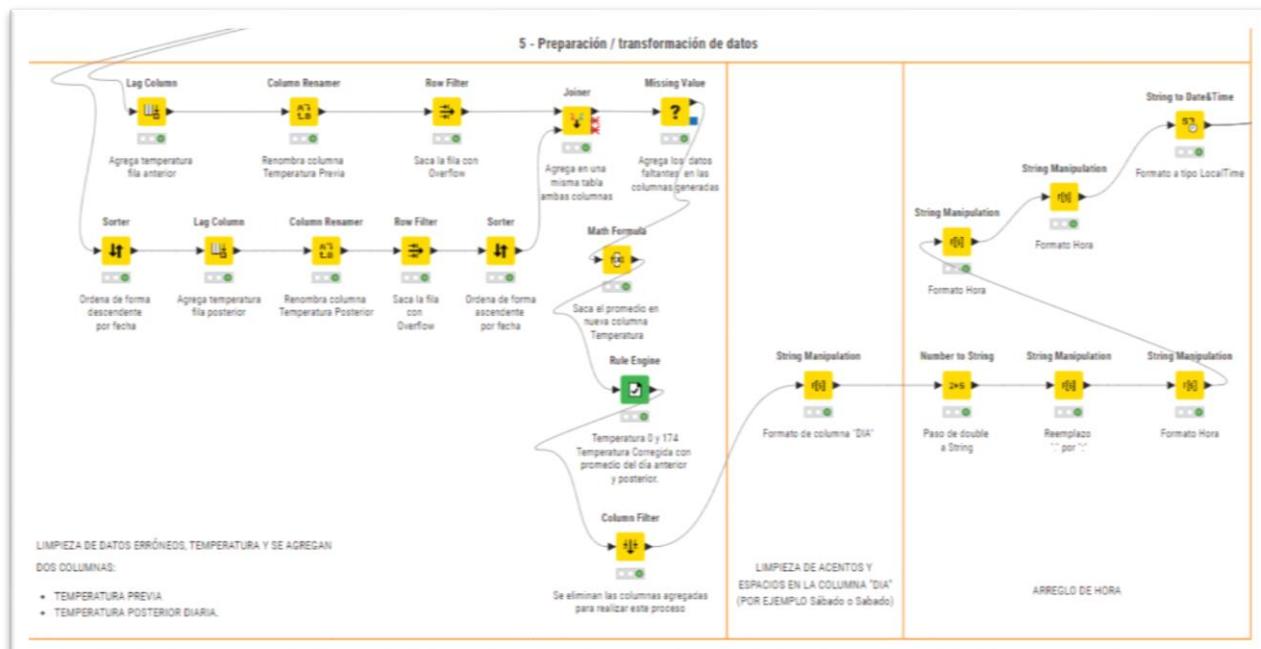
A continuación, se detallan los pasos aplicados para transformar los datos provistos por el dataset, de forma tal que sean más robustos para el análisis y asegurar una correcta interpretación. Se incluye la corrección de valores atípicos en la variable de temperatura, la limpieza y normalización de los datos del campo “Día”, y la conversión del formato de hora, que presentaba inconsistencias.

Lag Column ▶  ◀ ●	Se agrega una columna con la temperatura del día anterior.
Column Renamer ▶  ◀ ●	Se renombra la columna como “Temperatura Previa Media Diaria GBA”.
Row Filter ▶  ◀ ●	Se elimina el registro sin información que se generó en el paso anterior.
Sorter ▶  ◀ ●	Partiendo nuevamente de la tabla limpia, se ordena la misma por fecha en forma descendente.

Lag Column 	Se agrega una columna con la temperatura del día posterior.
Column Renamer 	Se renombra la columna como “Temperatura Posterior Media Diaria GBA”.
Row Filter 	Se elimina el registro sin información que se generó en el paso anterior.
Sorter 	Se ordena la tabla por fecha en forma descendente.
Joiner 	Unifica la tabla con los datos de “Temperatura Previa Media Diaria GBA” y “Temperatura Posterior Media Diaria GBA”.
Missing Value 	Agrega los datos faltantes en las columnas de “Temperatura Previa Media Diaria GBA” y “Temperatura Posterior Media Diaria GBA”.
Math Formula 	En una nueva columna “Promedio Temperatura Día anterior y posterior”, se calcula el promedio entre “Temperatura Previa Media Diaria GBA” y “Temperatura Posterior Media Diaria GBA”. Código: <code>(\$Temperatura Previa Media Diaria GBA\$ + \$Temperatura Posterior Media Diaria GBA\$ )/2</code>
Rule Engine 	Se corrigen los valores atípicos con el valor obtenido en el paso anterior. Código: <code>\$Temperatura Media Diaria GBA (°C)\$= 0 =&gt; \$Promedio Temperatura Dia anterior y posterior\$ \$Temperatura Media Diaria GBA (°C)\$= 174 =&gt; \$Promedio Temperatura Dia anterior y posterior\$ TRUE =&gt; \$Temperatura Media Diaria GBA (°C)\$</code>
Column Filter 	Para simplificar la tabla, se eliminan las columnas de “Temperatura Previa Media Diaria GBA”, “Temperatura Posterior Media Diaria GBA” y “Promedio Temperatura Día anterior y posterior”.
String Manipulation 	Para evitar problemas en el proceso o al hacer comparaciones, se aplicó una función que normaliza el texto, reemplazando los caracteres acentuados por el equivalente sin acento y elimina espacios al principio y final del texto. Código: <code>replaceChars(lowerCase(strip(\$DIA\$)), "áéíóúÁÉÍÓÚ", "aeiouAEIOU")</code>
Number to String 	Para manipular la columna, se convierte el formato de la columna “Hora Potencia Pico”, pasa de tipo double a String.

<p><b>String Manipulation</b></p> <pre> String Manipulation     ▶ [!]     ● ○○   </pre>	<p>Se aplicaron 4 nodos String Manipulation para ajustar la columna “Hora Potencia Pico” y dejar los valores normalizados.</p> <ul style="list-style-type: none"> <li>• Se reemplaza el punto por dos puntos (. → :). replace(\$Hora Potencia Pico\$, ".", ":")</li> <li>• Corrección de formatos irregulares: añadir ceros a los minutos cuando faltan.  <pre> regexReplace(   regexReplace(     regexReplace(\$Hora Potencia Pico\$, :(6 7 8 9)\$, :0\$1),     :(1 2 3 4 5)\$, :\$10   ),   :0\$, :00" ) </pre> </li> <li>• Corrección de formatos irregulares: asegurar dos dígitos en minutos. regexReplace(\$Hora Potencia Pico\$, :(\\d{2})\\d\$, :\$1")</li> <li>• Corrección de formatos irregulares: asegurar dos dígitos en horas. regexReplace(\$Hora Potencia Pico\$, ^(\\d):", "0\$1:")</li> </ul>
<p><b>String to Date&amp;Time</b></p> <pre> String to Date&amp;Time     ▶ [!]     ● ○○   </pre>	<p>Se convierte la columna “Hora Potencia Pico” de texto (String) a formato de hora local (LocalTime) utilizando el patrón HH:mm.</p>

La siguiente imagen ilustra el segmento del workflow en KNIME en el que se implementaron las transformaciones descriptas previamente, permitiendo observar cómo se integraron y encadenaron los distintos nodos.

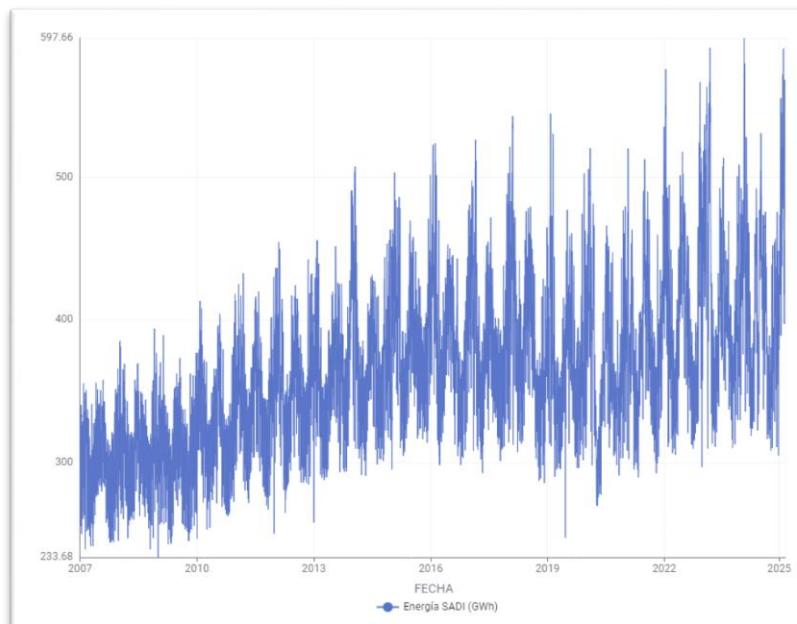


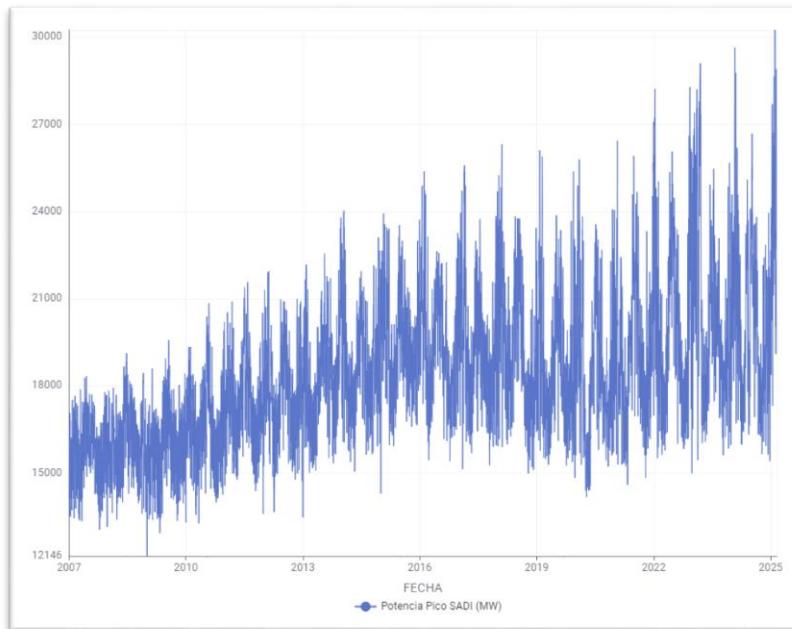
## 1.6. Técnicas gráficas utilizadas

<b>Metanode</b> 	Se utilizó para agrupar varios tipos de gráficos, los cuales se describen a continuación.
<b>Line Plot</b> 	Se visualiza las tendencias de las variables “Energía SADI (GWh)” y “Potencia Pico SADI (MW)” a lo largo del tiempo.
<b>Box Plot</b> 	Visualiza la distribución, asimetrías y outliers para detectar valores atípicos en las variables.
<b>Heatmap</b> 	Visualización de la correlación entre hora promedio de Potencia Pico y tipo de día.

- Line Plot:

Las curvas presentan similares tendencias acorde con su correlación, lo cual es coherente entre ambas variables. Estas graficas permiten identificar estacionalidades, años con mayor demanda y posibles anomalías en los datos.





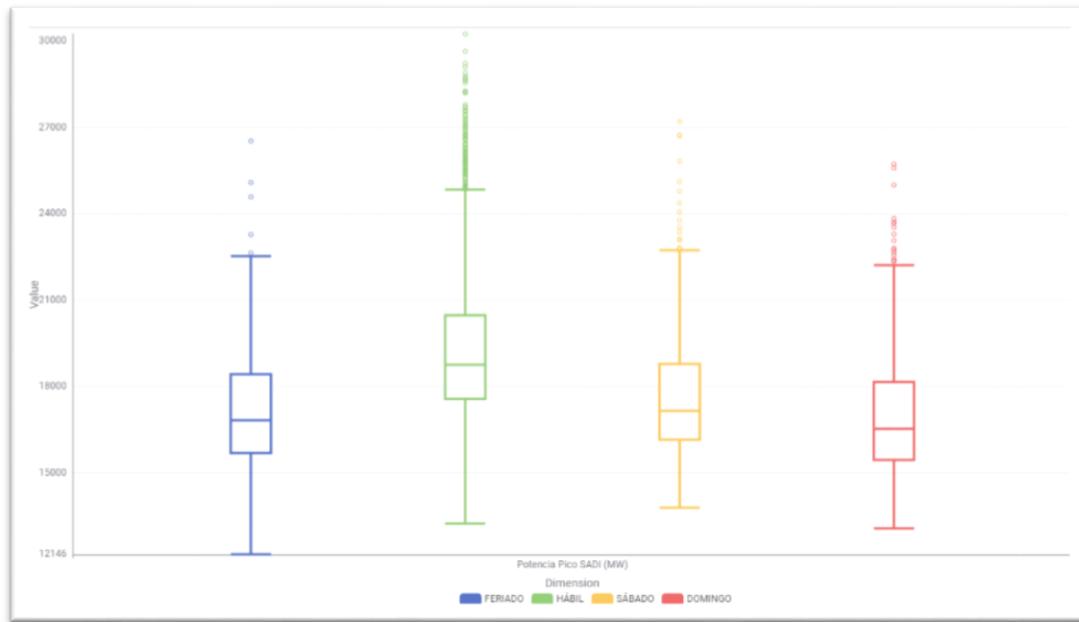
- Box Plot:

A través de este gráfico se pudo resumir visualmente como se distribuyen los datos. Nos muestra la mediana (línea dentro de la caja), cuartiles Q1 y Q3 en extremos de la caja, whiskers o bigotes (valores dentro de un rango razonable) y Outliers (valores por afuera de lo esperado).

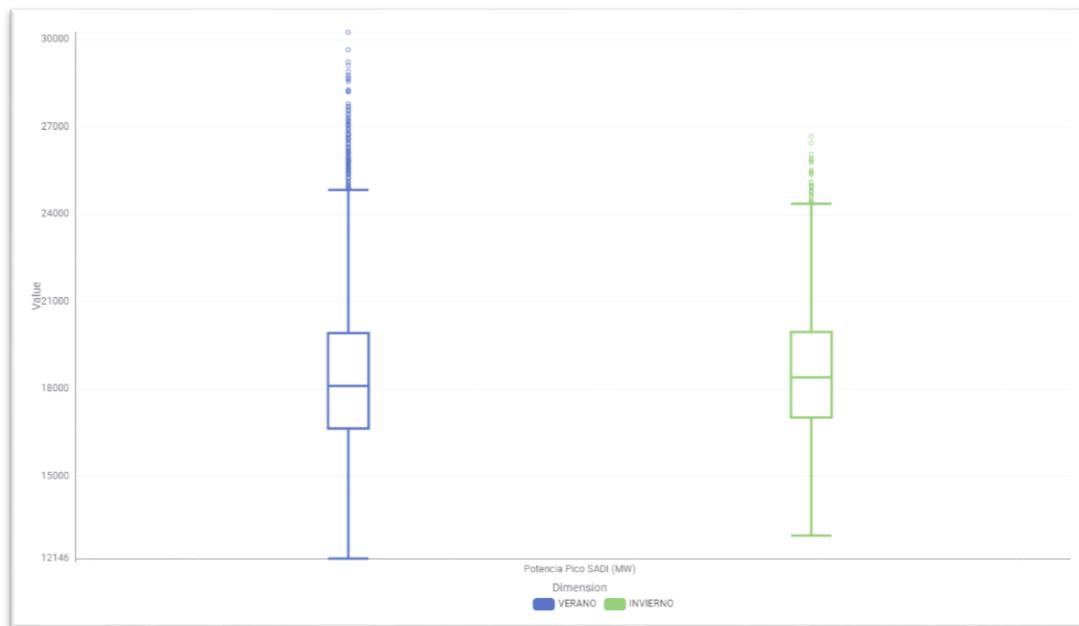
Considerando:

- Rango intercuartílico:  $IQR = Q3 - Q1$
- Límite inferior:  $Q1 - 1,5 * IQR$
- Límite superior:  $Q3 + 1,5 * IQR$

En el siguiente gráfico, se analizó la distribución de la variable “Potencia Pico SADI” según el tipo de día. Se observa un mayor consumo los días hábiles y menor consumo, los días feriados; los días sábados con menor dispersión, ya que se observan más concentrados; los días hábiles presentan mayor cantidad de outliers acorde a picos de demanda en jornadas de altas temperaturas.



Respecto a la “Potencia Pico SADI” según el clima (verano, invierno), se observa una mayor dispersión en verano, con mayor cantidad de outliers, en correspondencia a las últimas olas de calor. En invierno, la dispersión es menor, reflejando una demanda más estable.



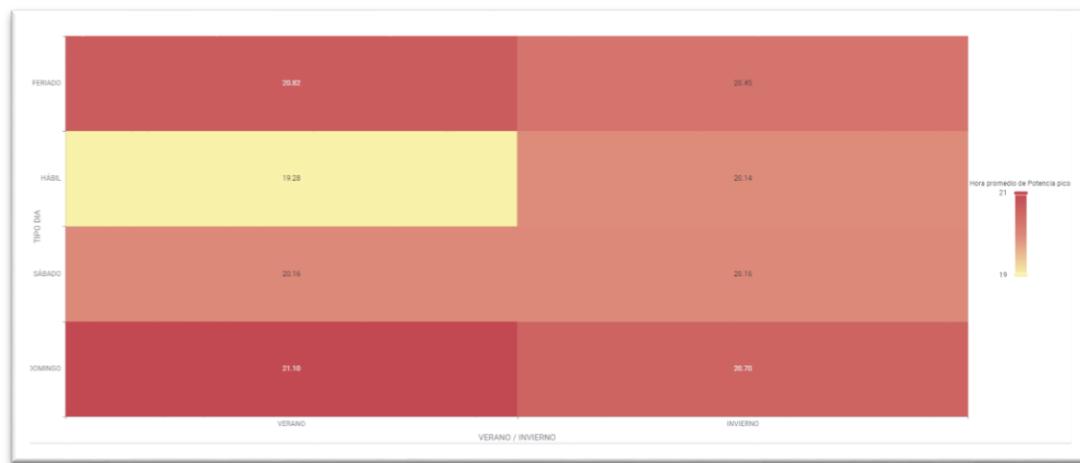
- Heatmap:

Mediante una matriz de colores se observa la correlación existente entre la hora promedio de “Potencia Pico SADI” y el tipo de día (Verano e Invierno), se aprecia que en invierno, la hora promedio de “Potencia Pico SADI” se da entre las 20 y 21 horas, mientras que en verano ocurre entre las 19 y 21 horas.

Los valores mínimos de la hora promedio en ambas estaciones se registran en los días hábiles, mientras que los máximos corresponden a los domingos.

Se aprecia que, en verano, los domingos y feriados presentan una hora promedio mayor que cualquier día de invierno.

*Cabe aclarar que este grafico se obtuvo antes de normalizar los valores de la columna “Hora Potencia Pico”.*



Las siguientes imágenes corresponden al metanodo que contiene las herramientas graficas descriptas anteriormente, empleadas para explorar visualmente el comportamiento de las variables del dataset. En una primera instancia se visualiza el metanodo cerrado, y al hacer click en el, se accede a su interior, donde pueden observarse los nodos específicos utilizados para las distintas visualizaciones.



## **PARTE 2 - MODELOS PREDICTIVOS Y TÉCNICAS DE EVALUACIÓN**

## 2.1. Regresión Lineal

El flujo de este modelo se dividió en tres partes, todas ellas comenzando a partir del nodo One to Many. Por un lado, para observar cómo se relacionan las variables independientes con la variable a predecir, se realizó un análisis utilizando el nodo Linear Regression Learner. Posteriormente, y en paralelo, se construyeron dos flujos: uno sin normalizar y el otro con los datos normalizados mediante el nodo Normalizer.

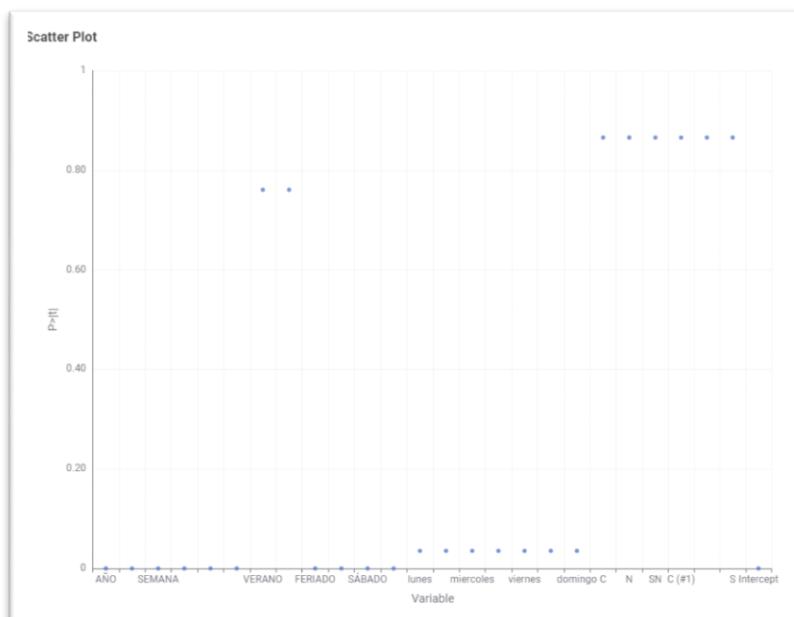
Inicio de los tres flujos:

 <b>One to Many</b>	Convierte las variables categóricas en variables binarias, permitiendo que el modelo de regresión lineal las interprete correctamente.
---	--

Flujo de relación entre variables independientes con la variable a predecir:

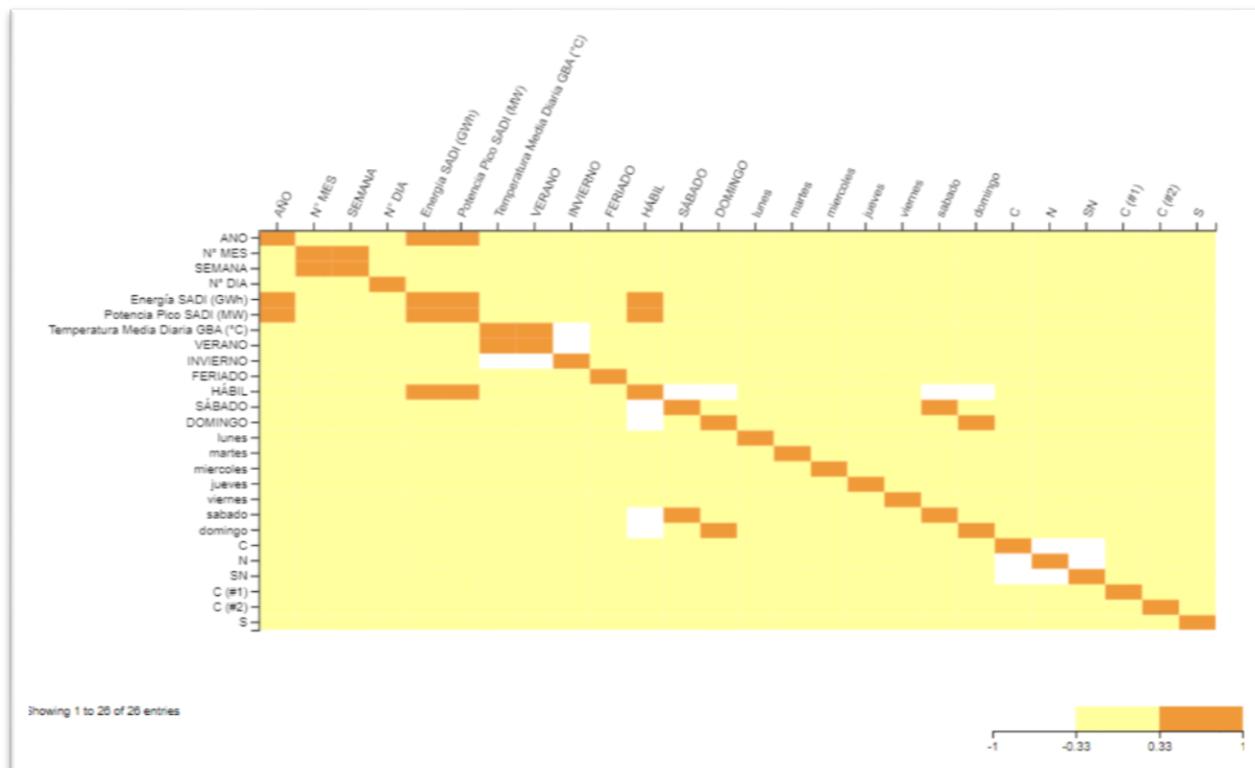
 <b>Linear Regression Learner</b>	Permite visualizar los coeficientes asociados a cada variable independiente, facilitando el análisis de su influencia sobre la variable objetivo.
 <b>Scatter Plot</b>	Permitió examinar el valor <i>p-value</i> asociado a cada variable, lo que posibilitó eliminar aquellas que no presentan una relación estadísticamente significativa con la variable dependiente. De esta forma, se depuró el conjunto de variables para optimizar la calidad del modelo.

En la siguiente imagen se observa gráficamente qué variables presentan una relación estadísticamente significativa con la variable Potencia Pico SADI, a partir del valor *p-value*. Aquellas cuyo *p-value* tiende a cero muestran una mayor influencia sobre la variable dependiente, mientras que las que se acercan a valores cercanos a 0.8 no presentan una relación significativa y pueden ser descartadas del modelo.



En esta imagen se representa gráficamente la matriz de correlación, cuyos valores oscilan entre 1 (correlación positiva perfecta), 0 (sin correlación lineal) y -1 (correlación negativa perfecta). Este método estadístico permite visualizar y comprender las relaciones entre las variables, lo cual resulta fundamental para detectar patrones, interpretar resultados y prevenir la multicolinealidad en modelos predictivos.

Con respecto a la variable dependiente (Potencia Pico SADI) vemos que Año, Energía SADI y Hábil son las variables con mayor correlación Positiva.



Aquí se Adjuntan los valores con respecto a la correlación de las variables independientes y la variable dependiente (Potencia Pico SADI).

<b>AÑO</b>	0.5347767647480463
<b>C</b>	-0.04797265351038727
<b>C (#1)</b>	-0.02007548538477315
<b>C (#2)</b>	-0.018647512784367257
<b>DOMINGO</b>	-0.2705491296244329
<b>Energía SADI (GWh)</b>	0.9726545525646464
<b>FERIADO</b>	-0.12502895855394383
<b>HÁBIL</b>	0.3767439320523418
<b>INVIERNO</b>	0.012363783391823209
<b>N</b>	0.029448583291241143
<b>N° DIA</b>	0.021921993215500254
<b>N° MES</b>	-0.09822286488196667
<b>Potencia Pico SADI (MW)</b>	1.0
<b>S</b>	0.028648448361817357

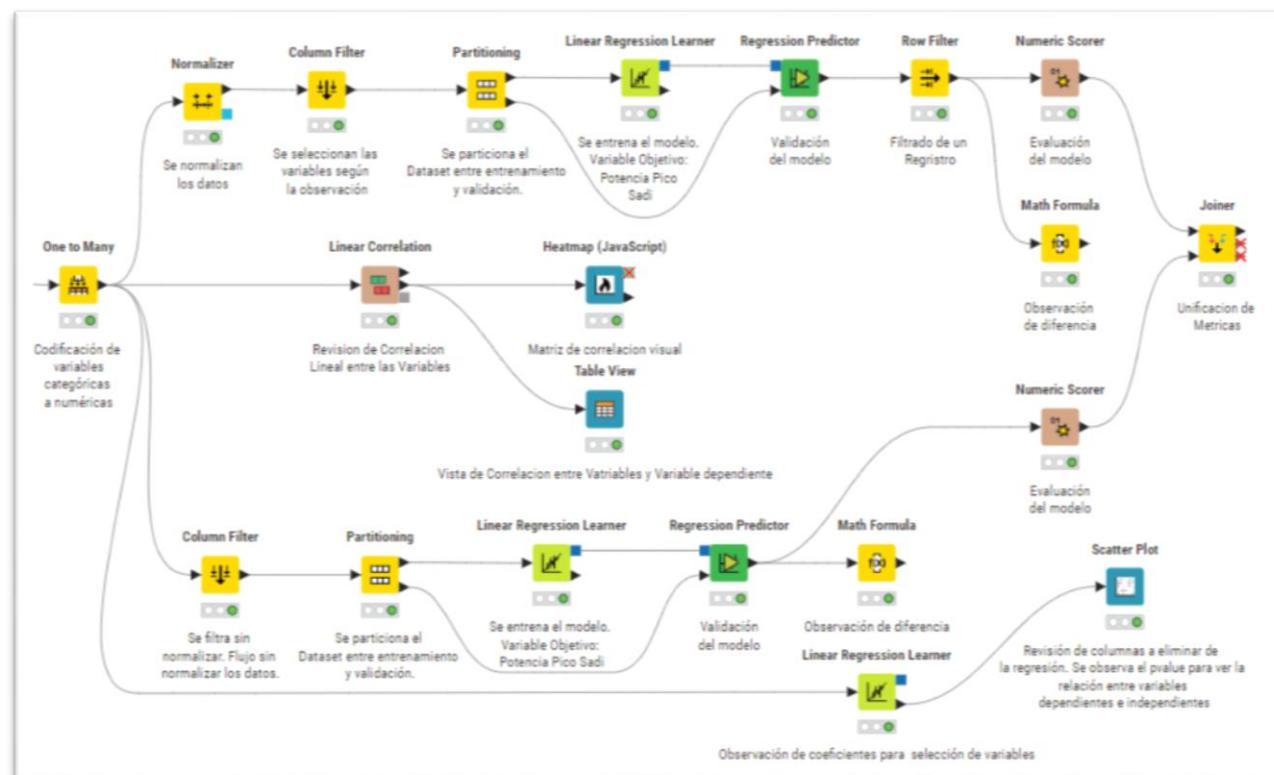
<b>SEMANA</b>	-0.09742539555027888
<b>SN</b>	0.01866361345864311
<b>SÁBADO</b>	-0.16309405418061315
<b>Temperatura Media Diaria GBA (°C)</b>	0.039366882501519575
<b>VERANO</b>	-0.012363783391823209
<b>domingo</b>	-0.2739831865846098
<b>jueves</b>	0.10569267376005113
<b>lunes</b>	0.06487401665584053
<b>martes</b>	0.10386617596848516
<b>miércoles</b>	0.1082770276621334
<b>sabado</b>	-0.16609664803582977
<b>viernes</b>	0.05717644614073911

Flujos de entrenamiento y predicción del modelo:

Normalizer 	Homogeneiza la escala de todas las variables numéricas, para evitar que las variables con rangos más altos dominen el modelo. Esta normalización se realizó en uno de los flujos, dejando el otro sin normalizar.
Column Filter 	Permite seleccionar únicamente las columnas relevantes para el análisis y el entrenamiento del modelo, eliminando aquellas que no aportan valor. Las variables incluidas en el modelo son: <ul style="list-style-type: none"> <li>AÑO - N° MES - SEMANA - FECHA - N° DÍA - Energía SADI (GWh) – Potencia Pico SADI (MW) - Temperatura Media Diaria GBA (°C) - Feriado - HÁBIL - SÁBADO - DOMINGO - lunes - martes - miércoles - jueves - viernes - sábado – domingo.</li> </ul>
Partitioning 	Divide la información en: 70% para entrenamiento y 30% para validación, asegurando así una adecuada evaluación del modelo.
Linear Regression Learner 	Entrenamiento del modelo de regresión lineal utilizando las variables seleccionadas previamente, permitiendo entender cómo influyen en la variable objetivo “Potencia Pico SADI (MW)”.
Regression Predictor 	Aplica el modelo entrenado sobre el conjunto de validación para obtener predicciones y poder evaluar su rendimiento.
Numeric Scorer 	Evalúa el desempeño del modelo utilizando métricas de error (como RMSE o MAE), permitiendo comparar los resultados entre diferentes enfoques (normalizado vs. no normalizado).
Math Formula 	Aplica una fórmula matemática para calcular el error relativo entre los valores reales y los valores predichos. Código: $(1 - (\$Potencia\ Pico\ SADI\ (MW)\$ / \$Prediction\$) ) * 100$

	Genera la correlación lineal entre las variables.
	Genera una tabla con la variable dependiente y las independientes para ver la correlación.
	Genera una matriz grafica para visualizar las relaciones de correlación.
	Filtrar un registro que obtuvo luego de la normalización un valor en 0.
	Se une y comparan las métricas obtenidas de ambos modelos.

A continuación se muestra el workflow con los nodos utilizados para entrenar y evaluar el modelo de Regresión Lineal, incluyendo la preparación de datos, entrenamiento y análisis de resultados.



### 2.1.1. Conclusiones preliminares

La comparación realizada entre los enfoques con y sin normalización permitió identificar diferencias significativas en el rendimiento predictivo.

En cuanto a los resultados obtenidos:

Modelo con normalización	Modelo sin normalización
File	File
R <sup>2</sup> : 0,962	R <sup>2</sup> : 0,964
Mean absolute error: 0,019	Mean absolute error: 346,952
Mean squared error: 0,001	Mean squared error: 224.422,319
Root mean squared error: 0,026	Root mean squared error: 473,732
Mean signed difference: -0,001	Mean signed difference: -25,069
Mean absolute percentage error: 0,063	Mean absolute percentage error: 0,019
Adjusted R <sup>2</sup> : 0,962	Adjusted R <sup>2</sup> : 0,964

Ambos modelos mostraron un alto poder explicativo, evidenciado por valores de R<sup>2</sup> superiores a 0.96, lo cual indica que son capaces de explicar gran parte de la variabilidad de la variable objetivo (Potencia Pico SADI).

Sin embargo, se observaron diferencias importantes en las métricas de error. El modelo con datos normalizados presentó errores absolutos considerablemente más bajos (MAE y RMSE), lo que sugiere una mejor capacidad de predicción en términos de magnitud.

Por otro lado, el modelo sin normalización mostró un MAPE inferior, lo que indica un mejor desempeño relativo en función del tamaño de los valores reales. Esto sugiere que, si bien el modelo con normalización fue más preciso en términos absolutos, el modelo sin normalización también logró buenos resultados cuando se evalúa el error de forma proporcional.

Por lo tanto podemos decir que la normalización de las variables ayudó a que la precisión absoluta mejorara el modelo de Regresión Lineal, no obstante ambos enfoques resultaron adecuados.

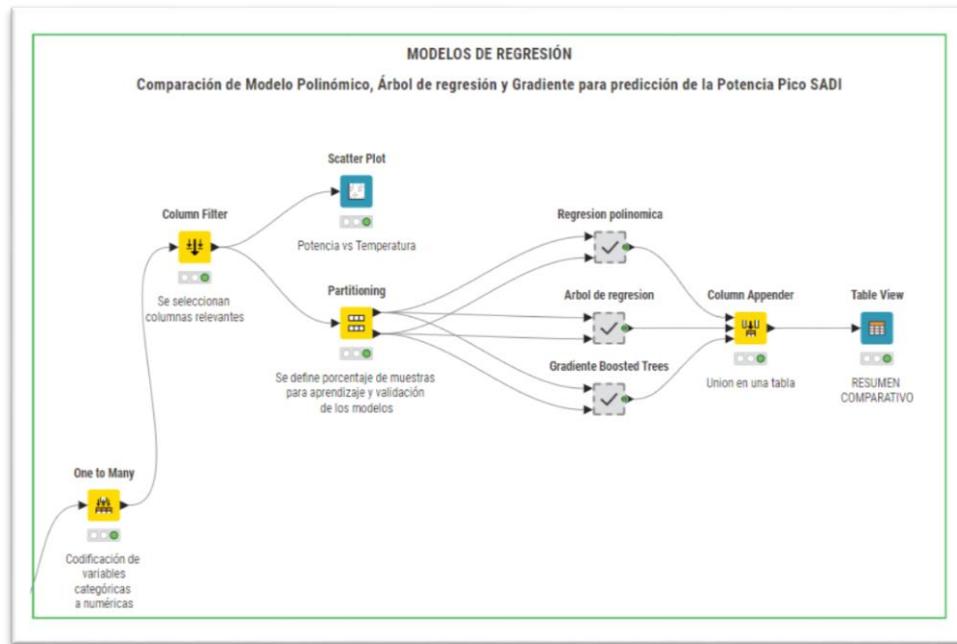
## 2.2. Flujo de Regresión con Estructura de Metanodos

Para que el flujo sea ordenado y visualmente comprensible, se han utilizado metanodos que desarrollan distintos métodos: Regresión Polinómica, Árbol de Regresión y Gradiente Boosted Trees. Posteriormente, estos análisis se integraron mediante el nodo Column Appender, lo que permitió visualizar un resumen comparativo utilizando el nodo Table view. A continuación, se describen los nodos utilizados.

 One to Many	Convierte las variables categóricas <i>Verano/Invierno</i> , <i>Tipo de Día</i> , <i>Día</i> y <i>Estado del Tiempo</i> en variables binarias (dummies), permitiendo que puedan ser utilizadas correctamente por los modelos de regresión (polinómica, árboles y Gradient Boosting)
--	---

	Permite seleccionar únicamente las columnas relevantes para el análisis y el entrenamiento del modelo. Las variables seleccionadas fueron: <ul style="list-style-type: none"> <li>• AÑO - Nº MES - Energía SADI (GWh) - Potencia Pico SADI (MW) - Temperatura Media Diaria GBA (°C) - VERANO - INVIERNO - FERIADO - HABIL - sábado - domingo.</li> </ul>
	Se utilizó para visualizar la relación entre la Potencia Pico SADI (MW) y la Temperatura Media Diaria GBA, permitiendo observar tendencias y patrones entre ambas variables.
	Divide el dataset en un 70% para entrenamiento y un 30% para validación, lo que permite evaluar el rendimiento del modelo de forma justa.
	Este metanodo implementa un modelo de regresión polinómica para predecir la Potencia Pico SADI. Permite modelar relaciones no lineales entre las variables independientes y la variable objetivo.
	Este metanodo construye un modelo de árbol de regresión para predecir la Potencia Pico SADI. El modelo divide los datos en distintos grupos según los valores de las variables, generando una estructura en forma de árbol. Esto permitirá descubrir relaciones no lineales y entender el resultado a través de reglas simples y fáciles de interpretar.
	Este metanodo aplica el algoritmo de Gradient Boosting para construir un modelo de regresión más preciso. El método combina varios árboles de decisión simples, que se van corrigiendo entre sí para reducir el error del modelo. Esto permite obtener predicciones más robustas y capturar relaciones complejas entre las variables. A diferencia de un único árbol, este enfoque mejora progresivamente el modelo a partir de los errores cometidos.
	Combina las predicciones generadas por los tres modelos de regresión (Polinómica, Árbol de Regresión y Gradient Boosted Trees) en una sola tabla. Esto permite comparar fácilmente los resultados de cada modelo para la misma instancia y analizar cuál ofrece mejores predicciones.
	Permite visualizar una tabla con las predicciones generadas por los distintos modelos de regresión utilizados. Facilitando la comparación de los resultados en paralelo.

A continuación, se presenta la estructura del workflow que implementa los modelos de Regresión Polinómica, Árbol de Regresión y Gradient Boosted Trees, organizados dentro de metanodos. Esta vista permite apreciar de manera global cómo se conectan los diferentes procesos y facilita la interpretación del esquema general antes descripto.



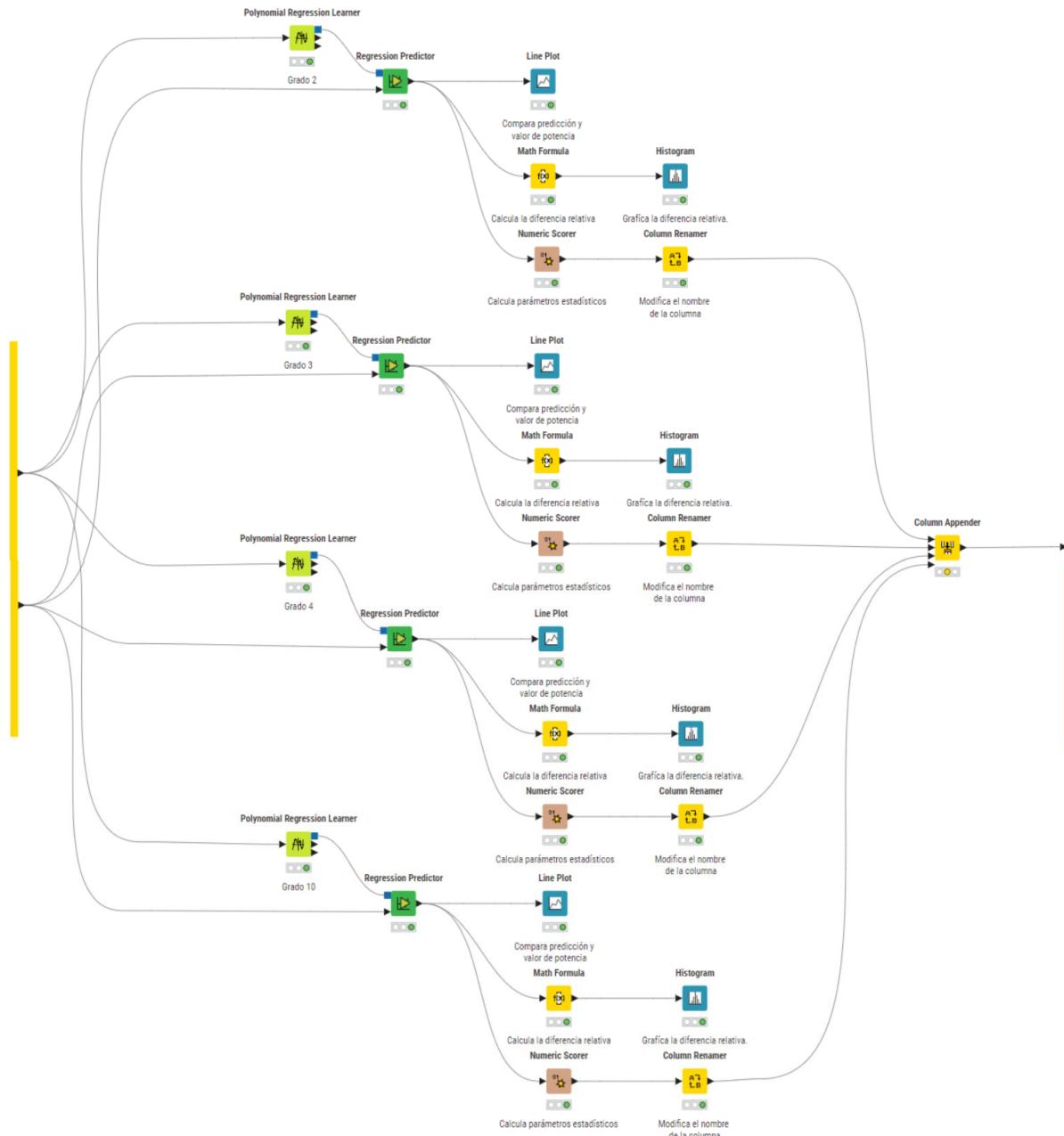
### 2.2.1. Regresión polinómica

Se utiliza el metanodo “Regresión polinómica” para efectuar la predicción mediante una regresión polinómica de grado 2, 3, 4 y 10, mediante los nodos de “Polynomial Regression Learner” y “Regression Predictor”.

<b>Polynomial Regression Learner</b> 	Entrena un modelo de regresión polinómica para predecir la Potencia Pico SADI, utilizando diferentes grados del polinomio: 2, 3, 4 y 10. En su configuración, se utilizó como variables independientes: <ul style="list-style-type: none"> <li>• AÑO - Nº MES - Energía SADI (GWh) - Temperatura Media Diaria GBA (°C).</li> </ul> Este modelo permite capturar relaciones no lineales entre las variables predictoras y la variable objetivo.
<b>Regression Predictor</b> 	Predice valores de potencia mediante la salida del nodo de aprendizaje y las muestras definidas para validación.
<b>Line Plot</b> 	Se grafican las curvas de potencia según el dataset inicial y los valores predichos según el modelo planteado.
<b>Math Formula</b> 	Calcula la diferencia relativa porcentual entre valor predicho de potencia y el valor real de las muestras. Calcula el error porcentual entre el valor predicho y el valor real de la Potencia Pico SADI. La expresión utilizada permite medir, en términos relativos, cuánto se desvía la predicción del modelo respecto al valor observado, facilitando la comparación del desempeño entre distintos modelos. Código:

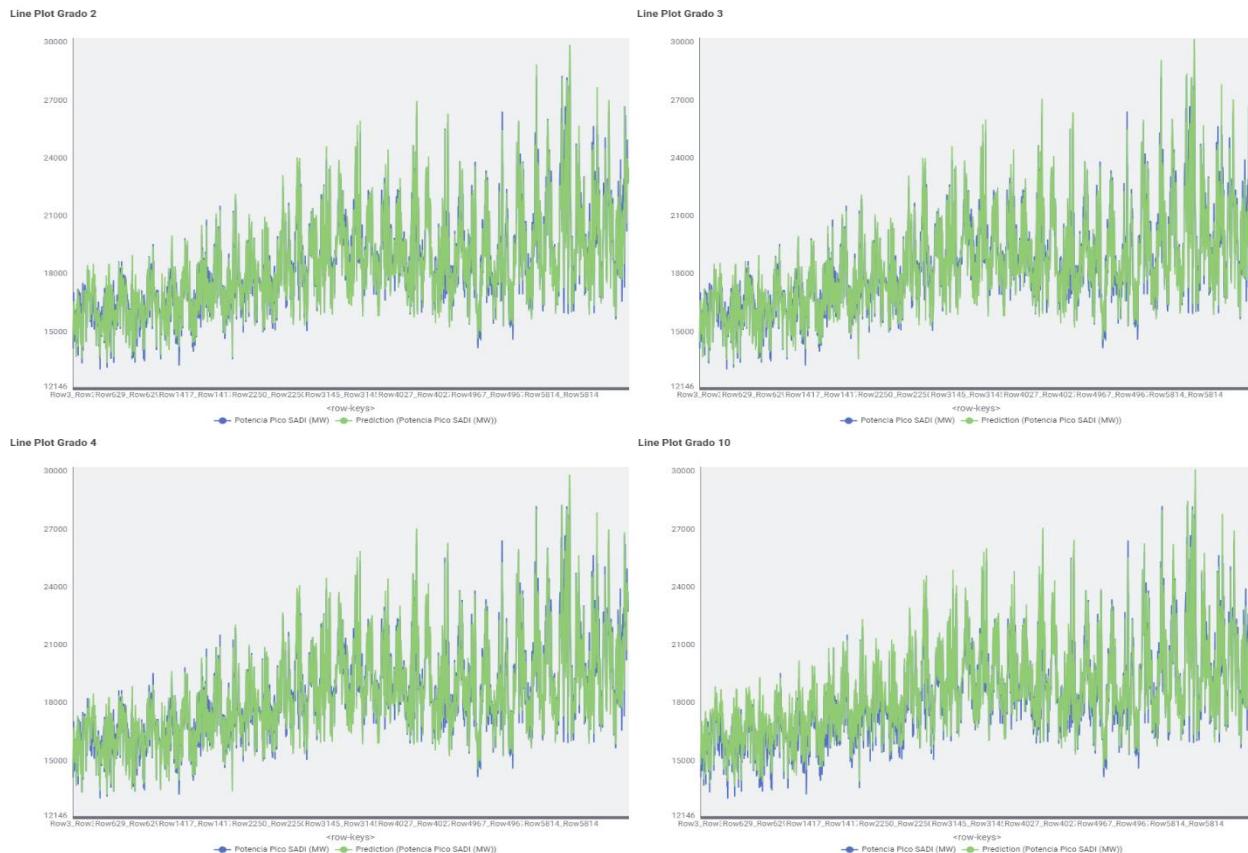
	<pre>(\$Prediction (Potencia Pico SADI (MW))\$ - \$Potencia Pico SADI (MW)\$ / \$Potencia Pico SADI (MW)\$ * 100)</pre>
Histogram 	Representa gráficamente la distribución de la variable <i>Diferencia %</i> , calculada como el error porcentual entre el valor predicho y el valor real de la Potencia Pico SADI.
Numeric Scorer 	Calcula parámetros estadísticos del modelo planteado.
Column Renamer 	Renombra la columna de predicción generada por el modelo, identificando a que regresión polinómica pertenece cada una. Se reemplaza el nombre genérico “Prediction (Potencia Pico SADI (MW))” por: <ul style="list-style-type: none"><li>• Regr_Pol_2º - Prediction (Potencia Pico SADI (MW))</li><li>• Regr_Pol_3º - Prediction (Potencia Pico SADI (MW))</li><li>• Regr_Pol_4º - Prediction (Potencia Pico SADI (MW))</li><li>• Regr_Pol_10º - Prediction (Potencia Pico SADI (MW))</li></ul> (Según corresponda al grado del modelo).
Column Appender 	Combina en una misma tabla las columnas de predicción generadas por los distintos modelos de regresión polinómica (grados 2, 3, 4 y 10). Permitiendo comparar de forma directa el desempeño de cada de ellos.

A continuación se muestra el workflow con los nodos utilizados para entrenar y evaluar los modelos de regresión polinómica de distintos grados (2º, 3º, 4º y 10º), incluyendo la preparación de datos, entrenamiento, renombrado de columnas y cálculo del error porcentual.



### 2.2.1.1. Conclusiones preliminares

Se comparan los valores predichos con los valores reales del dataset. La grafica que muestra el Line Plots, correspondiente a los diferentes grados de regresión polinómica (2º, 3º, 4º y 10º), presentan patrones similares, esto indica que el comportamiento general de la predicción no varía significativamente entre los distintos grados. A pesar de las diferencias en los valores calculados, las curvas muestran una tendencia común en la relación entre las variables independientes y la Potencia Pico SADI, lo que sugiere que, en este caso, la complejidad del modelo no tiene un impacto notable en la forma de la predicción.



Respecto a la comparación de la diferencia relativa de los modelos de regresión polinómica (2º, 3º, 4º y 10º) podemos indicar que se aprecian histogramas similares, aunque en los modelos de grados 4 y 10 se aprecia una reducción en los valores de la diferencia relativa. Es importante tener en cuenta que el modelo de grado 10 podría estar sufriendo sobreajuste (overfitting), ajustando sus oscilaciones de manera excesiva a los datos de entrenamiento, lo que podría comprometer su capacidad de generalización.

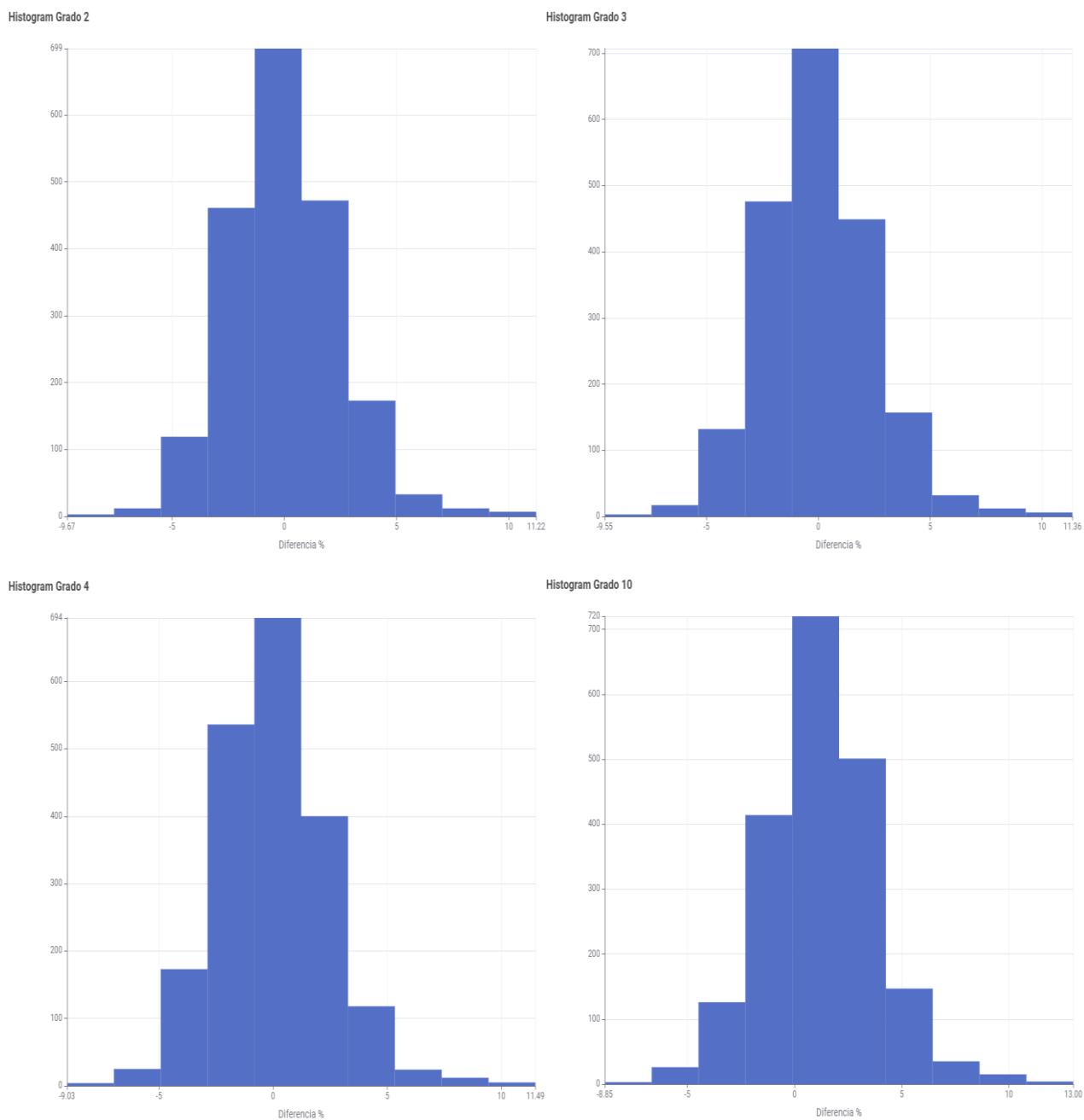
Las máximas diferencias relativas obtenidas no coinciden con los valores mínimo ni máximo de Potencia Pico consumida, sino a potencias medias, punto a destacar para evitar la saturación del sistema.

#	RowID	AÑO Number (inte...)	N° MES Number (inte...)	Energía S... Number (dou...)	Potencia ... Number (inte...)	Temperat... Number (dou...)	VERANO Number (inte...)	INVIERNO Number (inte...)	FERIADO Number (inte...)	HÁBIL Number (inte...)	sabado Number (inte...)	domingo Number (inte...)	Predictio... Number (dou...)	Difere... Number (dou...)
1551	Row51	2021	3	361.417	15298	22.5	1	0	0	0	0	1	17,902.741	17.027
#	RowID	AÑO Number (inte...)	N° MES Number (inte...)	Energía S... Number (dou...)	Potencia ... Number (inte...)	Temperat... Number (dou...)	VERANO Number (inte...)	INVIERNO Number (inte...)	FERIADO Number (inte...)	HÁBIL Number (inte...)	sabado Number (inte...)	domingo Number (inte...)	Predictio... Number (dou...)	Difere... Number (dou...)
991	Row33	2016	2	383.296	20400	26.8	1	0	0	0	0	1	18,825.436	-7.718

Los parámetros estadísticos generados por el nodo Numeric Scorer de los modelos de predicción se visualizarán más adelante mediante una tabla exportada desde cada uno de los metanodos, detallados en el apartado “[Comparación de modelos de predicción](#)”.

Los histogramas obtenidos para cada grado de polinomio presentan similares tendencias y varían en los valores de diferencia relativa máxima, destacando que el histograma de grado 10 presenta una frecuencia similar al de grado 3. De esta forma se observa como al aumentar la complejidad del modelo no se obtiene necesariamente una mayor exactitud, tendiendo incluso hacia una alta varianza y bajo sesgo.

El punto de equilibrio se definirá en función de los parámetros estadísticos, métrica, descartando el polinomio de grado 10 por sobreajuste.

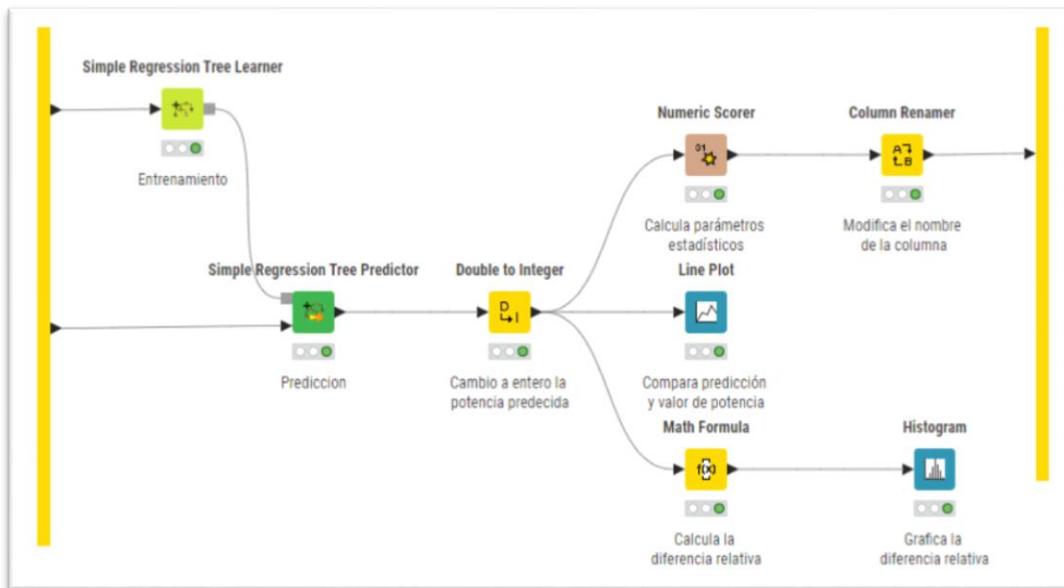


## 2.2.2. Árbol de regresión

De manera similar a lo realizado con la Regresión Polinómica, en este caso se entrena al modelo mediante el nodo “Simple Regression Tree Learner”, considerando el año, mes, energía, temperatura, verano-invierno, y tipo de día. Luego con el nodo “Simple Regression Tree Predictor”, se predicen los valores. A fin de que los valores predichos y la potencia presenten el mismo tipo de formato, se usa el nodo “Double to Integer”; se calcula la diferencia relativa y se exporta del metanodo para comparar con los otros modelos.

Simple Regression Tree Learner	Se utiliza para la etapa de aprendizaje del algoritmo. Su input es el 70 % del dataset inicial, definido en el nodo Partitioning.
Simple Regression Tree Predictor	Predice valores de Potencia Pico SADI mediante la salida del nodo de aprendizaje y las muestras definidas para validación.
Double to Integer	Convierte la columna de datos de tipo numeric double a integer para la potencia predicha por el algoritmo. En la configuración del nodo, se excluyen las columnas de Energía SADI y Temperatura Media Diaria GBA (°C), dejando solo la columna Prediction (potencia) para ser transformada a tipo integer.
Numeric Scorer	Calcula parámetros estadísticos del modelo planteado.
Line Plot	Se grafican las curvas de potencia según el dataset inicial y los valores predichos según el modelo planteado.
Math Formula	<p>Calcula la diferencia relativa porcentual entre valor predicho de potencia y el valor real de las muestras.</p> <p>Calcula el error porcentual entre el valor predicho y el valor real de la Potencia Pico SADI. La expresión utilizada permite medir, en términos relativos, cuánto se desvía la predicción del modelo respecto al valor observado, facilitando la comparación del desempeño entre distintos modelos.</p> <p>Código:</p> <pre>(\$Prediction (potencia)\$  - \$Potencia Pico SADI (MW)\$)  / \$Potencia Pico SADI (MW)\$  * 100</pre>
Column Renamer	Renombra la columna de predicción generada por el modelo, se reemplaza el nombre genérico “Prediction (potencia)” por “Arbol_Regresion - Prediction (potencia)”.
Histogram	Representa gráficamente la distribución de la variable <i>Diferencia %</i> , calculada como el error porcentual entre el valor predicho y el valor real de la Potencia Pico SADI.

A continuación, se muestra el workflow con los nodos utilizados para entrenar y evaluar el modelo Árbol de regresión, incluyendo la preparación de datos, entrenamiento, renombrado de columnas y cálculo del error porcentual.



### 2.2.2.1. Conclusiones preliminares

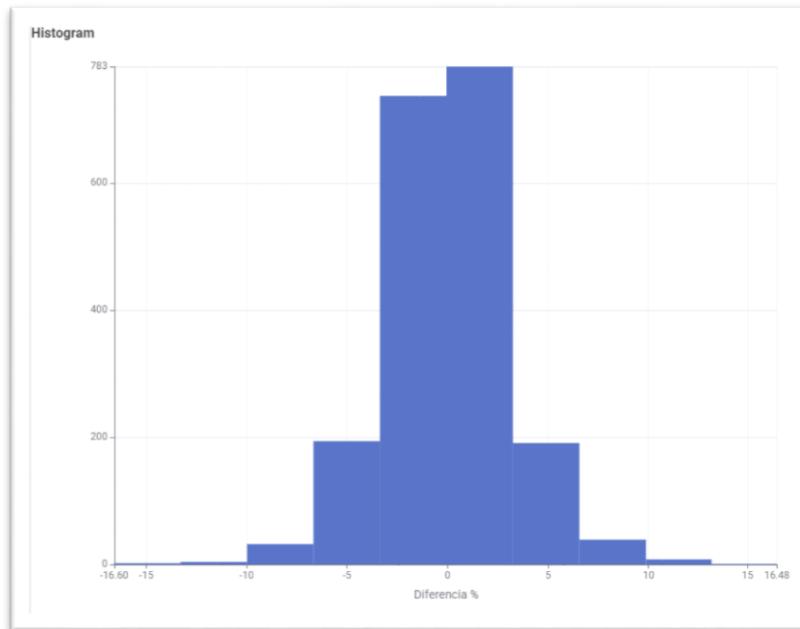
Este modelo presenta una mayor frecuencia de diferencia relativa en valores mayores al 7% y menores al -7% respecto a los alcanzados con regresión polinómica, y las máximas diferencias no corresponden con los valores mínimo ni máximo de potencia consumida. Además, a potencias máximas, la diferencia alcanzada es del orden del 2%.

#	RowID	AÑO	Nº MES	Energía S...	Potencia ...	Temperat...	VERANO	INVIERNO	FERIADO	HÁBIL	sabado	domingo	Predictio...	Diferencia %
		Number (inte...)	Number (inte...)	Number (dou...)	Number (inte...)	Number (dou...)	Number (inte...)	Number (double)						
1551	Row51	2021	3	361.417	15298	22.5	1	0	0	0	0	1	17645	15.342

#	RowID	AÑO	Nº MES	Energía S...	Potencia ...	Temperat...	VERANO	INVIERNO	FERIADO	HÁBIL	sabado	domingo	Predictio...	Diferencia %
		Number (inte...)	Number (inte...)	Number (dou...)	Number (inte...)	Number (dou...)	Number (inte...)	Number (double)						
110	Row36	2007	12	291.841	15956	27.8	1	0	0	0	0	0	13667	-14.346

#	RowID	AÑO	Nº MES	Energía S...	Potenc... ↓	Temperat... ↓	VERANO	INVIERNO	FERIADO	HÁBIL	sabado	domingo	Predictio...	Diferencia %
		Number (inte...)	Number (inte...)	Number (dou...)	Number (inte...)	Number (dou...)	Number (inte...)	Number (double)						
1863	Row62	2024	2	597.664	29653	31.5	1	0	0	1	0	0	29105	-1.848
1864	Row62	2024	2	596.615	29232	30.7	1	0	0	1	0	0	29105	-0.434
1760	Row59	2023	3	567.291	28643	31.4	1	0	0	1	0	0	28207	-1.522
1867	Row62	2024	2	558.99	28565	29.7	1	0	0	1	0	0	27796	-2.692

Dado el histograma resultante, se tiene una menor exactitud que las de regresión polinómica, no alcanzando un punto equilibrio entre sesgo y varianza con esta complejidad de modelo.



Por último, se considera el modelo de gradiente a fin de alcanzar una mejor métrica.

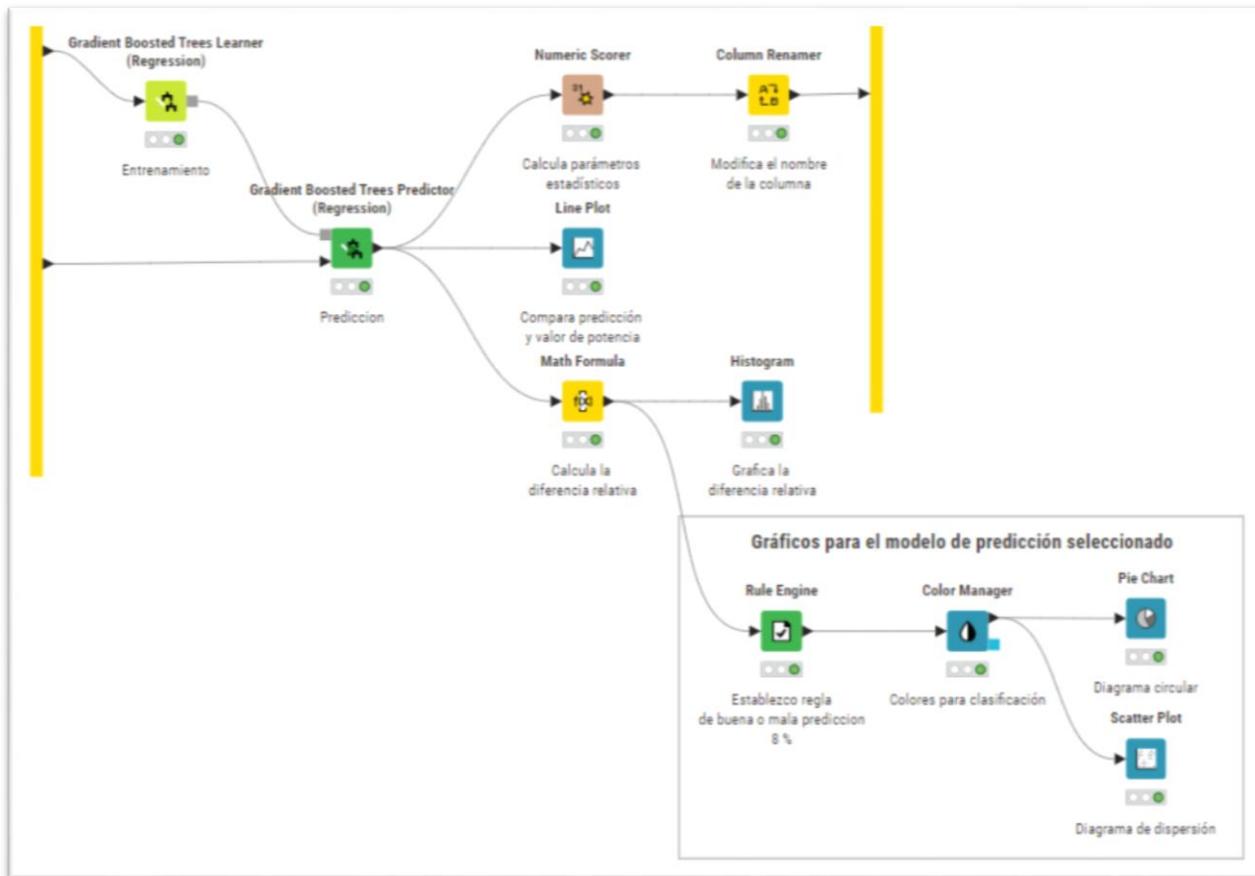
### 2.2.3. Gradiente Boosted Trees

De igual manera que los modelos ya descriptos, se entrena al modelo, en este caso con mediante el nodo “Gradient Boosted Trees Learner”. Luego, con el nodo “Gradient Boosted Trees Predictor”, se predicen los valores. Por último, se calcula la diferencia relativa y se exporta del metanodo.

Gradient Boosted Trees Learner (Regression) 	Se utiliza para la etapa de aprendizaje del algoritmo. Su input es el 70 % del dataset inicial, definido en el nodo Partitioning.
Gradient Boosted Trees Predictor (Regression) 	Predice valores de Potencia Pico SADI mediante la salida del nodo de aprendizaje y las muestras definidas para validación.
Double to Integer 	Convierte la columna de datos de tipo numeric double a integer para la potencia predicha por el algoritmo. En la configuración del nodo, se excluyen las columnas de Energía SADI y Temperatura Media Diaria GBA (°C), dejando solo la columna Prediction (potencia) para ser transformada a tipo integer.
Numeric Scorer 	Calcula parámetros estadísticos del modelo planteado.
Line Plot 	Se grafican las curvas de potencia según el dataset inicial y los valores predichos según el modelo planteado.
Math Formula 	Calcula la diferencia relativa porcentual entre valor predicho de potencia y el valor real de las muestras. Calcula el error porcentual entre el valor predicho y el valor real de la Potencia Pico SADI. La expresión utilizada permite medir, en términos

	<p>relativos, cuánto se desvía la predicción del modelo respecto al valor observado, facilitando la comparación del desempeño entre distintos modelos.</p> <p>Código:</p> <pre>(\$Prediction (Potencia Pico SADI (MW))\$ - \$Potencia Pico SADI (MW)\$) / \$Potencia Pico SADI (MW)\$ * 100</pre>
Column Renamer 	Renombra la columna de predicción generada por el modelo, se reemplaza el nombre genérico “Prediction (Potencia Pico SADI (MW))” por “Gradient Boosted Trees - Prediction (potencia)”.
Histogram 	Representa gráficamente la distribución de la variable <i>Diferencia %</i> , calculada como el error porcentual entre el valor predicho y el valor real de la Potencia Pico SADI.
Rule Engine 	Establece una regla para determinar una buena o mala predicción. Código: \$Diferencia %\$ <= 8 => "Buena Prediccion" \$Diferencia %\$ > 8 => "Mala Prediccion"
Color Manager 	Se asigna el color rojo para identificar una “Mala Predicción” y el color verde para una “Buena predicción” haciendo más fácil la visualización de los resultados.
Pie Chart 	Representa la Buena y Mala predicción en un gráfico de torta.
Scatter Plot 	Representa la Buena y Mala predicción en un diagrama de dispersión.

A continuación se muestra el workflow con los nodos utilizados para entrenar y evaluar el modelo Gradiente Boosted Trees, incluyendo la preparación de datos, entrenamiento, renombrado de columnas y cálculo del error porcentual.

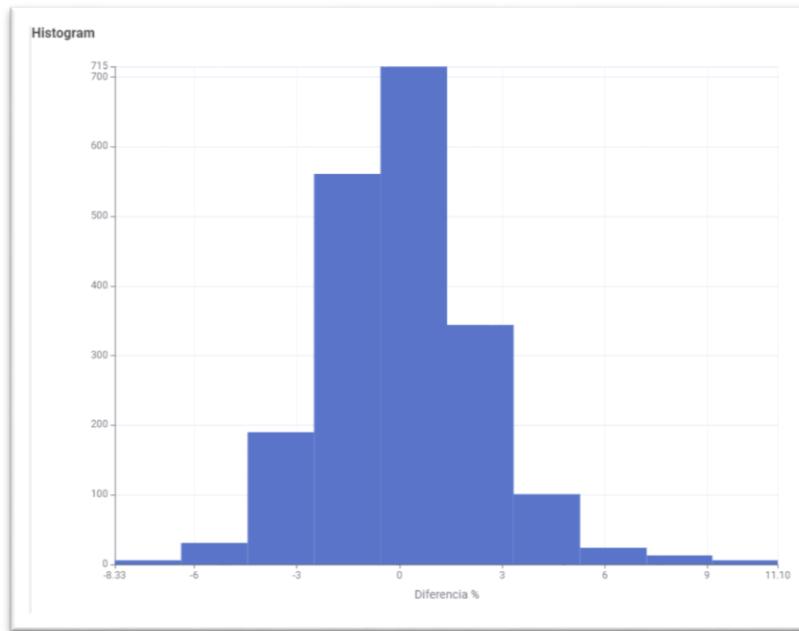


### 2.2.3.1. Conclusiones preliminares

Se alcanza una diferencia relativa similar al de regresión polinómica, pero más concentrado alrededor del  $\pm 5\%$ . Además, la máxima diferencia alcanzada del 17,97% coincide con valores bajos de potencia consumida, y para el caso de potencia máxima, de -6,91%. De esta forma, el caso de mayor diferencia relativa (mínima potencia pico) implicara un mayor costo de generación que lo necesario, destacándose que no peligra la saturación del sistema con un menor error.

A continuación, se presentan las métricas obtenidas y el histograma correspondiente.

#	RowID	AÑO Number (integer)	Nº MES Number (integer)	Energía SADI Number (double)	Potencia Pico ... Number (integer)	Temperatura ... Number (double)	VERANO Number (integer)	INVIERNO Number (integer)	FERIADO Number (integer)	HÁBIL Number (integer)	sábado Number (integer)	domingo Number (integer)	Prediction (Po... Number (double)	Diferencia % ↓ Number (double)
1551	Row51	2021	3	361.417	15298	22.5	1	0	0	0	0	1	18,046.527	17.967
1114	Row37	2017	2	430.612	22346	28.4	1	0	1	0	0	0	20,646.83	-7.604
1863	Row62	2024	2	597.664	29653	31.5	1	0	0	1	0	0	27,602.571	-6.915
1864	Row62	2024	2	596.615	29232	30.7	1	0	0	1	0	0	27,602.571	-5.574

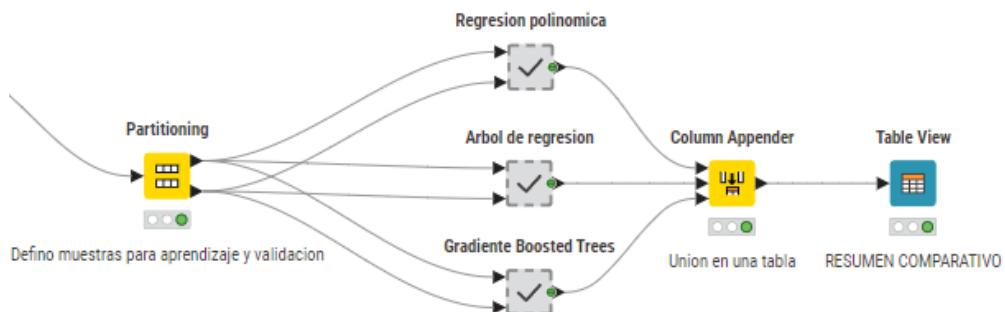


A partir de los análisis realizados, se concluye que los modelos con mejor desempeño son el modelo de gradiente y el modelo de regresión polinómico de grado 4.

La elección final del modelo más eficiente se determinara en función de los indicadores estadísticos obtenidos.

### 2.3. Comparación de modelos de predicción

Los valores estadísticos de cada modelo planteado, obtenidos con el nodo Numeric Scorer, se exportan del metanodo y se unen a una tabla comparativa.



RowID ↓	Regr_Pol_2º - Prediction (Number (double))	Regr_Pol_3º - Prediction (Number (double))	Regr_Pol_4º - Prediction (Number (double))	Regr_Pol_10º - Prediction (Number (double))	Arbol_Regresion - Prediction (Number (double))	Gradient Boosted Trees - Prediction (potencia) (Number (double))
root mean squared error	456.041	456.027	447.354	479.901	580.799	436.369
mean squared error	207,973.385	207,960.628	200,125.983	230,304.571	337,327.187	190,417.643
mean signed difference	-4.165	-3.974	-1.377	165.481	-5.518	-17.428
mean absolute percentage error	0.019	0.019	0.018	0.02	0.023	0.017
mean absolute error	348.354	347.435	341.595	369.311	423.233	322.752
adjusted R²	0.966	0.966	0.967	0.963	0.945	0.969
R²	0.966	0.966	0.967	0.963	0.945	0.969

Analizando las métricas, se concluye que la mejor opción es la del modelo de gradiente, dado que presenta el mayor valor de  $R^2$  (varianza de los datos que explica el modelo), y menores valores de MAE (promedio de los errores absolutos), MSE, RMSE y MAPE (error promedio porcentual). Además, se observa la semejanza con el modelo de regresión polinómica grado 4 que se analizó gráficamente, y cómo el aumento de la complejidad del modelo no conlleva a una notable mejora en la métrica.

Si bien el modelo de gradiente presenta una mayor diferencia relativa respecto al modelo polinómico, la misma corresponde a un caso especial dentro de las bajas potencias pico y, en base a las muestras analizadas, la misma logra concentrar la diferencia relativa en torno al  $\pm 5\%$ , alcanzando un error razonable para evitar costos innecesarios en generación. De esta manera se alcanza un punto de equilibrio técnico y económico entre generación y demanda, teniendo la posibilidad de informar a las generadoras de que ajusten su despacho para un determinado día.



El modelo seleccionado estima la potencia pico máxima con un error de  $-6.91\%$  (mayor al obtenido con el modelo de árbol de regresión). Está cubierta por las reservas operativas de generación e importación. Teniendo en cuenta que CAMMESA monitorea la demanda en tiempo real, la estimación de consumo para un determinado momento anticipa que se deberá actuar de manera preventiva, abasteciendo la demanda con reservas de distinto tiempo de respuesta y, de ser necesario, importar energía de interconexiones con otros países o recurrir al mercado mayorista de electricidad para comprarla a generadores privados.

Por último, al variar la semilla de selección de muestras en el nodo Partitioning, se aprecia leves variaciones en las métricas de evaluación, manteniéndose la misma tendencia y modelo seleccionado.

RowID ↓	Regr_Pol_2º - Prediction (...	Regr_Pol_3º- Prediction (...	Regr_Pol_4º - Prediction (...	Regr_Pol_10º - Prediction (...	Arbol_Regresion - Prediction (...	Gradient Boosted Trees - Prediction (potencia)
root mean squared error	469.265	468.692	461.096	616.37	591.953	449.411
mean squared error	220,209.844	219,672.379	212,609.711	379,911.863	350,408.923	201,970.009
mean signed difference	-3.418	-3.367	-5.856	358.154	-26.551	-28.513
mean absolute percentage error	0.019	0.019	0.019	0.028	0.023	0.018
mean absolute error	354.914	354.228	348.295	498.691	431.861	333.729
adjusted R²	0.965	0.965	0.966	0.94	0.944	0.968
R²	0.965	0.965	0.966	0.94	0.944	0.968

## 2.4. Clasificación Multiclasificación: Comparación Naive Bayes - Random Forest

En esta clasificación multiclasificación se realizó una comparación entre los algoritmos de Naive Bayes y Random Forest. La clasificación utilizada fue en tres niveles: baja, media y alta demanda, con el objetivo de predecir la demanda de Potencia Pico SADI.

Este enfoque permite obtener una predicción útil sobre cómo será la demanda energética en una fecha determinada, lo que representa una ventaja competitiva y económica, además de contribuir a la seguridad del sistema mediante la optimización de recursos y la reducción de costos operativos.

Las categorías de demanda se definen según posibles acciones operativas asociadas:

- Baja: permite realizar mantenimiento preventivo y reducir la generación innecesaria.
- Media: corresponde a una operación estándar con monitoreo normal.
- Alta: requiere activación de reservas y control reforzado del sistema.

Este modelo de clasificación también presenta la ventaja de ser más comprensible y fácil de validar por stakeholders no técnicos.

*Por ejemplo: "Para mañana, el modelo predice demanda Alta por ser un día hábil con 33°C de temperatura."*

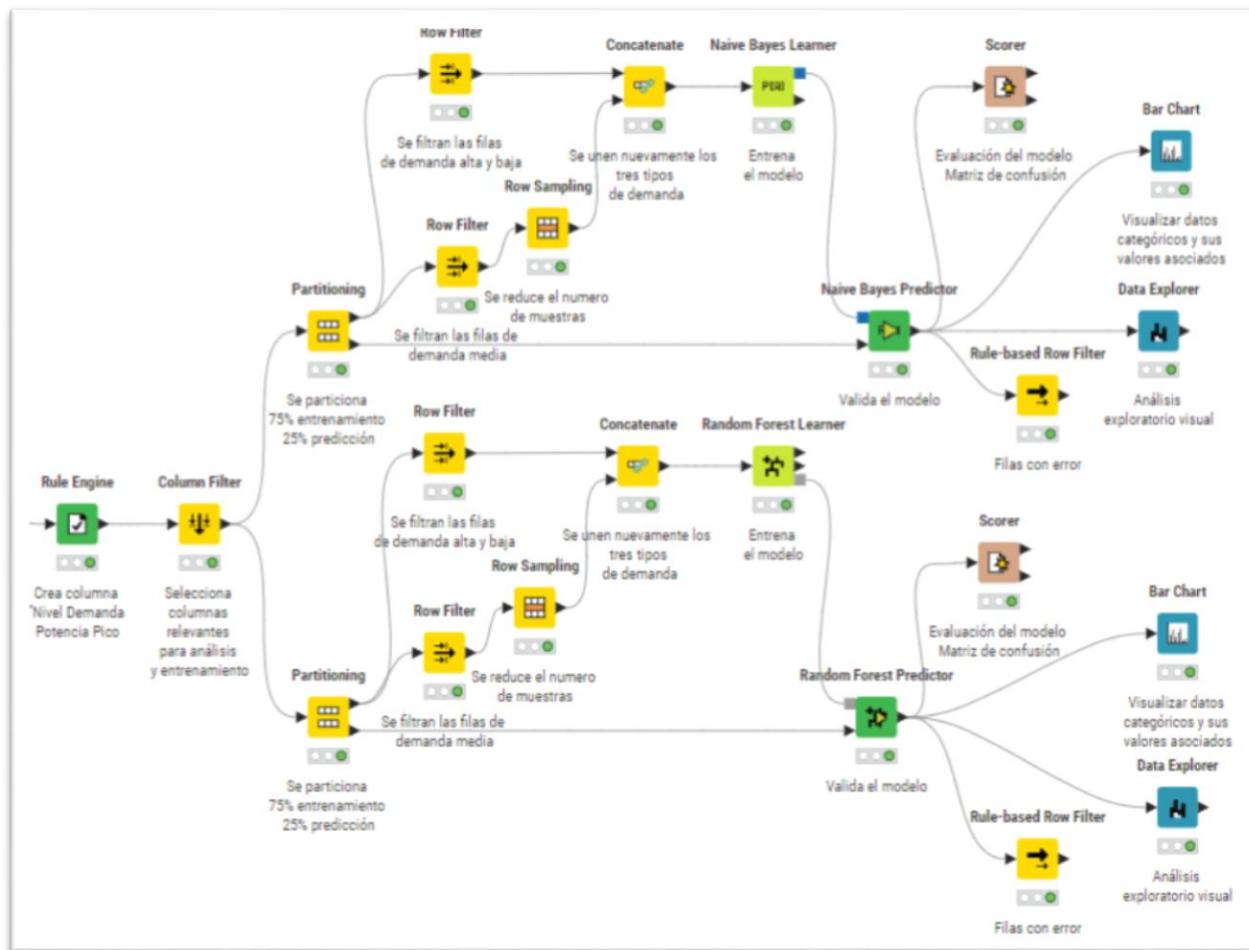
A continuación, se presenta el flujo de trabajo implementado para cada modelo, ambos modelos compartieron la misma estructura de flujos y nodos, con excepción del nodo de entrenamiento (Learner) y de predicción (Predictor), que fueron específicos para cada algoritmo.

 <b>Rule Engine</b>	<p>Crea una nueva columna categórica con el nombre “Nivel Demanda Potencia Pico”, se clasifica la Potencia Pico SADI en los siguientes límites: baja, si es menor o igual a 17000 MW; media, si es menor o igual a 20000 MW; alta, para los casos restantes.</p> <p>Código:</p> <pre>\$Potencia Pico SADI (MW)\$ &lt;= 17000 =&gt; "Baja" \$Potencia Pico SADI (MW)\$ &lt;= 20000 =&gt; "Media" TRUE =&gt; "Alta"</pre>
---	---

	Permite seleccionar únicamente las columnas relevantes para el análisis y el entrenamiento del modelo, eliminando aquellas que no aportan valor. Las variables incluidas en el modelo son: <ul style="list-style-type: none"> <li>• AÑO - VERANO/INVIERNO - SEMANA - TIPO DÍA - Temperatura Media Diaria GBA (°C) - Nivel Demanda Potencia Pico</li> </ul>
	Divide la información en: 75% para entrenamiento y 25% para validación, asegurando así una adecuada evaluación del modelo.
	Se filtran las filas de la columna "Potencia Pico SADI" conservando únicamente aquellas que no sean iguales a la demanda media, obteniéndose dos categorías: Baja (1369; 53.17%) y Alta (1206; 46.83%).
	Se filtran las filas de la columna "Potencia Pico SADI" conservando únicamente aquellas que sean iguales a la demanda media, obteniéndose la categoría Media (2400; 100.0%).
	Reduce la muestra de la demanda media a 1400 instancias, un número equiparable al de las categorías de Baja y Alta, con el fin de evitar distorsiones en el balance de las clases.
	Se concatenan nuevamente las tres categorías de la demanda (Baja, Media y Alta) en un único conjunto de datos.
	Entrena un modelo de clasificación utilizando el teorema de Bayes, bajo el supuesto de independencia entre las variables predictoras.
	Aplica el modelo entrenado para predecir la categoría de demanda en nuevos registros, asignando probabilísticamente cada observación a una de las clases definidas.
	Entrena el modelo de clasificación basado en un conjunto de árboles de decisión, donde cada árbol se construye a partir de una muestra aleatoria de los datos.
	Utiliza el modelo entrenado para predecir la clase de nuevos registros. La predicción final se determina por mayoría de votos entre los árboles individuales del bosque.
	Evalúa el modelo a través de la matriz de confusión, donde se pueden observar a simple vista los aciertos y fallos que tuvo el modelo para cada categoría de demanda.
	Visualiza la distribución de las predicciones, permitiendo ver cuántas veces el modelo clasificó la demanda como Baja, Media o Alta.
	Permite analizar y explorar estadísticamente el conjunto de datos.

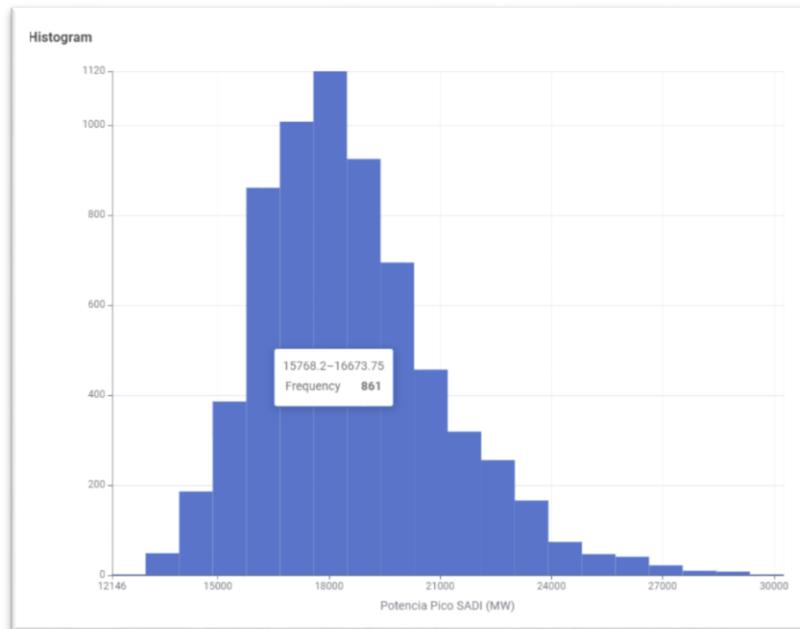
 <b>Rule-based Row Filter</b>	<p>Identifica la fila donde la clase real difiere de la clase predicha, a partir de la siguiente regla:</p> <p>Código:</p> <pre>\$Nivel Demanda Potencia Pico\$ MATCHES \$pred_Nivel Demanda Potencia Pico\$ =&gt; FALSE TRUE =&gt; TRUE</pre>
---	--

A continuación se muestra el workflow con los nodos utilizados para entrenar y evaluar los modelos de Clasificación (Naive Bayes - Random Forest), incluyendo la preparación de datos, entrenamiento y análisis de resultados.



### Criterios de selección de los límites para las clasificaciones

Los umbrales de clasificación en demanda Alta, Media y Baja utilizados en el nodo Rule Engine, se definieron a partir del análisis visual de la distribución de los datos mediante el siguiente histograma generado en la exploración de los datos.



Esta visualización permitió observar concentraciones naturales en los valores, facilitando la elección de los puntos de corte significativos. Como resultado, las clases quedaron distribuidas de la siguiente manera:

- Media (3200; 48.24%),
- Baja (1826; 27.52%),
- Alta (1608; 24.24%).

#### Criterios de selección de variables para la clasificación

Luego de realizar múltiples pruebas y evaluaciones con distintos conjuntos de datos, se concluyó que ciertas columnas debían ser incluidas o excluidas en los modelos de clasificación. Esta selección se basó en su relevancia predictiva, la redundancia entre variables y los resultados obtenidos en las métricas de desempeño.

- Variables excluidas:

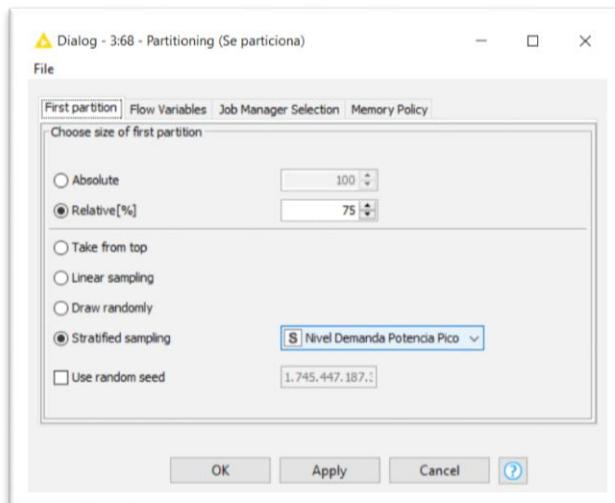
VARIABLE	JUSTIFICACIÓN
MES, FECHA, DIA, NRO DIA	Información redundante puede introducir ruido al modelo.
ENERGÍA SADI (GWH)	Está correlacionada con la Potencia, puede generar <i>leakage</i> (fuga de información) durante el entrenamiento.
POTENCIA PICO SADI (MW)	Es la variable base a partir de la cual se construyó la clase objetivo. No debe formar parte del conjunto de entrenamiento.
ESTADO DEL TIEMPO	No se encontró una relación clara entre el estado del cielo (claro, nublado o seminublado) y la demanda de potencia. Su inclusión generaba ruido.
HORA POTENCIA PICO	Conocer la hora exacta de la potencia pico implicaría ya conocer el resultado, lo cual invalida el modelo en un contexto de predicción real.

- Variables incluidas:

VARIABLE	JUSTIFICACIÓN
AÑO	Se observa una tendencia creciente de la demanda a lo largo de los años. Aporta información temporal útil al modelo.
VERANO / INVIERNO	Las estaciones extremas suelen presentar mayor demanda energética. Esta variable ayuda a capturar esa estacionalidad.
SEMANA	Su inclusión mejora el rendimiento del modelo. Aporta contexto temporal intermedio.
TIPO DIA	Permite distinguir entre días hábiles, feriados, sábados y domingos; los cuales presentan patrones distintos de consumo.
TEMPERATURA	Es una variable disponible con antelación y relevante, ya que la temperatura tiene un impacto directo con la demanda energética.
MEDIA DIARIA	
NIVEL DEMANDA	
POTENCIA PICO	Variable objetivo utilizada para entrenar el modelo de clasificación.

### Criterios de selección del porcentaje de partición

En ambos modelos (Naive Bayes y Random Forest) se optó por una partición inicial del conjunto de datos utilizando un 75% para el entrenamiento y un 25% para la validación del modelo.



La distribución de las clases resultantes fue:

	MEDIA	BAJA	ALTA
ENTRENAMIENTO	2400; 48.24%	1369; 27.52%	1206; 24.24%
VALIDACIÓN	800; 48.22%	457; 27.55%	402; 24.23%

Inicialmente, en ambos modelos, se utilizó el nodo SMOTE para balancear las clases, generando clases sintéticas de las clases minoritarias. Sin embargo, se observó que esta técnica introducía ruido a los datos, ya que se generaba información artificial que no representaba adecuadamente el comportamiento real. Dado el carácter real de los datos trabajados se decidió no utilizar esta técnica.

Por tal motivo se aplicó el nodo Rule Sampling para realizar una reducción de la clase mayoritaria “Media”. El conjunto final quedo equilibrado de la siguiente manera:

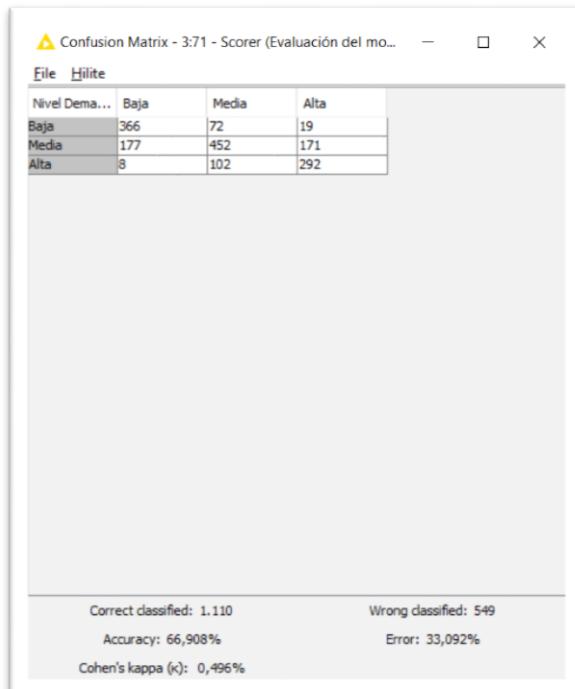
	MEDIA	BAJA	ALTA
ENTRENAMIENTO	1.400; 35.22%	1.369; 27.52%	1.206; 24.24%
VALIDACIÓN	800; 48.22%	457; 27.55%	402; 24.23%

#### 2.4.1. Conclusiones preliminares

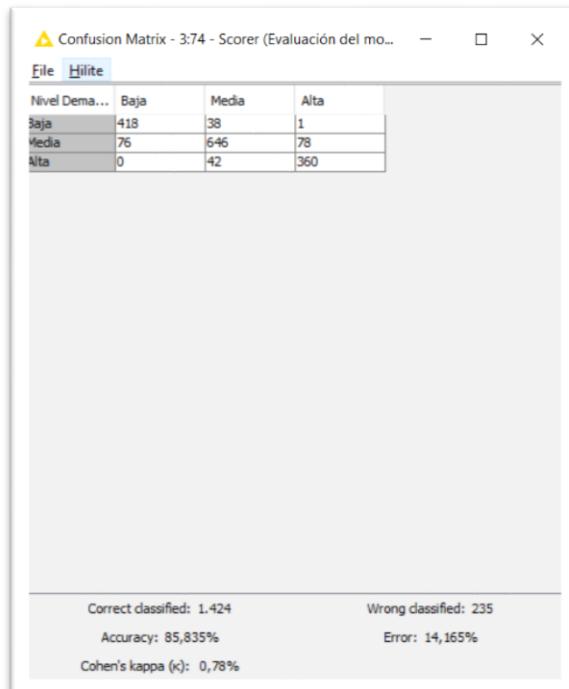
Al comparar las métricas obtenidas, se observa un mejor rendimiento del algoritmo Random Forest, tanto en términos de precisión general como en su capacidad de generalización frente a nuevos datos. Esta conclusión se fundamenta en los valores que se detallan a continuación:

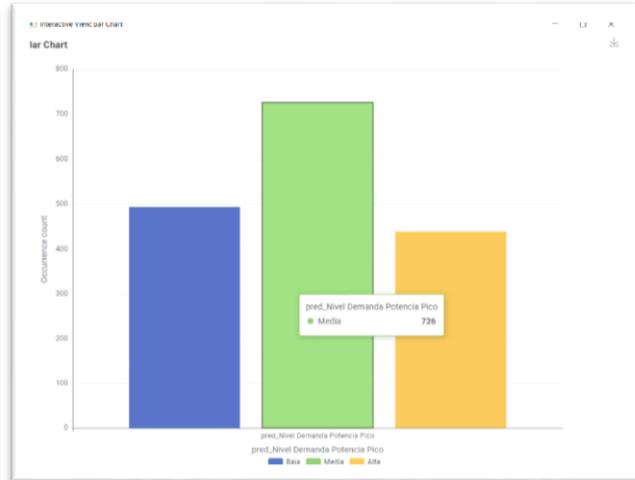
	NAIVE BAYES	RANDOM FOREST
CLASIFICACIÓN CORRECTA	1.110	1.424
CLASIFICACIÓN INCORRECTA	549	235
EXACTITUD	66,908%	85,835%
COEFICIENTE DE COHEN'S KAPPA	0,496%	0,78%
PORCENTAJE DE ERROR	33,092%	14,165%

Matriz de Confusión Naives Bayes



Matriz de Confusión Random Forest



**Bar Chart Naives Bayes****Bar Chart Random Forest**

Las predicciones de cada modelo tambien se realizó mediante el nodo Bar Chart mostrando las distribuciones de las clases predichas.

Observando estos gráficos y la matriz de confusión podemos inferir, en el caso del modelo Naives Bayes, que se sobreestima la clase Media y se subestima la clase Alta. Lo cual puede indicar una menor capacidad del modelo para discriminar correctamente entre clases cuando los valores son más cercanos.

Respecto al modelo de Random Forest, muestra una distribución de predicciones más cercanas al equilibrio real observado en el conjunto de validación. Lo cual demuestra que el modelo tiene una mejor capacidad del modelo para capturar la variable objetivo y mantener una representación más proporcional entre las clases.

Estos resultados refuerzan la elección del modelo Random Forest como el modelo más adecuado para la clasificación del nivel de demanda.

## 2.5. Conclusiones Modelos Predictivos

### Comparación entre los modelos de regresión desarrollados

La Regresión Lineal mostró un alto poder explicativo ( $R^2 > 0,96$ ) para un modelo con datos normalizados y para otro modelo sin normalizar. La normalización mejoró significativamente los errores absolutos (MAE - RMSE), mientras que en el modelo sin normalización se obtuvo un MAPE inferior. Esto nos sugiere que, si bien la normalización optimiza la precisión en magnitud, el modelo sin normalizar mantiene buen desempeño relativo.

En el conjunto de métodos más complejos, Regresión Polinómica grado 4 y Gradient Boosted Trees resultaron los de mejor balance sesgo-varianza; el polinomio de grado 10 demostró señales de sobreajuste. Finalmente, Gradient Boosted Trees fue seleccionado por exhibir el mayor  $R^2$  y los menores MAE, MSE, RMSE y MAPE; estimando la Potencia Pico SADI con un error máximo de -6,91%, dentro de los márgenes de reserva operativa.

## Relación entre complejidad y beneficio

Aumentar la complejidad del modelo más allá de grado 4 no proporcionó mejoras sustanciales en la métrica. Lo cual indica que para el caso de predicción de potencia pico, el modelo de complejidad intermedia (grado 4) es suficiente para capturar la no linealidad sin incurrir en overfitting.

## Comparación entre modelos de regresión y clasificación

El enfoque de regresión es el más adecuado cuando se requiere una estimación continua de la potencia pico, minimizando costos de generación/importación.

El enfoque de clasificación multiclase (baja/media/alta) brinda una herramienta operacional útil para decisiones rápidas de despacho. En este contexto, el modelo de Random Forest superó al de Naive Bayes (85,835% contra 66,908% de exactitud) y demostró mejor capacidad de generalización sin necesidad de utilizar el nodo SMOTE, optándose por una reducción de la clase mayoritaria mediante muestreo.

## Variables clave

Las variables con mayor influencia en la predicción de potencia pico fueron:

- Temperatura media diaria (impacto estacional)
- Energía diaria SADI (correlación con potencia pico)
- Tipo de día (hábil/feriado/sábado/domingo)
- Estación (verano/invierno) y número de semana/año, que aportan contexto temporal

## Equilibrio técnico-económico

El modelo final (Gradient Boosted Trees) permite anticipar la demanda de potencia pico con un error controlado ( $\pm 7\%$ ) y orienta a CAMMESA a ajustar despachos de manera preventiva, evitando costos innecesarios y garantizando la seguridad del sistema.

En conjunto, estos hallazgos confirman que, para la predicción de demanda energética en el SADI, es posible lograr un punto de equilibrio entre simplicidad de modelo, precisión operativa y costos, apoyando la toma de decisiones en tiempo real.

## **PARTE 3 - MODELOS DESCRIPTIVOS Y TÉCNICAS DE EVALUACIÓN**

### 3.1. Clustering

#### 3.1.1. Análisis de la técnica del codo

Para calcular el número adecuado de clúster (k) se aplica la técnica del codo, buscando minimizar el error cuadrático (la distancia cuadrada entre cada punto y su centroide dentro de los grupos). Fue necesario calcular el WSS (Within-Cluster Sum of Squares) total para cada valor de k dentro del rango [1;10]. Se utiliza una tabla con los posibles valores que va a tener k, que son iterados en un bucle (loop).

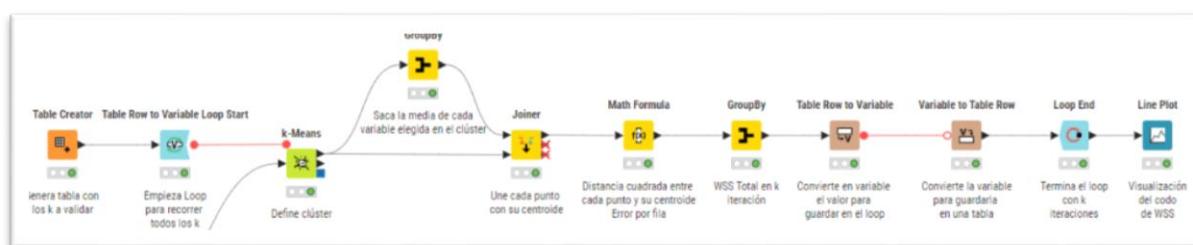
En cada iteración, se define el clúster utilizando las variables seleccionadas y se varía el número de k. Dentro de cada iteración se obtiene la media de cada atributo por clúster, se asigna cada punto a su centroide, y se calcula la distancia cuadrada entre cada punto y su centroide (error por fila).

Por último, se realiza la sumatoria de esos errores para calcular el WSS y, con ello, graficar el codo.

	Genera una tabla de una columna que contiene los valores de k a evaluar (de 1 a 10), para determinar el número óptimo de clústeres.
	Unifica un loop, donde convierte cada fila de la tabla de valores k en una variable de flujo. Ejecuta el algoritmo k-means una vez por cada valor de k definido.
	Ejecuta el algoritmo de k-means utilizando las variables seleccionadas como entrada. Se configura para que asigne los datos a k clústeres utilizando las variables: <ul style="list-style-type: none"> <li>AÑO - Energía SADI (GWh) - Potencia Pico SADI (MW) - Temperatura Media Diaria GBA (°C) - VERANO - INVIERNO - FERIADO - HÁBIL - SÁBADO - DOMINGO.</li> </ul>
	Calcula la media de cada variable seleccionada dentro de cada clúster, buscando resumir las características de cada grupo para análisis luego.
	Une cada instancia del conjunto de datos con su centroide.
	Calcula el error cuadrático por fila. Código: $\text{pow}(\$Potencia Pico SADI (MW) - \$Mean(Potencia Pico SADI (MW)), 2) + \text{pow}(\$Energía SADI (GWh) - \$Mean(Energía SADI (GWh)), 2) + \text{pow}(\$Temperatura Media Diaria GBA (°C) - \$Mean(Temperatura Media Diaria GBA (°C)), 2) + \text{pow}(\$AÑO - \$Mean(AÑO), 2) + \text{pow}(\$VERANO - \$Mean(VERANO), 2) + \text{pow}(\$INVIERNO - \$Mean(INVIERNO), 2) + \text{pow}(\$HÁBIL - \$Mean(HÁBIL), 2) + \text{pow}(\$SÁBADO - \$Mean(SÁBADO), 2) + \text{pow}(\$DOMINGO - \$Mean(DOMINGO), 2)$
	Suma el error cuadrático de todas las filas para poder obtener el WSS.
	Convierte el valor del WSS total en una variable de flujo, guardándolo en cada iteración del loop.

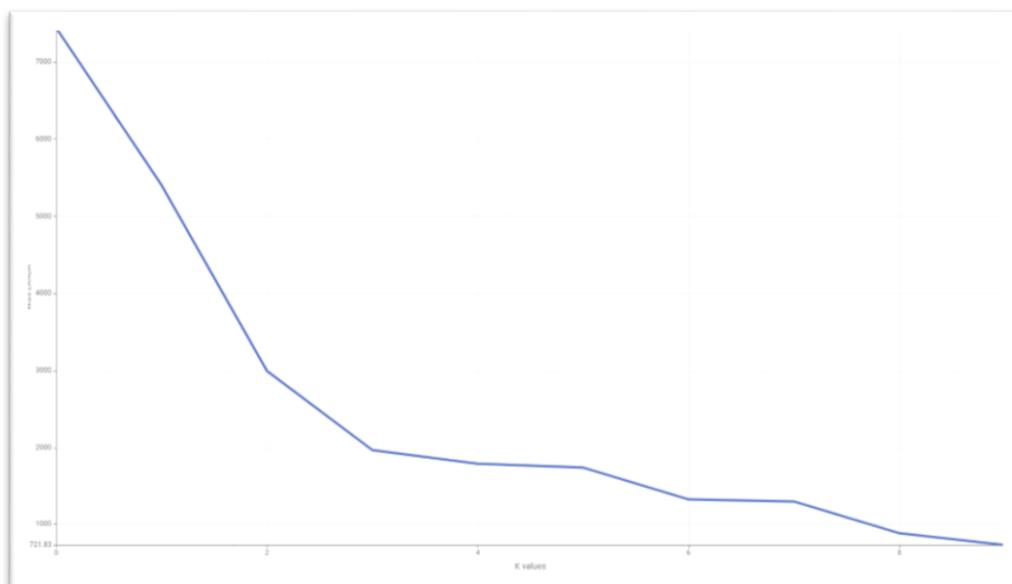
	Convierte la variable del WSS total en una fila de una tabla.
	Finaliza el loop.
	Permite visualizar el codo de WSS.

A continuación se muestra el workflow con los nodos utilizados para utilizar la técnica del codo.



### 3.1.2. K-means, Silhouette y Entropía

Una vez utilizada la técnica del codo para conocer cuántos clústeres de forma natural hay para agrupar los datos sin supervisión, y conocer su agrupación natural, se obtiene que la misma es entre 3 y 6. Para evaluar la separación de estos clústeres, se procede a calcular los coeficientes de Silhouette (métrica interna). Si este coeficiente es cercano a 1, los clústeres son compactos y están bien separados; si es 0, se considera que el punto está en la frontera entre dos clústeres; y si es -1, el punto está mal asignado a un clúster.



Para el cálculo del coeficiente de Silhouette se utilizaron los siguientes nodos:

<b>One to Many</b> 	Codifica las variables de categóricas a numéricas.
<b>Normalizer</b> 	Normaliza los datos entre 0 y 1.
<b>k-Means</b> 	Define una cantidad de clústeres considerando todas las variables de trabajo excepto la Potencia pico (por ser la variable objetivo) y la energía, fuertemente correlacionada a ella.
<b>Siluete</b> 	Metanodo generado con el fin de concentrar diversos nodos para el cálculo del coeficiente de Silhouette.
<b>Silhouette Coefficient</b> 	Calcula el coeficiente de Silhouette correspondiente a la cantidad de clústeres definidos en k-means en función de la distancia entre ellos.
<b>Color Manager</b> 	Se definen colores a cada uno de los clústeres a fin de poder diferenciarlos en las gráficas.
<b>Scatter Plot</b> 	Grafica la relación entre la Potencia y el coeficiente de Silhouette.
<b>Histogram</b> 	Histograma para visualizar la frecuencia de valores del coeficiente de Silhouette.

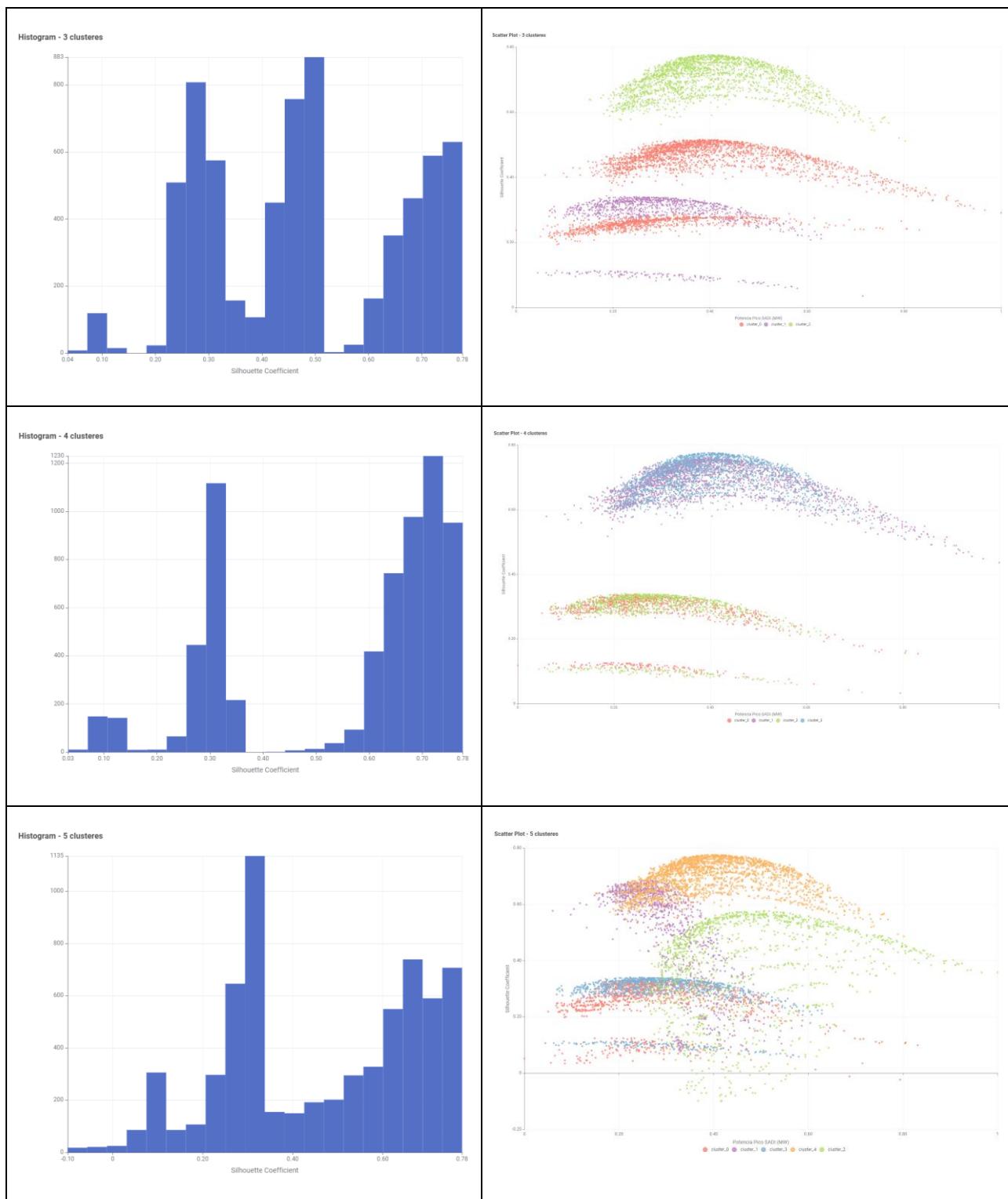
Se calculó el coeficiente de Silhouette para los diferentes valores de k considerados, teniendo en cuenta que el óptimo es aquel en el que la curva se aplana (valor óptimo de k).

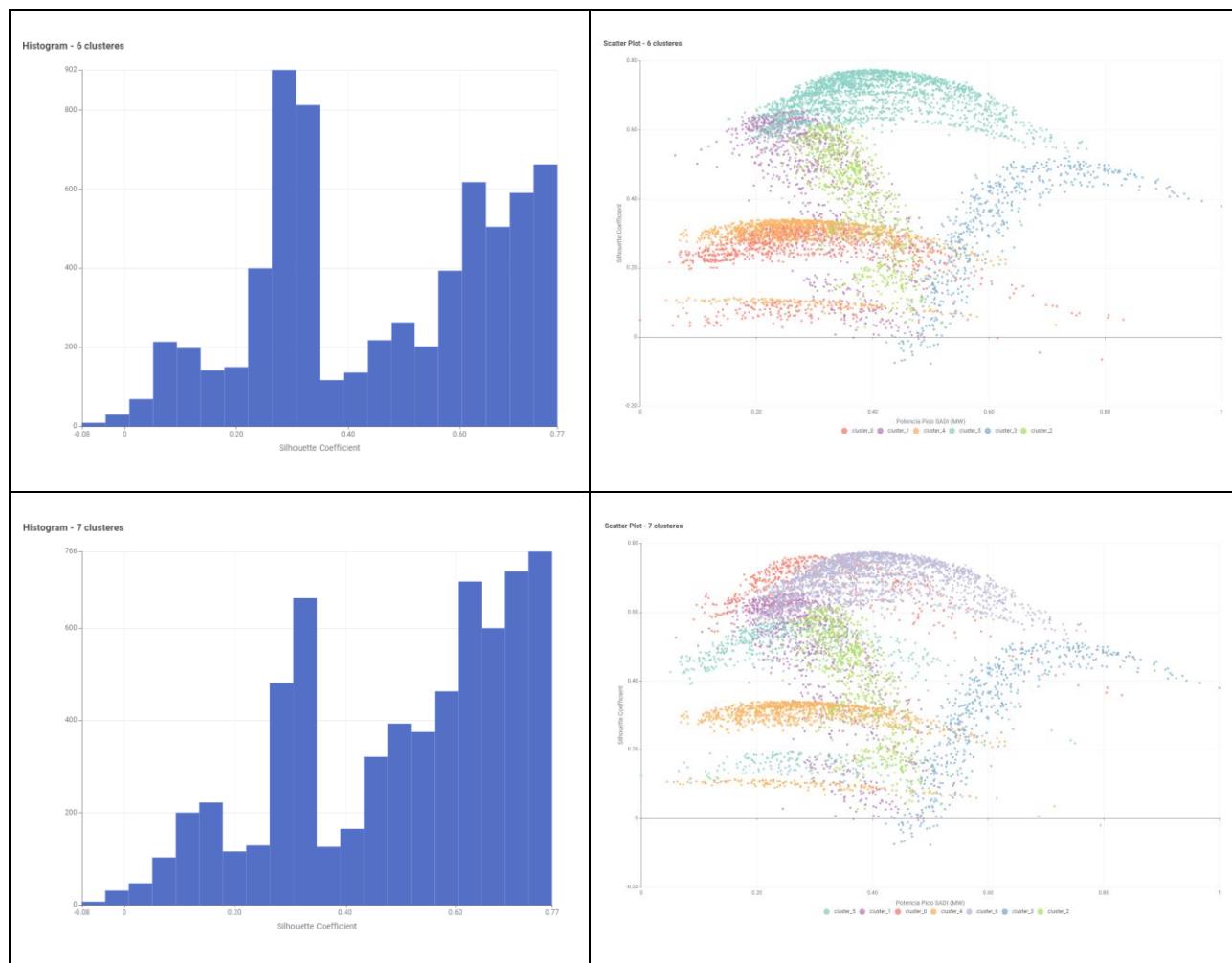
Luego, se analizará cuál es el valor más conveniente, siendo para un k menor al óptimo representa un modelo simplificado con menor precisión, mientras que un k mayor, genera un modelo más detallado, con posible sobreajuste.

Los valores del coeficiente varían de la siguiente forma según la cantidad de clústeres:

Cantidad de clústeres (k)	Coeficiente de Silhouette
3	Entre 0,035 y 0,776
4	Entre 0,033 y 0,776
5	Entre -0,1 y 0,776
6	Entre -0,077 y 0,775
7	Entre -0,077 y 0,775

Las mismas presentan las siguientes distribuciones del coeficiente, según la cantidad de clústeres considerados:





Se aprecia cómo los clústeres definidos se superponen en mayor medida al aumentar su cantidad, lo que sugiere valores altos de entropía en esas zonas. Además, para los casos de 3 y 4 clústeres, se observa que los valores mínimos del coeficiente de Silhouette son positivos, a diferencia de los demás. En consecuencia, se debe decidir si conviene simplificar el modelo o aumentar su complejidad, con el riesgo de un posible sobreajuste. A fin de seleccionar la cantidad de clústeres más conveniente, se procede a calcular la entropía de cada una de ellos.

### Entropía

Dentro de un metanodo, se calcula la entropía correspondiente a cada clúster y se evalúa, considerando que una alta entropía indica que el clúster contiene distintas clases de una variable categórica mezcladas (menos informativo), mientras que una entropía baja indica un clúster compuesto principalmente por una sola clase (más puro).

Para este estudio, se clasificará la Potencia en Baja, Media y Alta según el nivel de consumo, y se busca que los clústeres, respecto de esta variable, sean lo más homogéneos, a fin de reducir la entropía y mejorar la segmentación.

A continuación, se explica el metanodo Entropía:

	Define numéricamente niveles de potencia según el consumo: \$Potencia Pico SADI (MW)\$ <= 17000 => "0" \$Potencia Pico SADI (MW)\$ <= 20000 => "0.5" TRUE => "1"
	Detalla clases de consumo de potencia: \$Potencia Pico SADI (MW)\$ <= 17000 => "BAJA" \$Potencia Pico SADI (MW)\$ <= 20000 => "MEDIA" TRUE => "ALTA"
	Convierte la variable categórica numérica de Potencia de String a number.
	Normaliza (0-1) las variables independientes de todo el dataset, excepto la Potencia Pico (variable objetivo), la energía SADI (correlacionada a Potencia Pico), y variables que se consideran de bajo peso para el análisis.
	Define la cantidad de clúster a considerar con las variables normalizadas.
	Edita la clasificación que genera k-means, quedándose solamente con el número de clúster en una nueva columna.
	Convierte de String a número, el número que figura en columna clúster.
	Se agrupa por tipo de demanda (alta, media, baja) y clúster correspondiente, y se cuenta la cantidad de nivel de demanda en cada caso. En paralelo se agrupa por clúster y se cuenta la cantidad de muestras que hay por clúster.
	Se unen las 2 tablas generadas anteriormente para facilitar el cálculo de la entropía.
	Cálculo de proporción relativa de clase por clúster. Y cálculo de término de entropía de cada clase por clúster.
	Sumatoria de los términos de entropía de cada clase por clúster para tener la entropía correspondiente a cada clúster ( $H_i$ ).

Valores de entropía por clúster Hi:

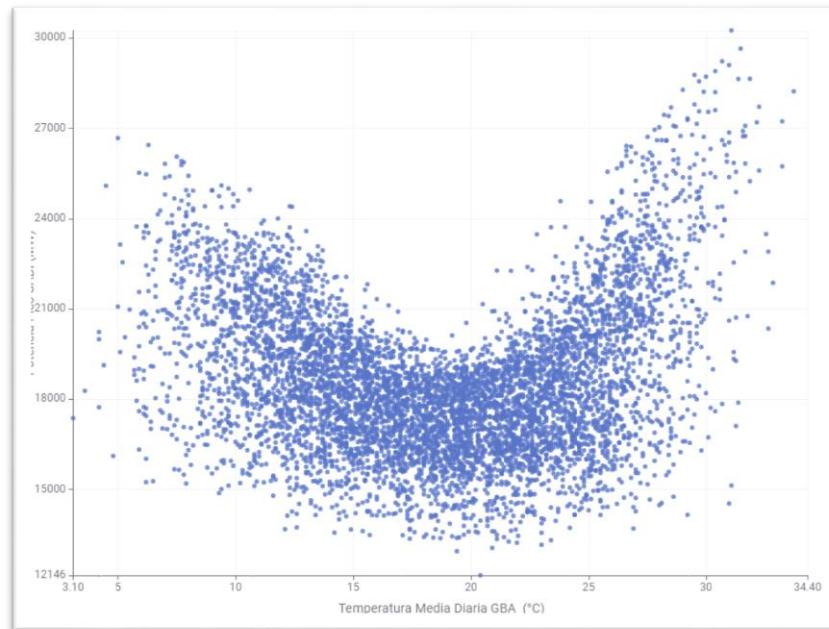
RowID	New Cluster Number (integer)	Entropía de Cluster Hi Number (double)	RowID	New Cluster Number (integer)	Entropía de Cluster Hi Number (double)
Row0	0	1.535	Row0	0	1.366
Row1	1	1.367	Row1	1	1.455
Row2	2	1.372	Row2	2	1.367
			Row3	3	1.372
New Cluster ↓ Number (integer)		Entropía de Cluster Hi Number (double)	New Cluster Number (integer)		Entropía de Cluster Hi Number (double)
4		1.372	0		1.366
3		1.367	1		1.248
2		1.082	2		0.745
1		1.215	3		0.021
0		1.366	4		1.367
			5		1.372
					1.274
					1.372

Los valores obtenidos indican que la cantidad de clústeres más conveniente es 6, dado que presenta menores valores de entropía. Se destaca el clúster 3, con una clase más definida.

### 3.1.3. Conclusiones preliminares

Por una parte, los valores obtenidos de coeficiente de Silhouette indican que un k entre 3 o 4 sería el más conveniente, dado que alcanzan valores mayores a 0, evitando que queden puntos mal asignados a un clúster. Sin embargo, si se considera la entropía, se aprecia que al aumentar k, esta se reduce, resultando un más conveniente un k=6, ya que disminuye la presencia de distintas clases de la variable categórica Potencia dentro de un mismo clúster.

Estos elevados valores de entropía pueden comprenderse dada la complejidad de nuestro problema, en el que existen casos donde la Potencia pico y la Energía están correlacionadas, pero también situaciones en las que, si bien una aumenta, la otra se reduce. Así mismo, se observa una variación parabólica del consumo de Potencia con la Temperatura, alcanzando valores similares de Potencia para diferentes valores de temperatura.



A fin de buscar un equilibrio entre la separación de los clústeres y la pureza de una clase dentro de un clúster, se opta por un k=4. Trabajos futuros podrían mejorar estos parámetros, ya sea modificando la cantidad de clases de la variable categórica Potencia o ajustando los valores que definen cada clase.

En caso de optar por una normalización del tipo z-score, se aprecian leves variaciones en la entropía, sin afectar en gran medida al k seleccionado.

RowID	New Cluster Number (integer)	Entropia de Cluster Hi Number (double)
Row0	0	1.27
Row1	1	1.457
Row2	2	1.417
Row3	3	1.373

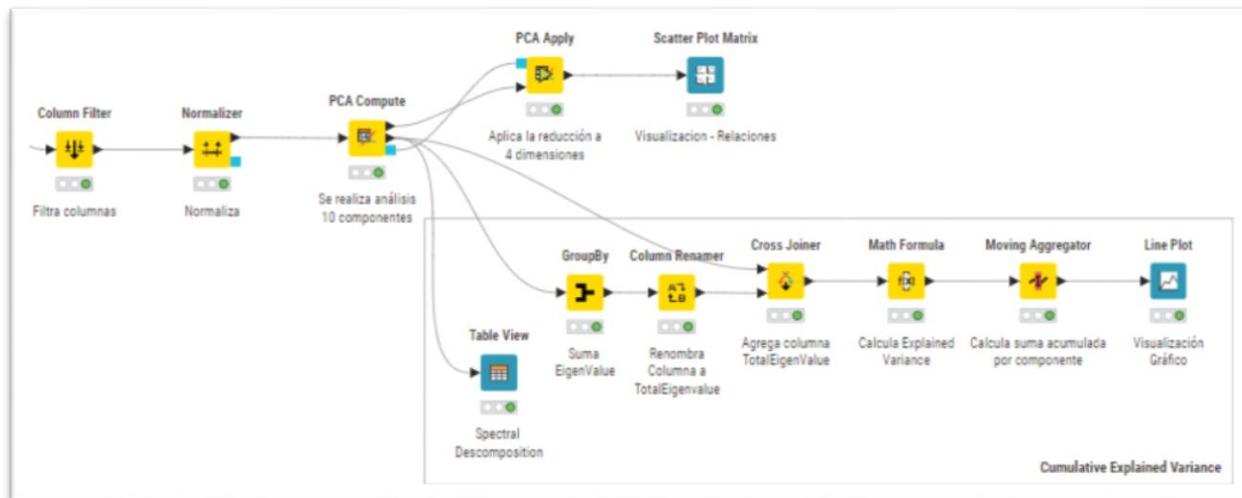
### 3.2. PCA

Se llevó a cabo un análisis PCA (Análisis de componentes principales) para reducir la dimensionalidad del conjunto de datos mientras se busca preservar las relaciones más importantes. Además, se busca reducir la cantidad de dimensiones facilitando el agrupamiento, reduciendo ruido y redundancia. La desventaja de la técnica de PCA es que puede llegar a perder información. En este análisis se busca preservar al menos un 90% de información tras la reducción de dimensión.

Para tal fin, se desarrolla el siguiente flujo de trabajo:

	Se filtran las variables relevantes que se incluirán en el análisis: <ul style="list-style-type: none"><li>• AÑO - Energía SADI (GWh) - Potencia Pico SADI (MW) - Temperatura Media Diaria GBA (°C) - VERANO - INVIERNO - FERIADO - HÁBIL - SÁBADO - DOMINGO.</li></ul>
	Se normalizan las variables seleccionadas, escalando los valores numéricos a una misma escala.
	Ejecuta el análisis de componentes principales (PCA), generando hasta 10 componentes principales, buscando reducir la dimensionalidad del conjunto de datos, buscando mantener la mayor cantidad posible de varianza.
	Aplica la transformación PCA previa, reduciendo los datos originales a las primeras 4 componentes principales, que explican la mayor cantidad de varianza.
	Visualización grafica
	Visualización de los resultados de la descomposición espectral del PCA
	Calcula la suma de los eigenvalues (varianza) que resultan del análisis de componentes principales.
	Renombra la columna “Sum(eigenvalue)” a “TotalEigenvalue”.
	Agrega la columna de los eigenvalues a la tabla original de componentes.
	Calcula la varianza explicada por cada componente principal. Código: \$eigenvalue\$ / \$TotalEigenvalue\$
	Calcula la suma acumulada de la varianza explicada por componente
	Visualización grafica

A continuación, se muestra el workflow completo, donde se puede observar los nodos y su secuencia.

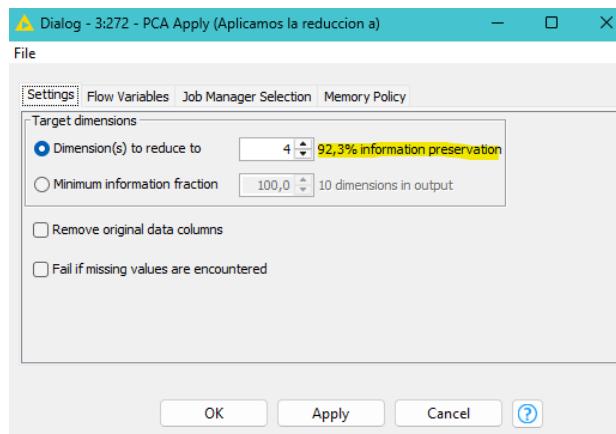


### 3.2.1. Conclusiones preliminares

Se utilizó la métrica Varianza explicada acumulada para validar el PCA. Esta métrica nos informa cuánto de la información original se conserva tras reducir dimensiones. Para saber cuántos componentes se necesitan se realizó el workflow mencionado. En el nodo “Moving Aggregator” se puede observar lo siguiente:

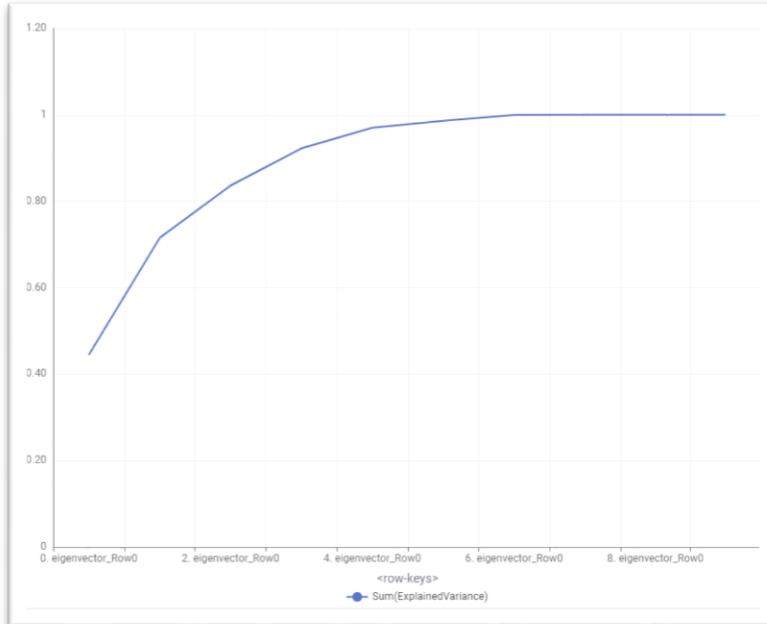
Sum(ExplainedV... Number (double))
0.445
0.715
0.836
0.923
0.97
0.986
1
1
1

Se observa que al cuarto componente obtenemos un 92,3% de información preservada. También observamos que el nodo “PCA Apply” cuando se selecciona 4 dimensiones realiza automáticamente el valor y arroja el mismo número de preservación. 92,3%.

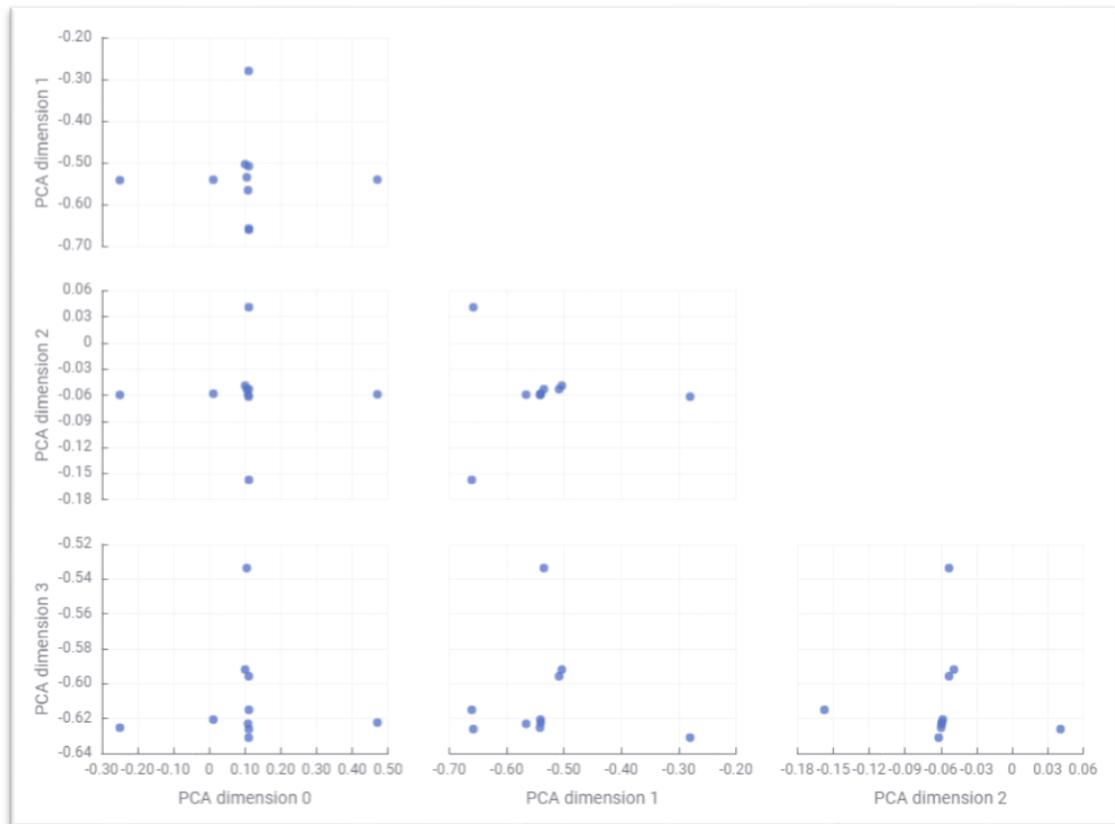


**Se confirma entonces que la reducción de 10 a 4 dimensiones se puede realizar logrando un 92,3% de información preservada con los componentes disponibles.**

Adicionalmente, se obtiene un gráfico donde se puede visualizar cómo se preserva la información a medida que se agregan los componentes de forma ordenada hasta llegar a un 100% de información preservada. Los primeros componentes explican la mayor parte de la varianza, y se observa que cada componente adicional aporta menos que el anterior. Esto genera una curva que sube rápido al principio y se aplana, generando un punto de codo. Agregar más dimensiones a partir del punto de codo observamos que no aporta a la preservación de información.



Además, se agregó un Scatter Plot con las 4 dimensiones, para observar las relaciones de los distintos componentes. Se puede observar separaciones claras entre verano – invierno y entre sábado/domingo y días hábiles.



### 3.3. Reglas de Asociación

En este tópico se busca, por medio de reglas de asociación, patrones frecuentes y/o correlaciones entre conjuntos de datos para responder a las preguntas planteadas la sección “[Objetivo de negocio e hipótesis](#)”.

El nodo principal para el desarrollo del workflow es el “Association Rule Learner”, su función principal es encontrar reglas de asociación interesantes entre los datos. Este toma las variables categóricas en formato vector y devuelve la frecuencia con la que se presentan estos parámetros entre sí, siempre que sea superior a un umbral mínimo de confianza configurable.

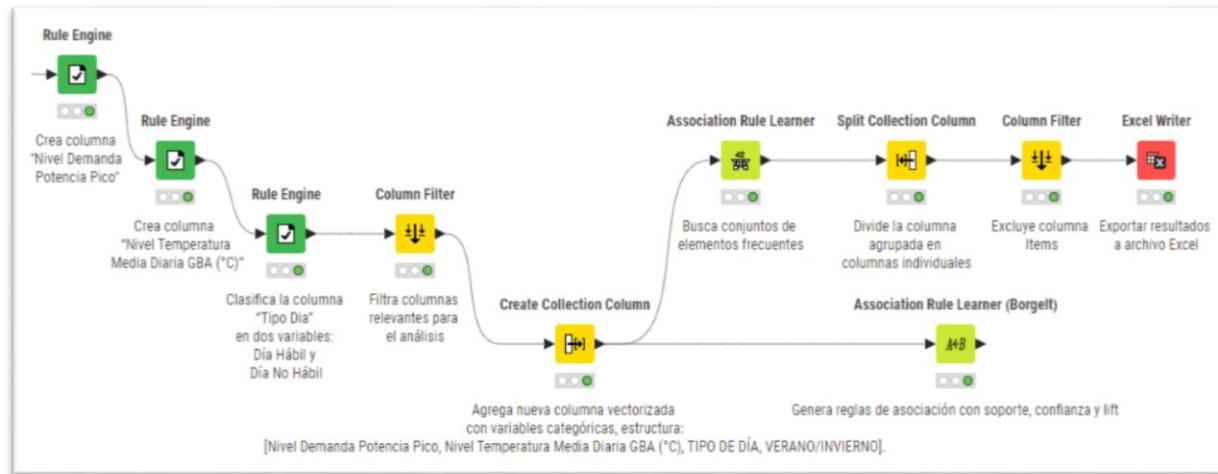
Debido a los requisitos del formato de entrada se necesita realizar un preprocesamiento de los datos para que el nodo “Association Rule Learner” pueda procesarlos correctamente y luego, un postprocesamiento para poder extraer conclusiones de lo obtenido.

A continuación, se presenta el flujo de trabajo implementado:

 <b>Rule Engine</b>	<p>Crea una nueva columna categórica con el nombre “Nivel Demanda Potencia Pico”, se clasifica la Potencia Pico SADI en los siguientes límites: baja, si es menor o igual a 17000 MW; media, si es menor o igual a 20000 MW; alta, para los casos restantes. Código:</p> <pre>\$Potencia Pico SADI (MW)\$ &lt;= 17000 =&gt; "Potencia Pico Baja" \$Potencia Pico SADI (MW)\$ &lt;= 20000 =&gt; "Potencia Pico Media" TRUE =&gt; "Potencia Pico Alta"</pre>
------------------------	--

 Rule Engine	<p>Crea una nueva columna categórica con el nombre “Nivel Temperatura Media Diaria GBA (°C)”, se clasifica la temperatura en los siguientes límites: temperatura baja, si en menor o igual a 15°C; temperatura media, si es menor o igual a 25°C; temperatura alta, para los casos restantes. Código:</p> <pre>\$Temperatura Media Diaria GBA(°C)\$ &lt;= 15 =&gt; "Temperatura Baja" \$Temperatura Media Diaria GBA(°C)\$ &lt;= 25 =&gt; "Temperatura Media" TRUE =&gt; "Temperatura Alta"</pre>
 Rule Engine	<p>Se clasifica la columna “Tipo Dia” en dos variables: Día Hábil y Día No Hábil, reemplazando los días Sábado, Domingo y Feriado por Día No Hábil, para aumentar la frecuencia entre variables y tener una mejor comprensión de nuestros datos. Código:</p> <pre>\$TIPO DIA\$ = "HÁBIL" =&gt; "HÁBIL" TRUE =&gt; "NO HÁBIL"</pre>
 Column Filter	<p>Se filtran únicamente las columnas relevantes para nuestro análisis:</p> <ul style="list-style-type: none"> <li>Nivel Demanda Potencia Pico - Nivel Temperatura Media Diaria GBA (°C) - TIPO DE DÍA - VERANO/INVIERNO</li> </ul>
 Create Collection Column	<p>Se agrega una nueva columna vectorizada de las variables categóricas, para cumplir con el formato de entrada requerido por el nodo Association Rule Learner. La estructura resultante es:</p> <p>[Nivel Demanda Potencia Pico, Nivel Temperatura Media Diaria GBA (°C), TIPO DE DÍA, VERANO/INVIERNO].</p>
 Association Rule Learner	<p>Busca conjuntos de elementos frecuentes que cumplan con el criterio de soporte mínimo definido por el usuario y, opcionalmente, genera reglas de asociación a partir de ellos.</p> <p>Como resultado, se crea una nueva columna llamada “Support(0-1)”, que muestra para cada fila, el valor del soporte como una proporción entre 0 y 1. Este valor indica la frecuencia con la que aparece ese conjunto de elementos en el total de datos analizados.</p> <p>Se estableció un umbral mínimo soporte de 0,01 (1%).</p>
 Split Collection Column	<p>Separa la columna que contiene varios valores agrupados en una sola celda, dividiéndolos nuevamente en columnas individuales. Esto permite ver cada variable por separado y facilita la exportación a Excel.</p>
 Excel Writer	<p>Se utiliza para exportar los resultados a un archivo de Excel, permitiendo una mejor compresión de los resultados y facilita la redacción de conclusiones.</p>
 Association Rule Learner (Borgelt)	<p>Genera reglas de asociación de forma eficiente, mostrando medidas como soporte, confianza y lift. Permite identificar relaciones relevantes entre variables categóricas y evaluar su importancia.</p>

A continuación se muestra el workflow con los nodos utilizados para el preprocesamiento de los datos, la transformación de las variables categóricas, la aplicación de reglas de asociación y la exportación de los resultados.



A modo ilustrativo, a continuación se puede observar algunas de las filas generadas y exportadas a Excel para su análisis. En ellas pueden observarse los conjuntos identificados junto a su valor de soporte.

Support(0-1):	Split Value 1	Split Value 2	Split Value 3	Split Value 4
0,010099487	Temperatura Alta	HÁBIL	Potencia Pico Baja	
0,010853181	Potencia Pico Media	VERANO	Temperatura Baja	
0,011305396	HÁBIL	VERANO	Temperatura Baja	
0,012360567	Temperatura Alta	Potencia Pico Baja	VERANO	NO HÁBIL
0,012812783	Temperatura Alta	Potencia Pico Baja	NO HÁBIL	
0,01416943	HÁBIL	INVIERNO	Potencia Pico Baja	Temperatura Baja
0,016430509	Potencia Pico Alta	INVIERNO	Temperatura Baja	NO HÁBIL
0,016581248	HÁBIL	Potencia Pico Baja	Temperatura Baja	
0,017787157	Potencia Pico Media	INVIERNO	Temperatura Media	NO HÁBIL

### 3.3.1. Conclusiones preliminares

A partir de las 113 reglas obtenidas se pueden identificar las siguientes conclusiones principales:

Las variables individuales tienen una alta incidencia por sí mismas, sin necesidad de combinarse con otros atributos. Por ejemplo, los días Hábiles representan el 67,41% de las observaciones, seguido por la Temperatura Media con un 53,38%, el Verano con 50,35% y el Invierno con un 49,65%.

Support(0-1):	Split Value 1	Split Value 2	Split Value 3	Split Value 4
0,496533012	INVIERNO			
0,503466988	VERANO			
0,533765451	Temperatura Media			
0,674103105	HÁBIL			

Respecto a la distribución del soporte, los valores oscilan aproximadamente entre 0,0101 y 0,6741. La media del soporte es de ~0,1317 lo cual indica que las reglas no son muy frecuentes y que la media esta sesgada hacia arriba por algunos valores altos. Por otro lado, la mediana es de ~0,0983 lo cual refuerza la idea de una distribución asimétrica.

Support(0-1):	Split Value 1	Split Value 2	Split Value 3	Split Value 4
0,010099487	Temperatura Alta	HÁBIL	Potencia Pico Baja	
0,674103105	HÁBIL			

El 75% de las reglas tiene un soporte inferior a aproximadamente 0,1739 lo que indica que las combinaciones de múltiples variables (3 o 4 variables) son menos frecuentes que las reglas de una sola variable. Esto se puede observar en la tabla siguiente, donde se detallan las reglas que combinan 4 variables:

Support(0-1):	Split Value 1	Split Value 2	Split Value 3	Split Value 4
0,012360567	Temperatura Alta	Potencia Pico Baja	VERANO	NO HÁBIL
0,01416943	HÁBIL	INVIERNO	Potencia Pico Baja	Temperatura Baja
0,016430509	Potencia Pico Alta	INVIERNO	Temperatura Baja	NO HÁBIL
0,017787157	Potencia Pico Media	INVIERNO	Temperatura Media	NO HÁBIL
0,018239373	Temperatura Alta	Potencia Pico Media	VERANO	NO HÁBIL
0,019294543	Temperatura Alta	Potencia Pico Alta	VERANO	NO HÁBIL
0,025776304	Temperatura Alta	HÁBIL	Potencia Pico Media	VERANO
0,026982213	HÁBIL	Potencia Pico Alta	VERANO	Temperatura Media
0,028188122	INVIERNO	Potencia Pico Baja	Temperatura Baja	NO HÁBIL
0,028489599	HÁBIL	INVIERNO	Potencia Pico Baja	Temperatura Media
0,03165511	Potencia Pico Media	VERANO	Temperatura Media	NO HÁBIL
0,047030449	Potencia Pico Media	INVIERNO	Temperatura Baja	NO HÁBIL
0,048387097	HÁBIL	Potencia Pico Baja	VERANO	Temperatura Media
0,050949653	INVIERNO	Potencia Pico Baja	Temperatura Media	NO HÁBIL
0,074464878	Temperatura Alta	HÁBIL	Potencia Pico Alta	VERANO
0,075520048	Potencia Pico Baja	VERANO	Temperatura Media	NO HÁBIL
0,082001809	HÁBIL	Potencia Pico Media	INVIERNO	Temperatura Baja
0,097829364	HÁBIL	Potencia Pico Alta	INVIERNO	Temperatura Baja
0,105517033	HÁBIL	Potencia Pico Media	INVIERNO	Temperatura Media
0,142297257	HÁBIL	Potencia Pico Media	VERANO	Temperatura Media

Para interpretar las reglas de asociación obtenidas, se decidió ordenar la tabla en función del Consecuente, ya que el objetivo principal de este análisis es identificar las condiciones bajo las cuales se produce una Potencia Pico Alta, Medio o Baja.

Nos enfocaremos especialmente en la Potencia Pico Alta debido a sus posibles consecuencias operativas, como la sobrecarga del sistema y potenciales cortes del suministro eléctrico.

En este sentido, se identificaron una regla relevante donde la combinación de día Habil, temperatura alta y Verano, permite predecir una potencia Pico Alta en un 67,8% de confianza. Además, con un Lift del 2.796, indica que esta relación ocurre casi tres veces más de lo que se esperaría si las variables fueran independientes, lo que sugiere una relación estadísticamente significativa y potencialmente útil para tareas de planificación y prevención.

Por último, presenta un soporte del 7,45% (494 ocurrencias sobre el total de días analizados), lo cual indica que se trata de un patrón con una frecuencia razonable. Por todas estas razones, esta regla puede considerarse tanto confiable como relevante desde una perspectiva operativa.

#	RowID	Consequent String	Antecedent List	Item Number	RelativeItemSetSupport% Number (double)	RuleConfidence% Number (double)	AbsoluteBodyNumber	RelativeBodyNumber	RuleLift	RuleLift% Number (double)
Row	1 selected	Antecedent	ItemS	RelativeItemSetSupport%	RuleConfidence%	AbsoluteBodySt	RelativeBodySel	RuleLift	RuleLift%	
3	Row2	Potencia Pico Alta	[HABIL,Temperatura Alta,VERANO]	494	7.447	67.8	729	11	2.796	279.57
7	Row6	Potencia Pico Alta	[HABIL,Temperatura Alta]	499	7.522	67.3	742	11.2	2.775	277.45
5	Row4	Potencia Pico Alta	[Temperatura Alta,VERANO]	622	9.376	58.7	1,060	16	2.421	242.09
9	Row8	Potencia Pico Alta	[Temperatura Alta]	627	9.451	58.1	1,079	16.3	2.397	239.74

### 3.4. Conclusiones Modelos Descriptivos

Luego de analizar el dataset con diferentes técnicas descriptivas y su posterior evaluación, con el objetivo de lograr una mayor comprensión de los datos y detectar patrones relevantes para la predicción de la demanda energética SADI, arribamos a las siguientes conclusiones.

#### Clustering

Se identificaron agrupamientos naturales en los datos utilizando el nodo K-means. Se utilizaron métricas internas (Coeficiente de Silhouette) y métricas externas (cálculo de la entropía). Todo esto llevo a seleccionar un k=4, valor que mostró una separación aceptable de los clústeres respecto al nivel de potencia. Sin embargo, los valores de entropía obtenidos no muestra una buena homogeneidad dentro de los clústeres (mezcla de clases), lo cual se debe a la complejidad del fenómeno energético, influido por múltiples factores.

#### Análisis de Componentes Principales (PCA)

Este análisis permitió reducir la dimensionalidad del conjunto de datos a 4 componentes principales, manteniendo un 92,3% de la varianza original. Lo que permitió visualizar la relación existente entre las variables, revelando agrupamientos como la diferenciación entre días hábiles y no hábiles, o entre verano e invierno.

#### Reglas de asociación

Mediante esta técnica se identificaron combinaciones de variables categóricas con alta relevancia operativa. Se destaca una regla con un 67,8% de confianza y un lift de 2,796, que vincula días hábiles, temperaturas altas y verano con una alta demanda de potencia, proporcionando información clave para la planificación energética.