

Análisis y Predicción de la Demanda Energética del SADI

Universidad Nacional De La Matanza

Especialización Ciencia de Datos

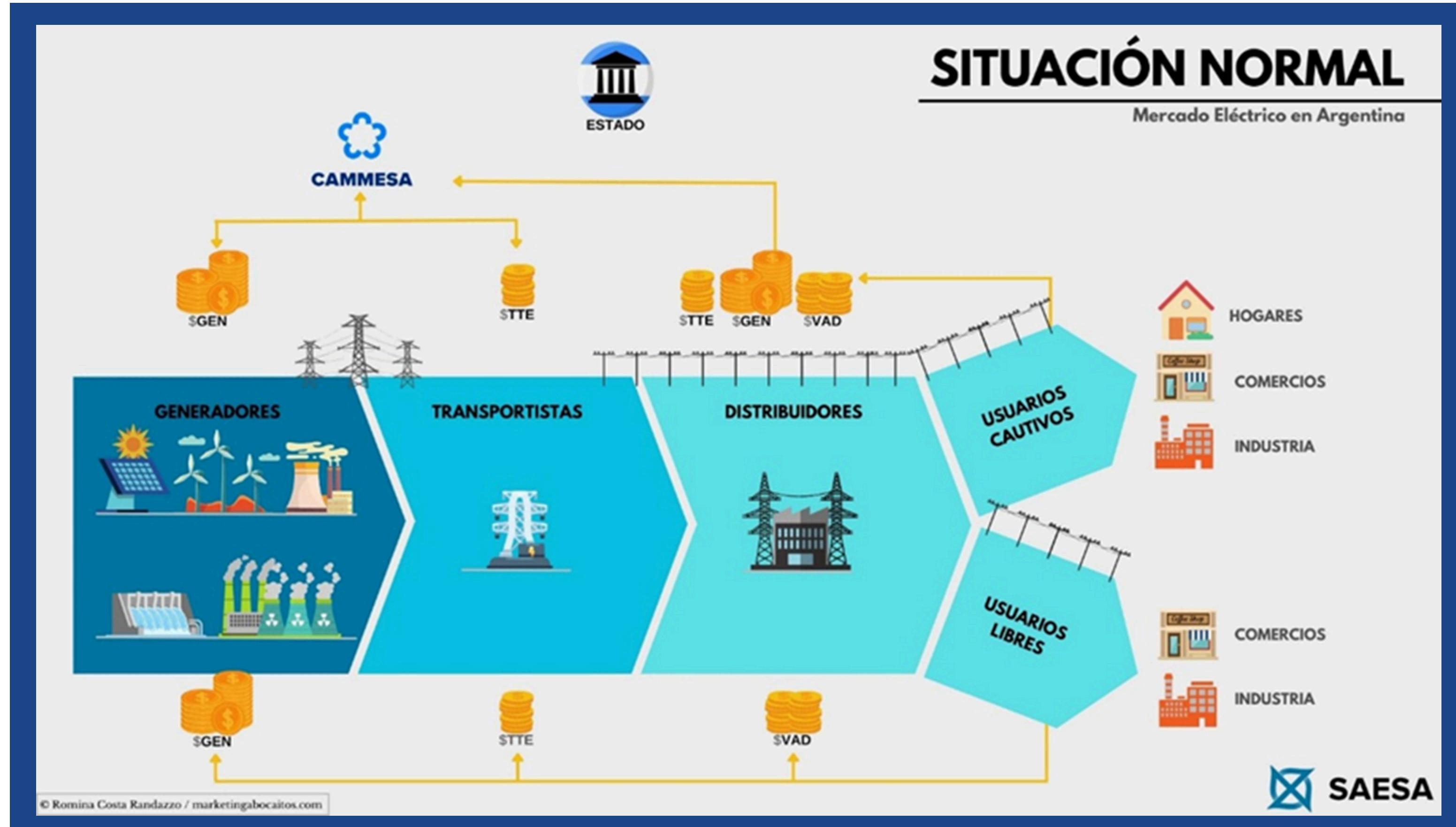
Materia: Minería de Datos

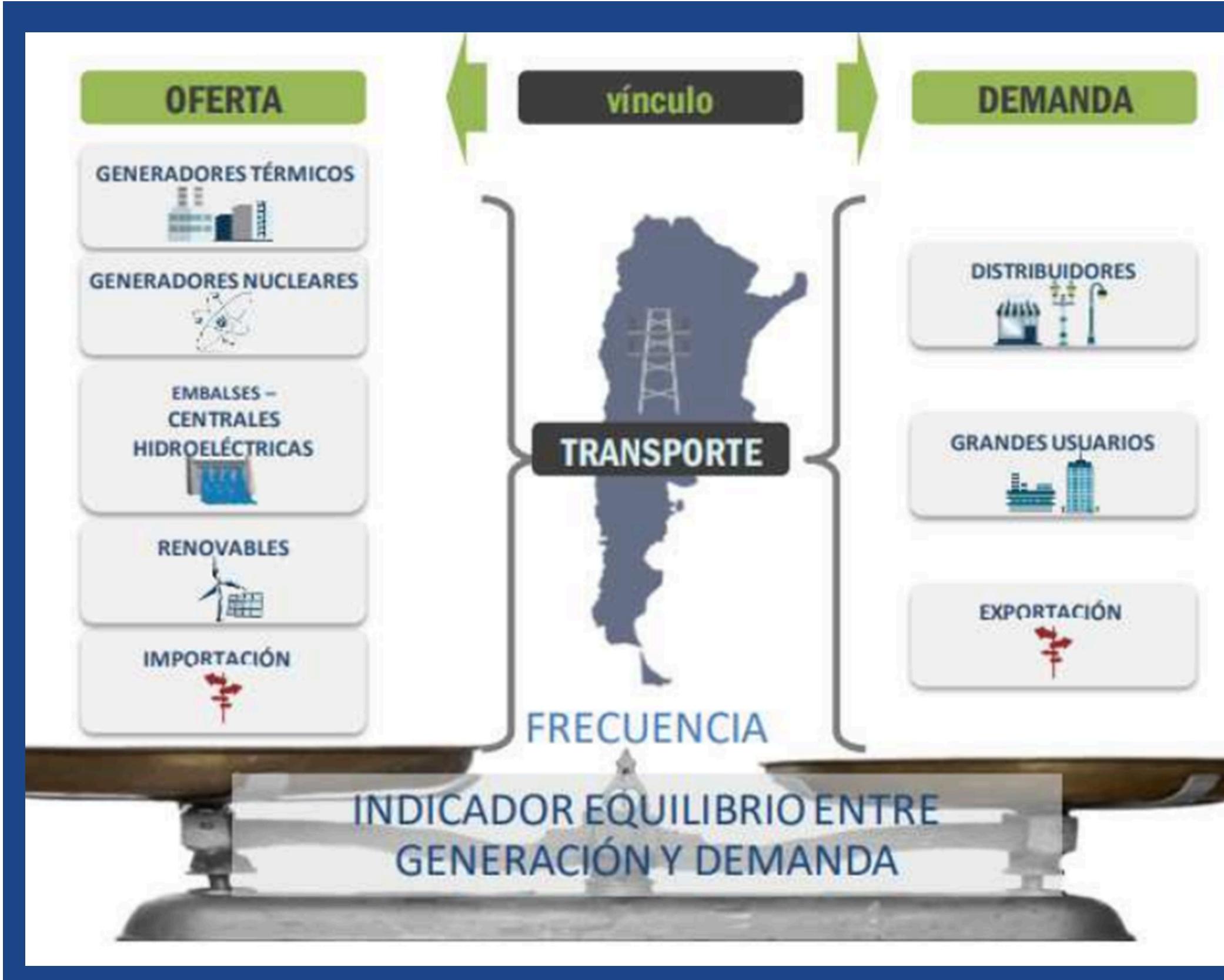
Grupo: A - Integrantes:

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



Caso de Estudio de negocio





Objetivos del Caso de Estudio

La demanda de energía eléctrica debe ser cubierta en tiempo real.

El equilibrio entre generación y consumo es crítico para mantener la estabilidad del sistema.

OBJETIVO: Aplicar Minería de Datos para analizar patrones de consumo y predecir la demanda.

Dataset Utilizado

Fuente: CAMMESA - Sección "Máximos Históricos de Energía y Potencia Estacionales del SADI"

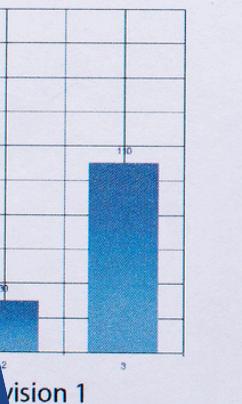
Acceso desde: [dataset](#)

VARIABLES PRINCIPALES:

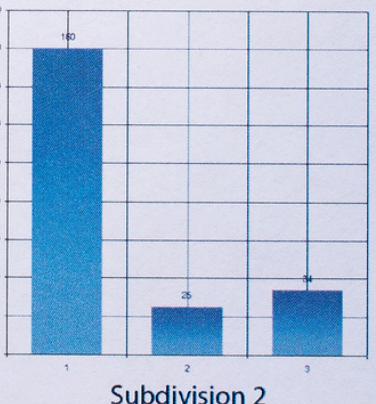
- Fecha
- Tipo de Día (hábil/no hábil)
- Energia SADI
- Potencia Pico SADI
- Hora Potencia Pico
- Temperatura Media Diaria GBA (°C)
- Estado del Tiempo



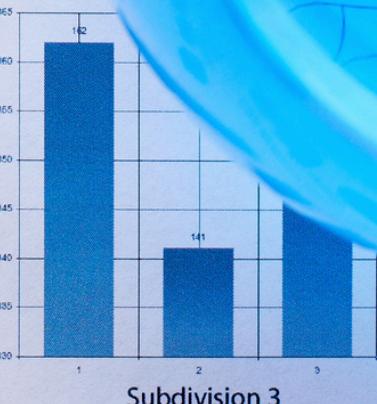
Detailed information of changing business activity of subdivisions of main



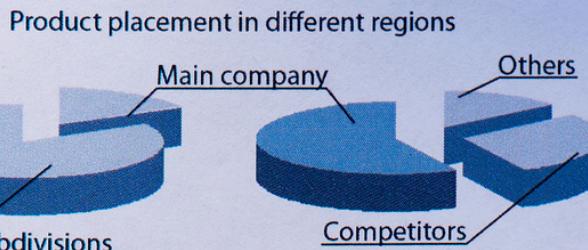
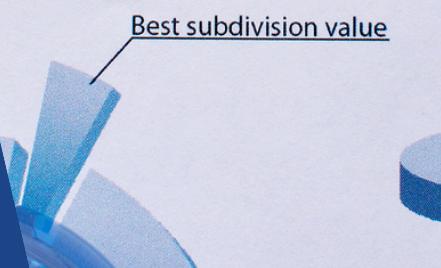
Subdivision 1



Subdivision 2



Subdivision 3



The given analytical report allows to estimate to the full a current situation both in all company, and in its divisions separately. It will allow to predict more precisely immediate prospects of development of the company at the account of preservation of positive dynamics of growth.

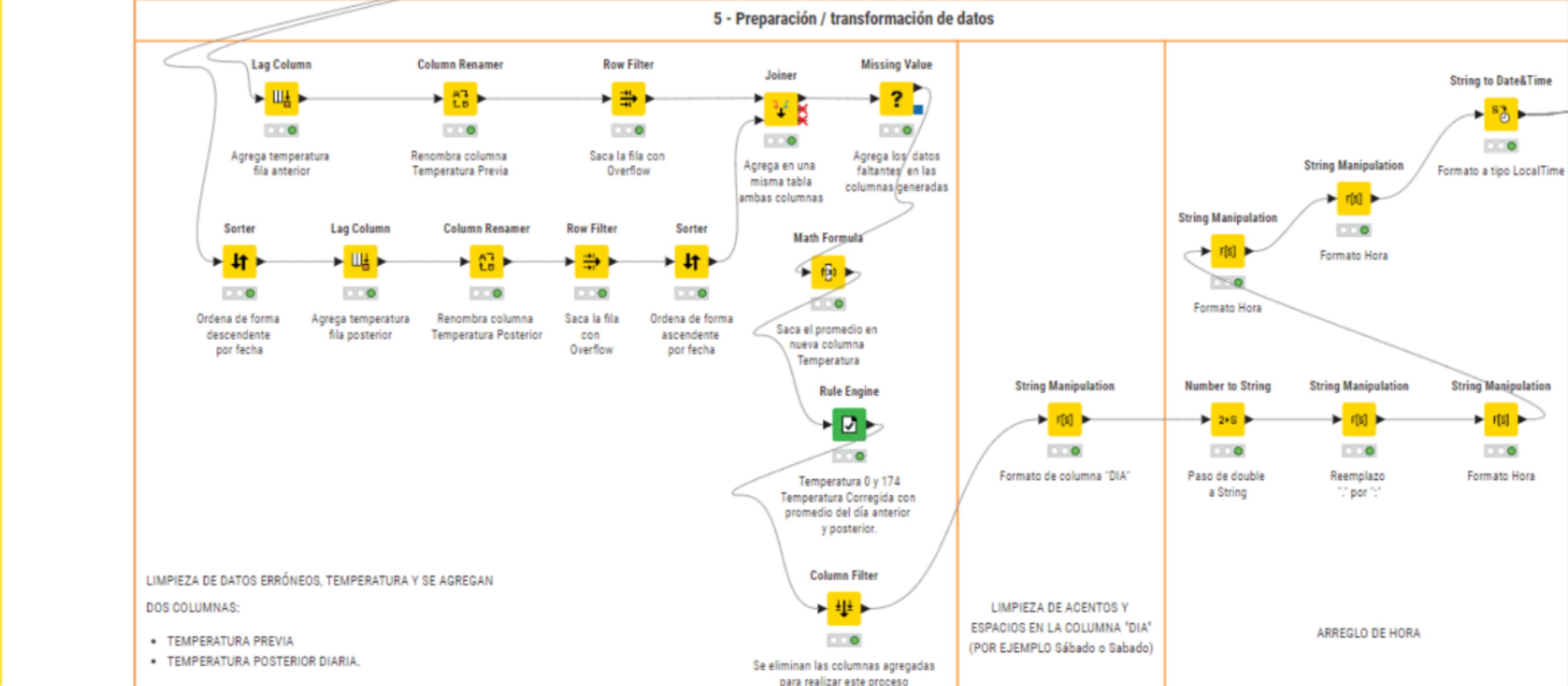
As a result of investigation of period to do next: raise a break-even sales level, increase incomes of direct sales, reduce costs to transportation, strengthen sale divisions, carry out personnel training.

Selección, Preprocesamiento y Transformación de Datos

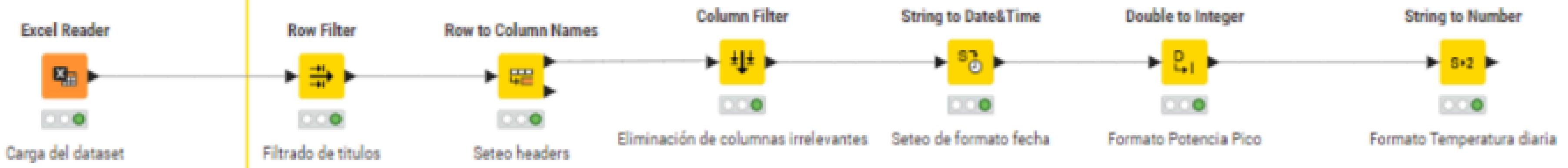
1- Carga del Dataset



2- Limpieza de datos



1- Carga del Dataset



2- Limpieza de datos

Filtrado inicial

- **Row Filter:** Elimina las dos primeras filas por irrelevantes.
- **Column Filter:** Descarta las columna "MES" y la columna nº 14, no útiles para el análisis.

Estructura del dataset

- **Row to Column Names:** Asigna una fila como nombres de columna.

Conversión de datos

- **String to Date&Time:** Convierte la columna "Fecha" a formato de fecha.
- **Double to String:** Redondea "Potencia Pico" a valores enteros.
- **String to Number:** Transforma "Temperatura Media Diaria GBA" a número decimal.

Normalización de temperatura

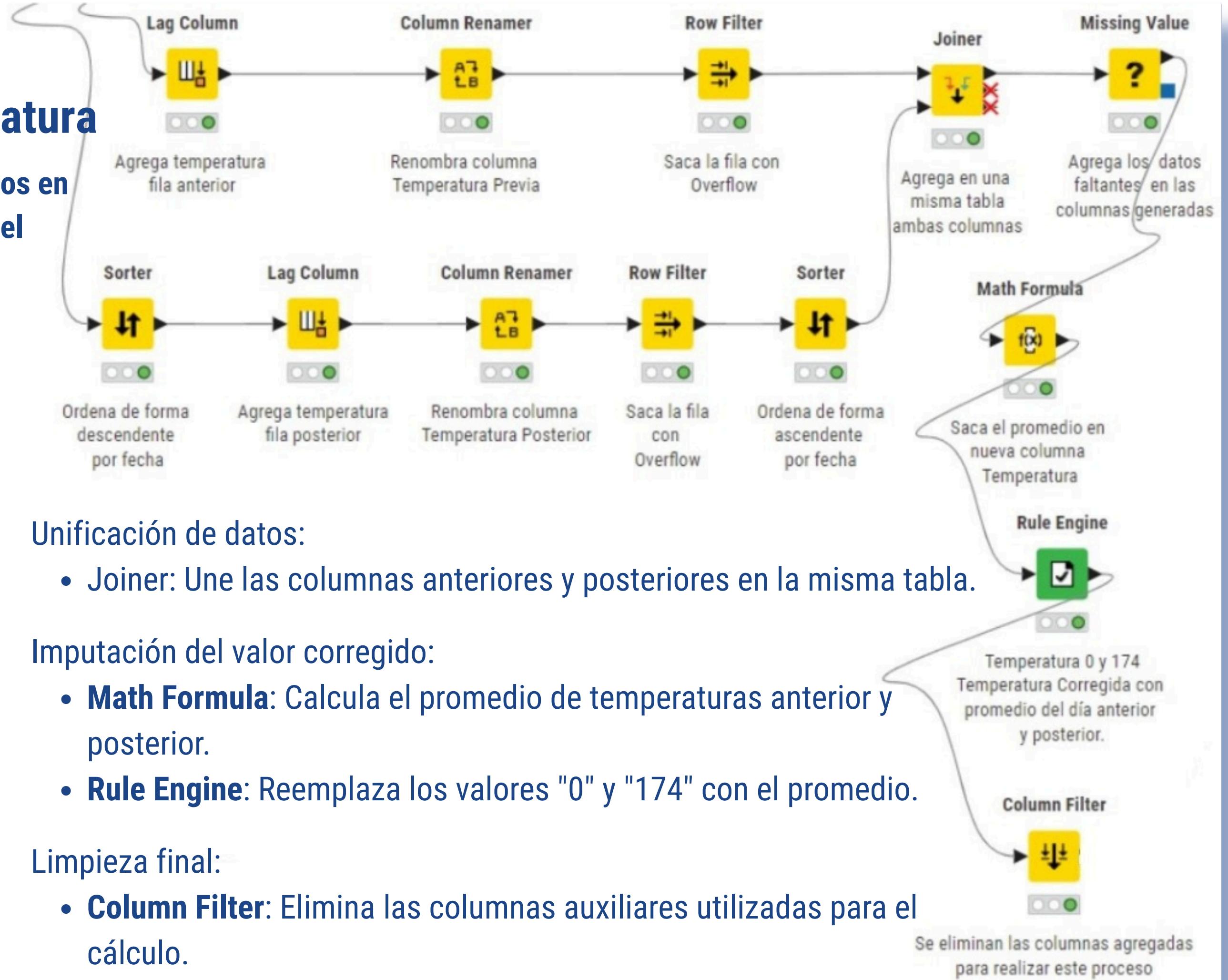
Objetivo: Reemplazar valores anómalos en la columna de temperatura mediante el promedio de días adyacentes.

Generación de columnas auxiliares:

- Se crean columnas con la temperatura del día anterior y posterior (usando **Lag Column**).

Limpieza intermedia:

- **Row Filter:** Se eliminan filas con errores por desbordes de índice.
- **Sorter:** Reordena el dataset para aplicar correctamente los lags.



Limpieza y normalización DIA

- **String Manipulation:** Limpia y normaliza el texto del campo DIA: quita espacios, convierte a minúsculas y elimina tildes.

```
replaceChars(lowerCase(strip($DIA$)), "áéíóúÁÉÍÓÚ", "aeiouAEIOU")
```

Normalización del formato de hora

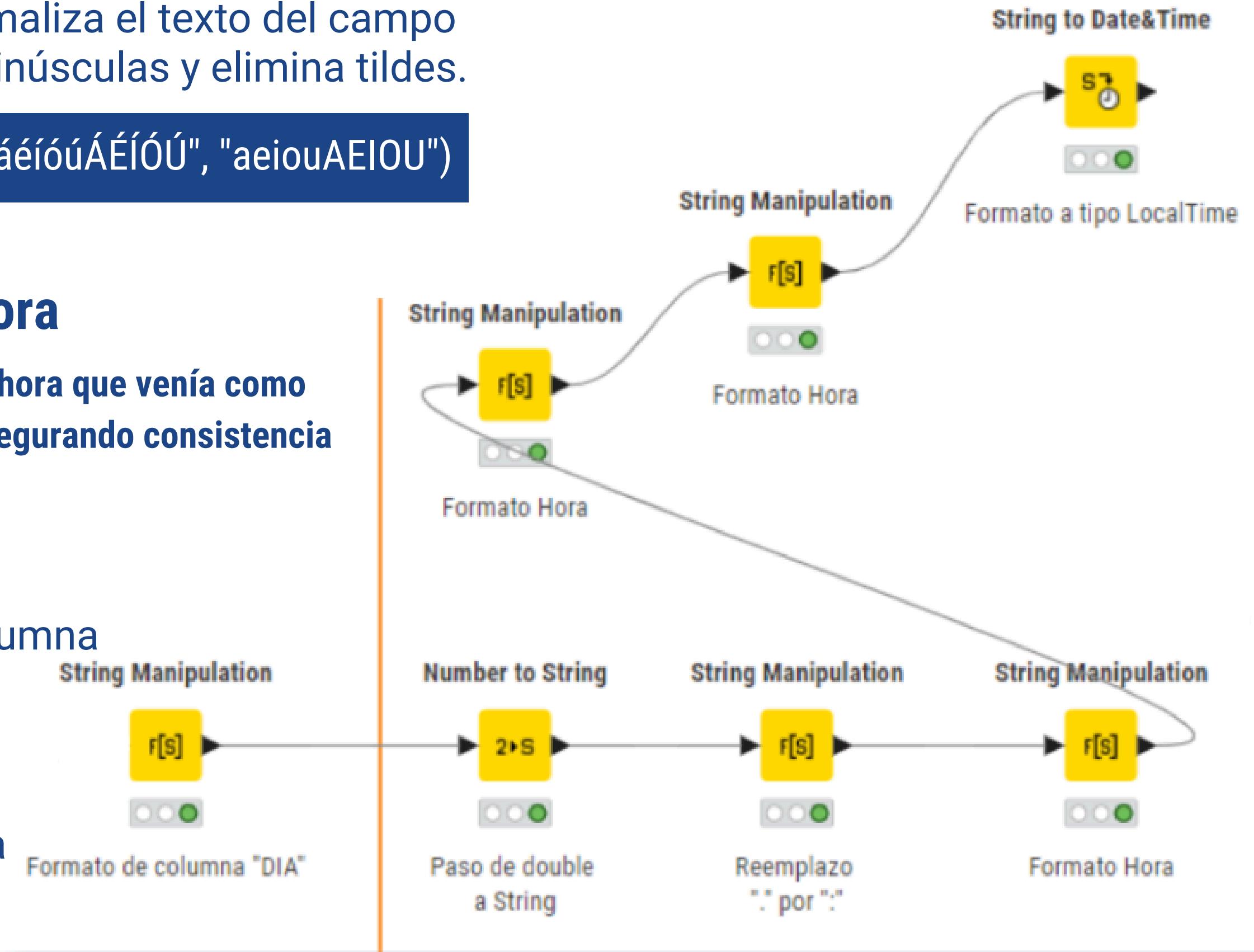
Objetivo: Corregir y estandarizar el formato de hora que venía como número decimal (hh.mm) al formato hh:mm, asegurando consistencia en todos los casos.

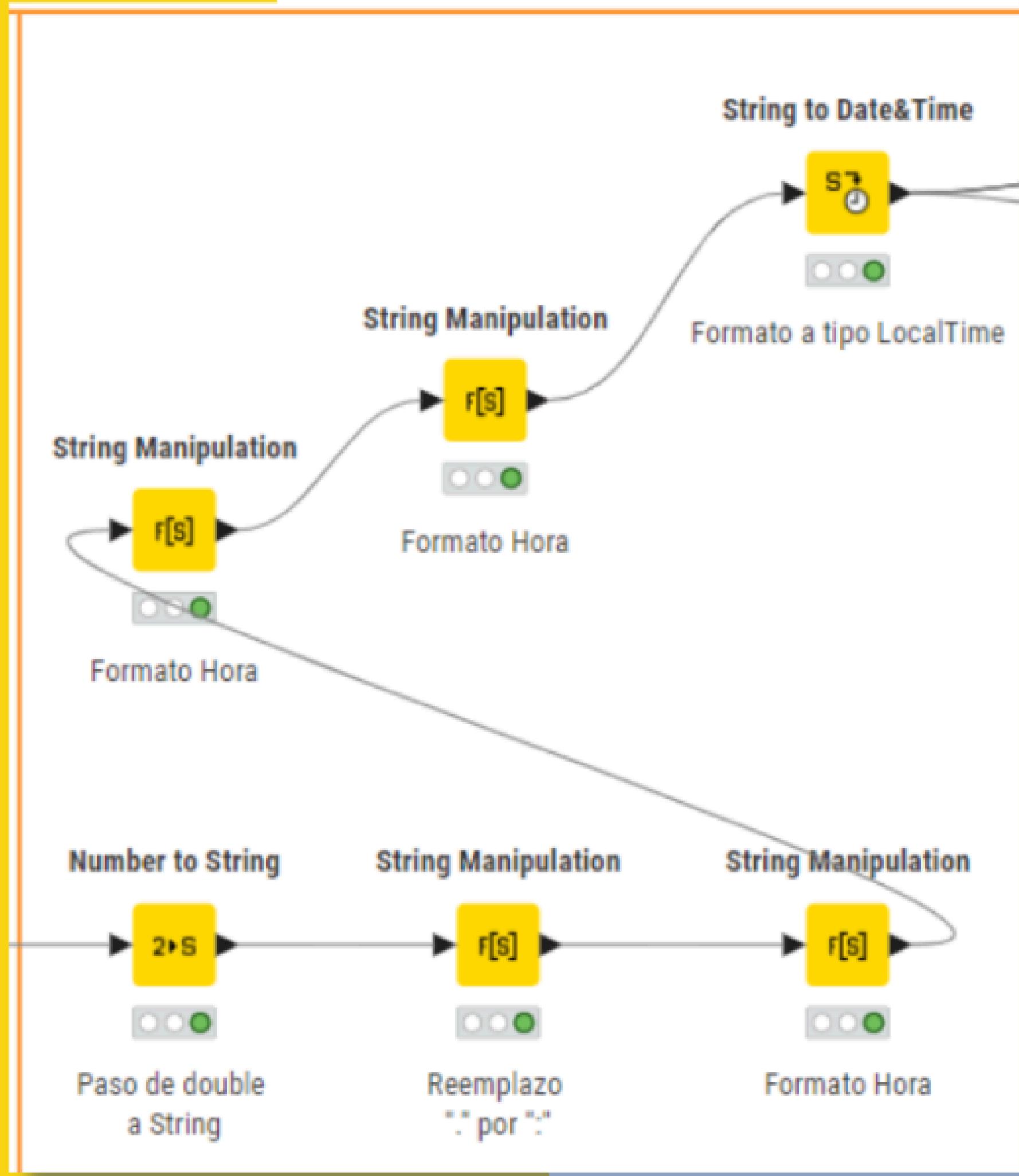
Conversión inicial:

- **Number to String:** Convierte la columna de hora (formato double) a texto.

Reemplazo de separador

- **String Manipulation:** Se reemplaza el punto por dos puntos (. → :).





Corrección de formatos irregulares:

- **String Manipulation:** se corrigen casos como:

12:2 → 12:20 (ejemplo en fila 5891)

9:409 → 09:40 (ejemplo en fila 4737)

15:0 → 15:00 (ejemplo en fila 5919)

Expresiones regulares utilizadas:

- Asegurar dos dígitos en horas y minutos.
- Eliminar decimales extra.
- Añadir ceros cuando faltan.

Conversión final

- **String to Date&Time:** Se convierte a tipo de dato hora local (Local Time).

Modelos Predictivos y Técnicas de Evaluación

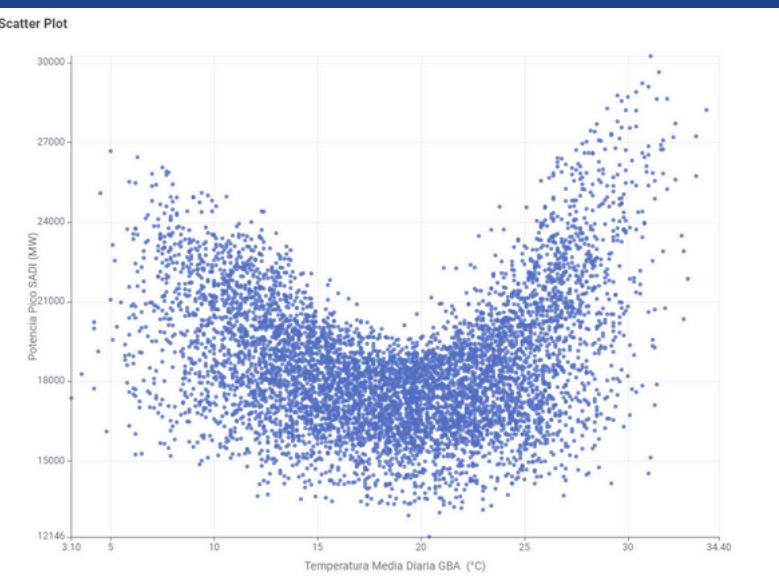
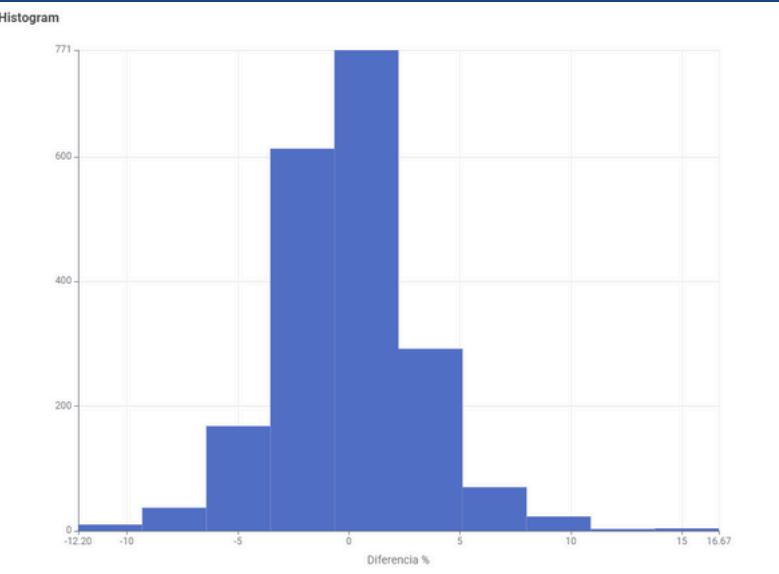
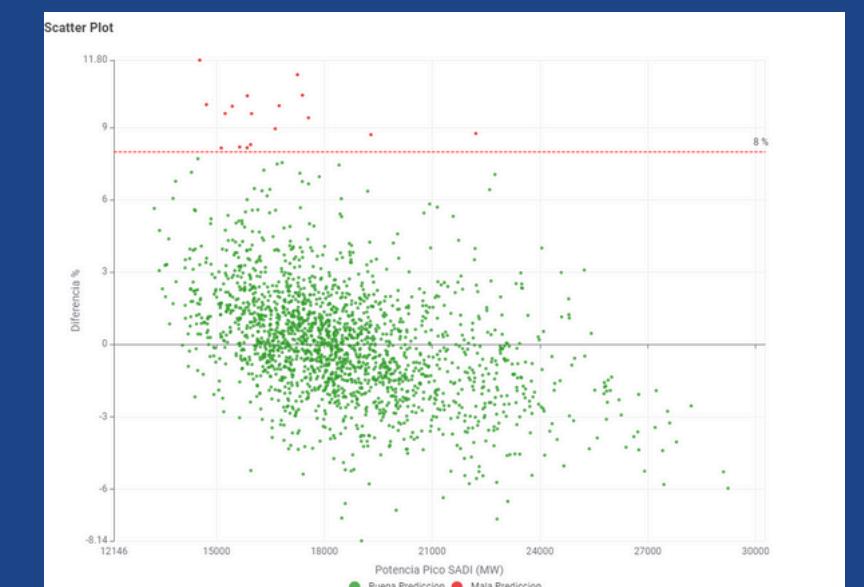
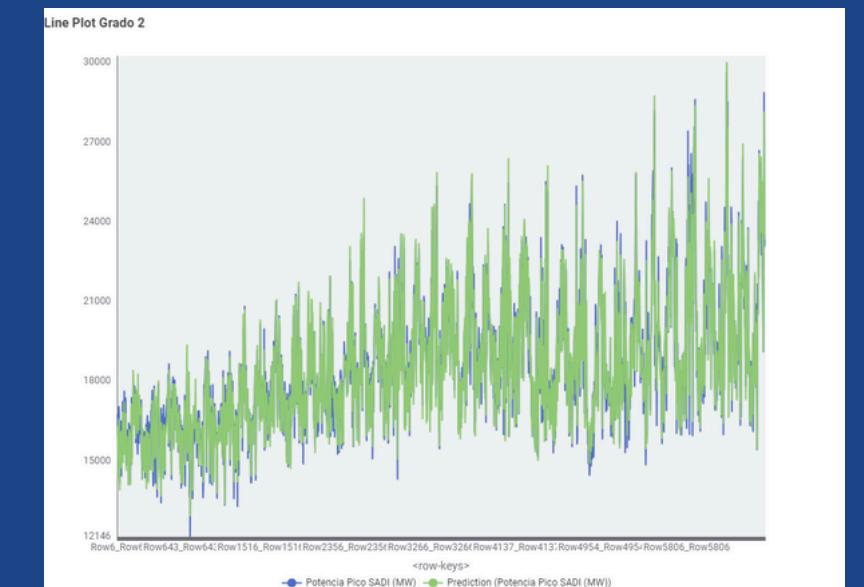
Modelos desarrollados

Modelos de Regresión

- Regresión Lineal
- Regresión Polinómica (2º, 3º, 4º y 10º)
- Árbol de Regresión
- Gradiente Boosted Trees

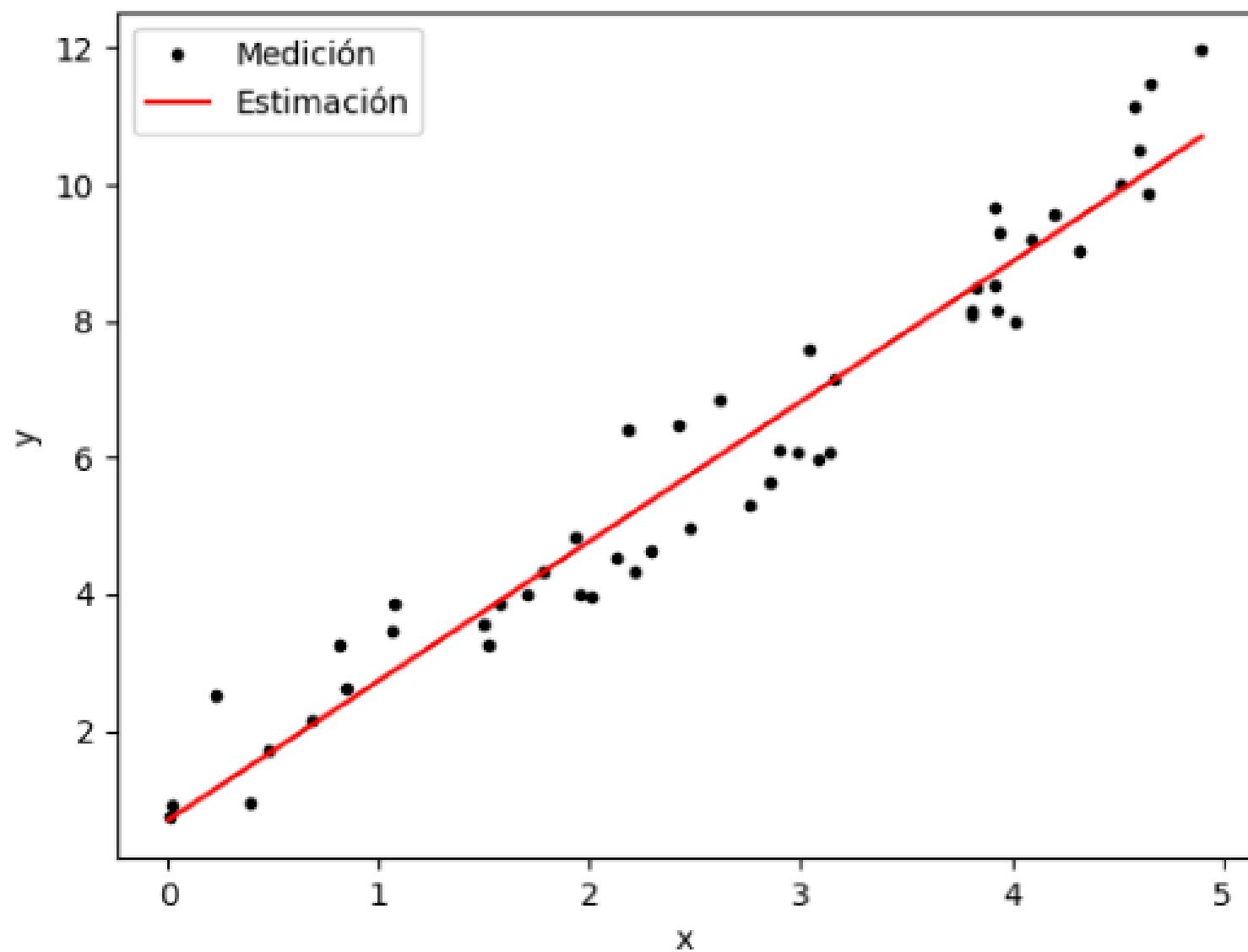
Modelos de Clasificación

- Naive Bayes
- Random Forest



Regresión Lineal Múltiple

Imagen Ilustrativa



Relación múltiple:

- Analiza cómo varias variables independientes influyen sobre una dependiente.

Selección de variables

- Se eligen las más significativas (valor $p < 0.05$), se evita multicolinealidad.

Predictión e interpretación

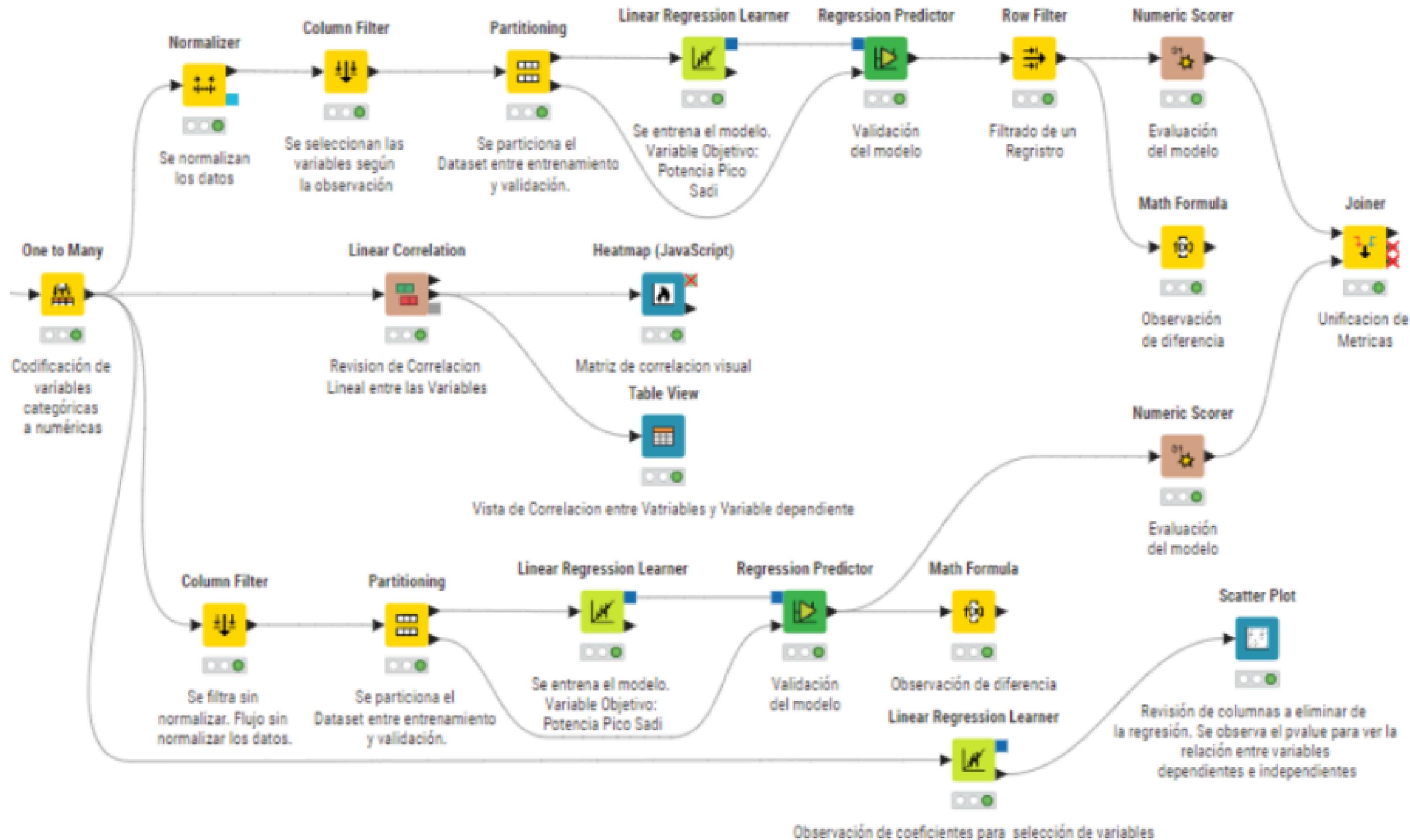
- Los coeficientes indican impacto de cada variable.

Evaluación:

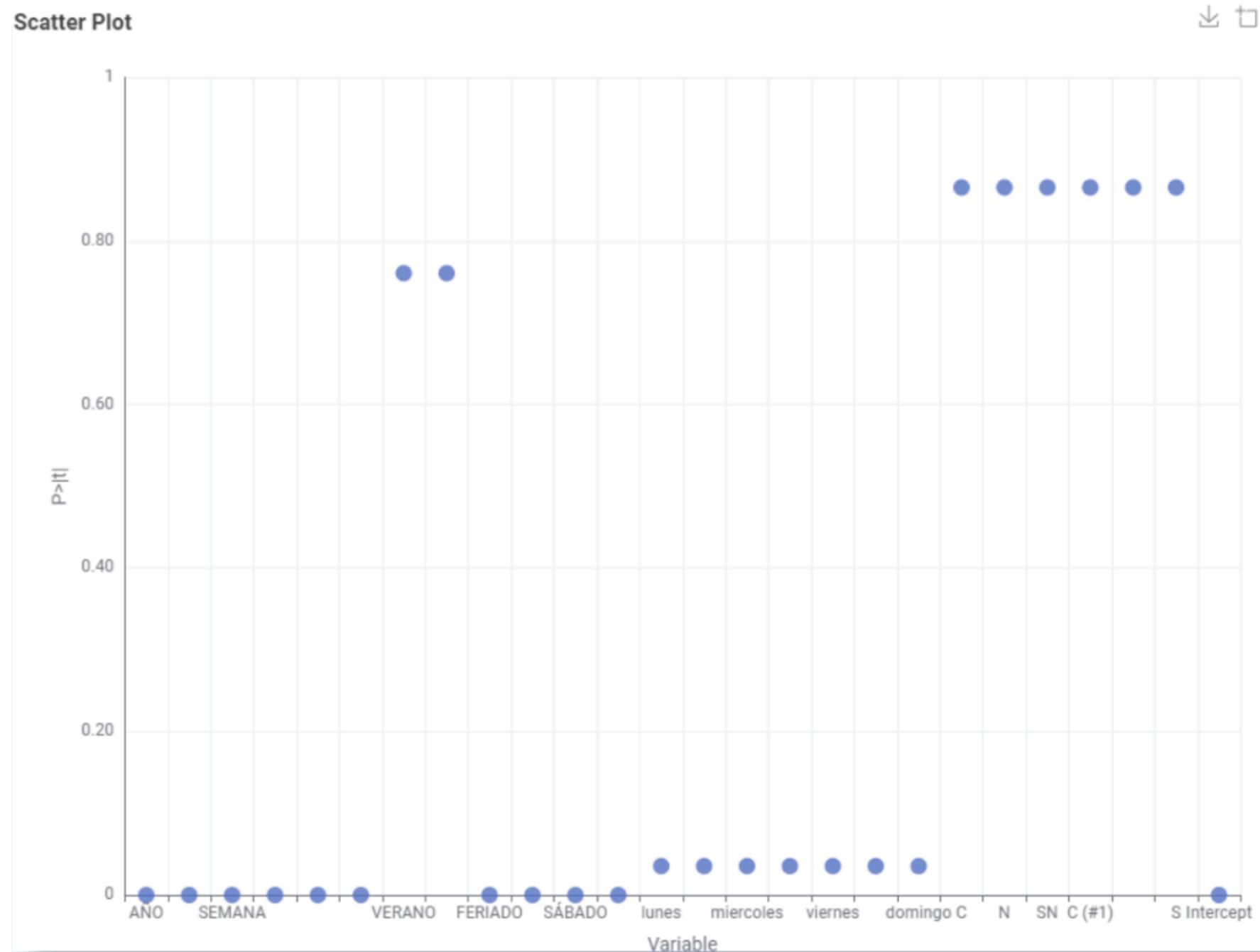
- Se uso R^2 y métricas de error (MAE, RMSE, MAPE).

Regresión Lineal Múltiple

Comparación de Modelo Normalizado y no Normalizado



Entrenamiento del modelo y significancia estadística



Valor p:

Mide la significancia de cada variable en el modelo.

- $p < 0.05 \rightarrow$ variable significativa
- $p \geq 0.05 \rightarrow$ variable no significativa

Importante

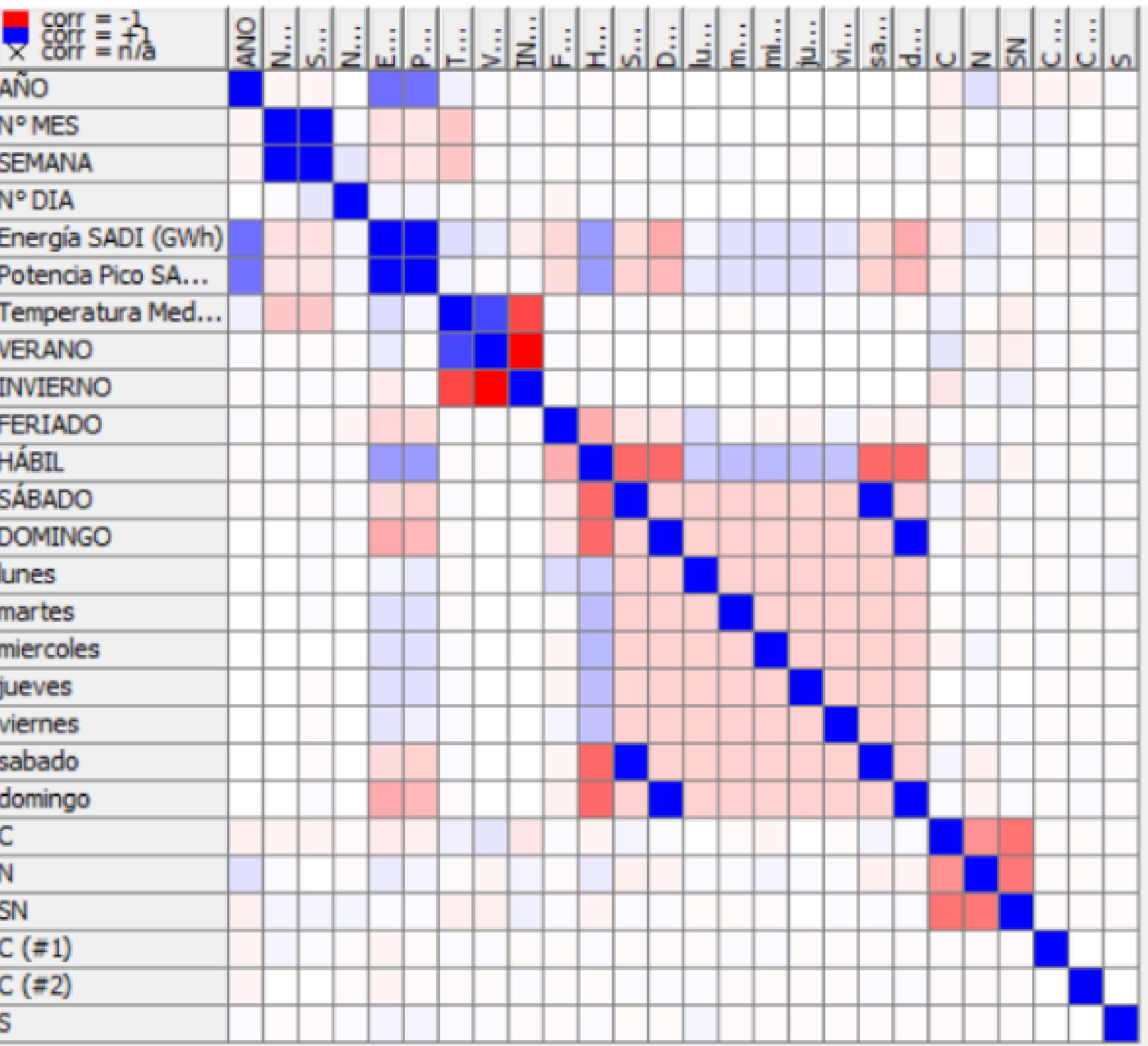
- Complementar con R^2 , análisis de residuos, y detección de multicolinealidad.
- Evalúa la calidad y validez del modelo.

Matriz de Correlación

- Permitió identificar relaciones fuertes entre variables.
- Ayudó a seleccionar variables relevantes y evitar redundancias

Variables con mayor correlación con Potencia Pico SADI

Año, Energía SADI, Hábil



Validación de los Modelos

Con Normalización de Variables

Statistics - ...	
File	
R ² :	0,962
Mean absolute error:	0,019
Mean squared error:	0,001
Root mean squared error:	0,026
Mean signed difference:	-0,001
Mean absolute percentage error:	0,063
Adjusted R ² :	0,962

Sin Normalización de Variables

Statistics - ...	
File	
R ² :	0,964
Mean absolute error:	346,952
Mean squared error:	224.422,319
Root mean squared error:	473,732
Mean signed difference:	-25,069
Mean absolute percentage error:	0,019
Adjusted R ² :	0,964

Poder explicativo

- Ambos modelos con $R^2 > 0.96 \rightarrow$ excelente ajuste.

Precisión absoluta

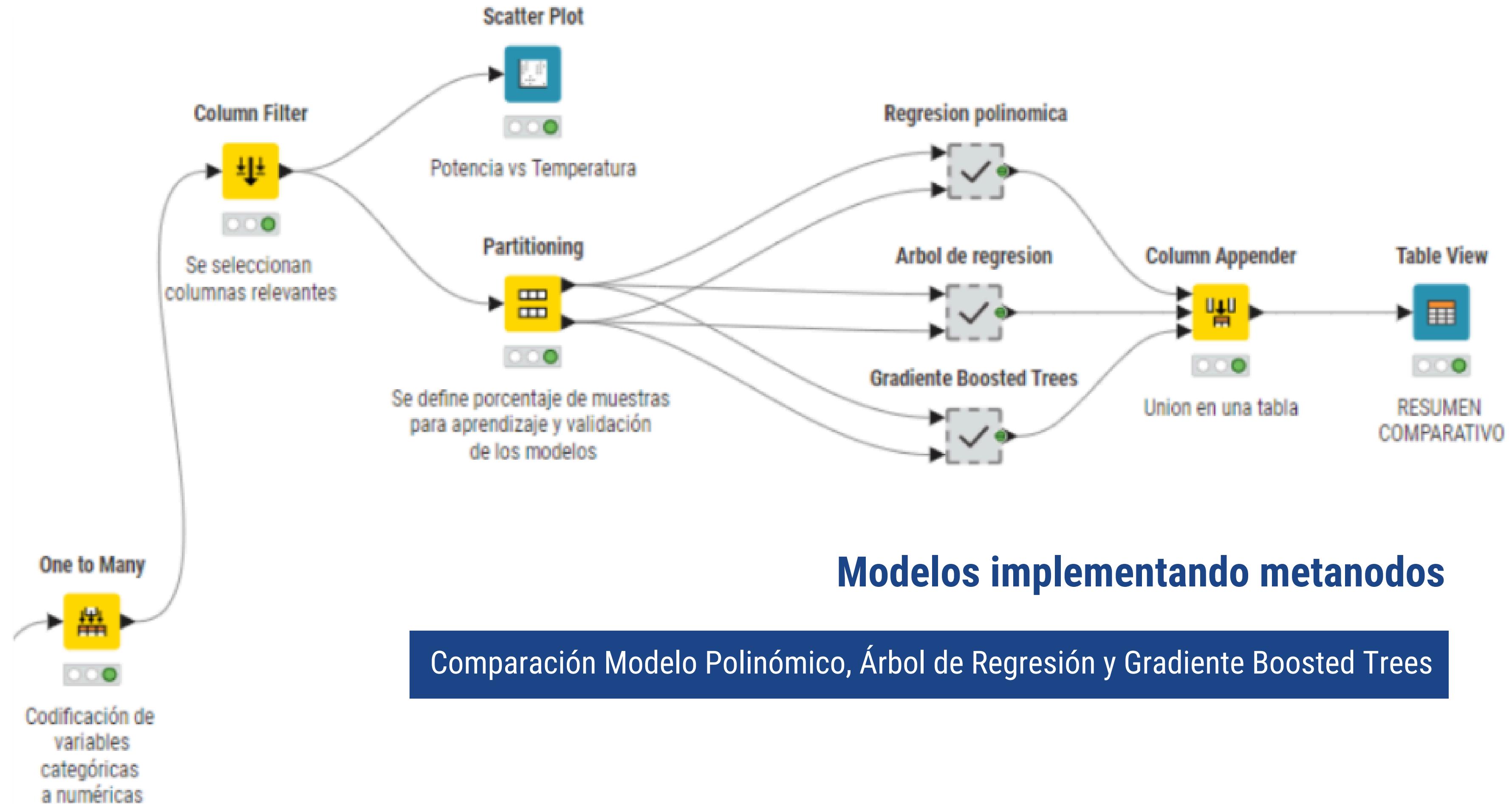
- Modelo normalizado con menor MAE y RMSE.

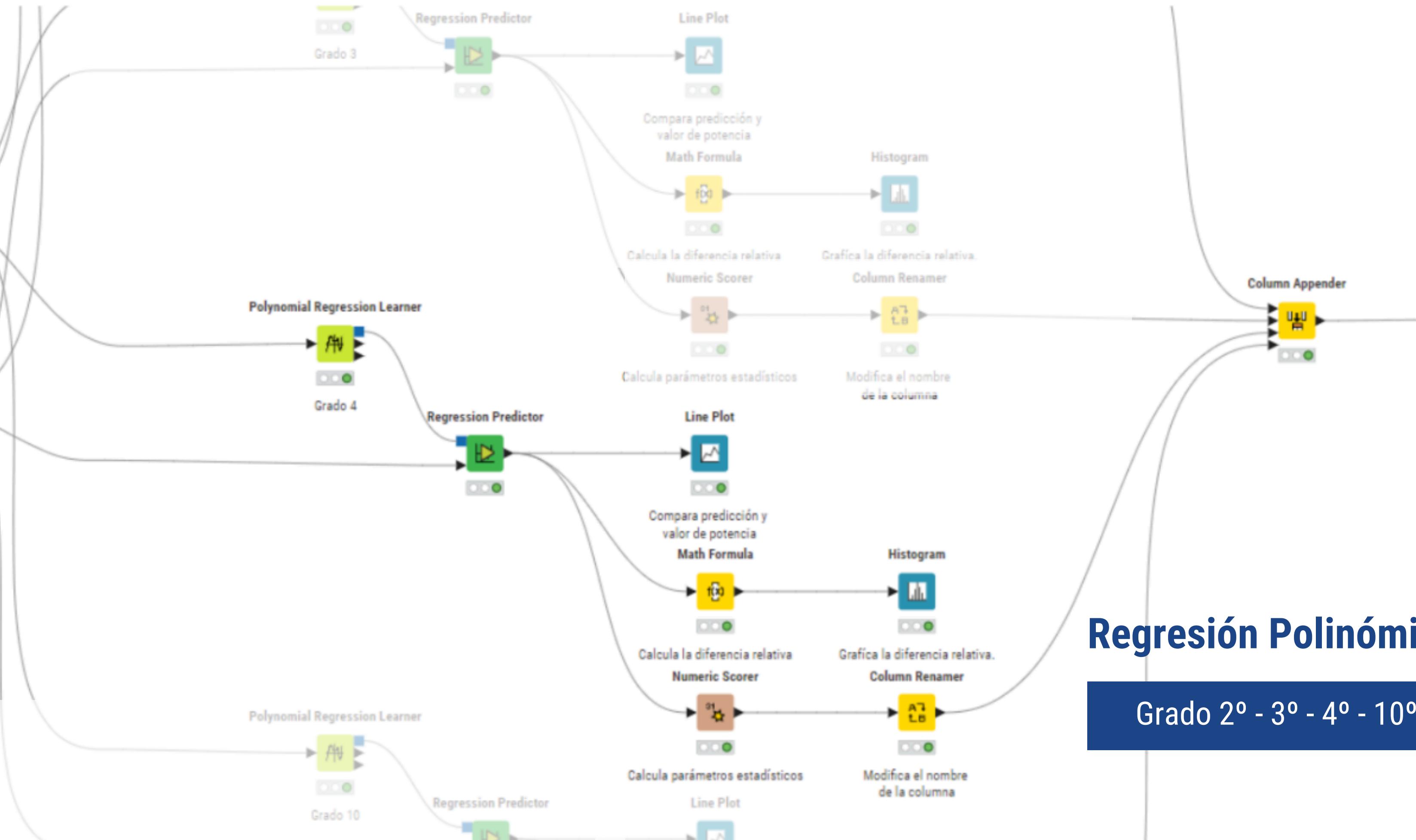
Precisión relativa

- Modelo sin normalizar con menor MAPE.

Conclusión

- Modelo normalizado \rightarrow mejor precisión absoluta.
- Ambos modelos explican bien la demanda.



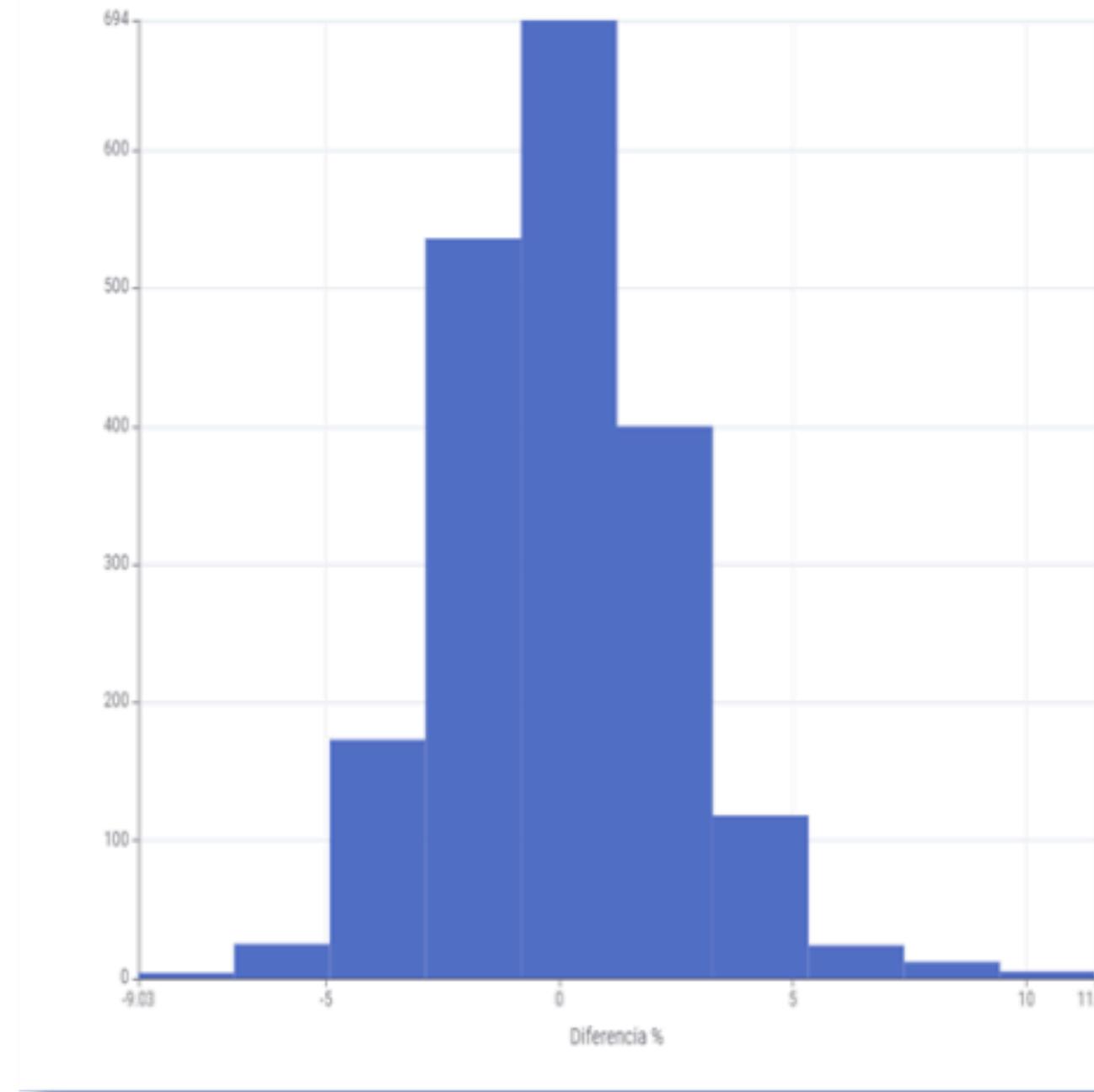


Regresión Polinómica

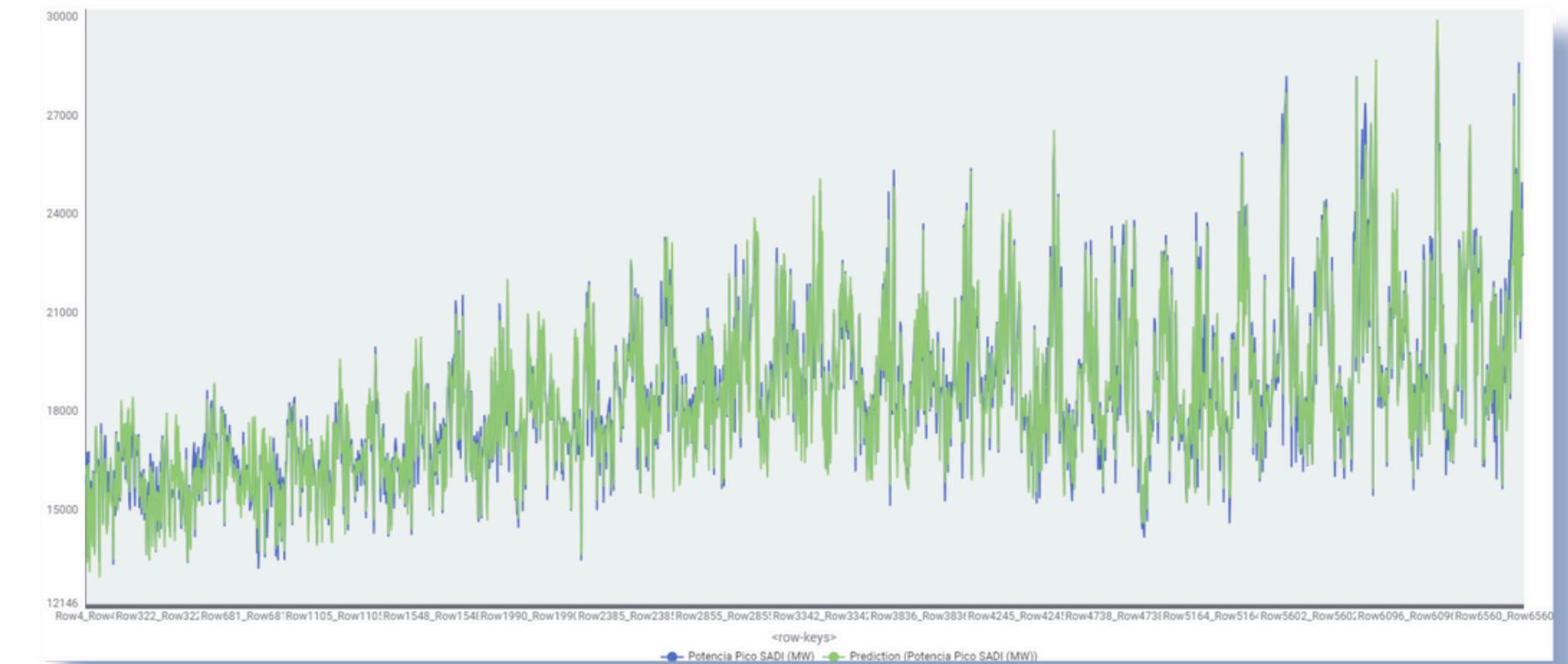
Grado 2º - 3º - 4º - 10º

Regresión polinómica grado 4

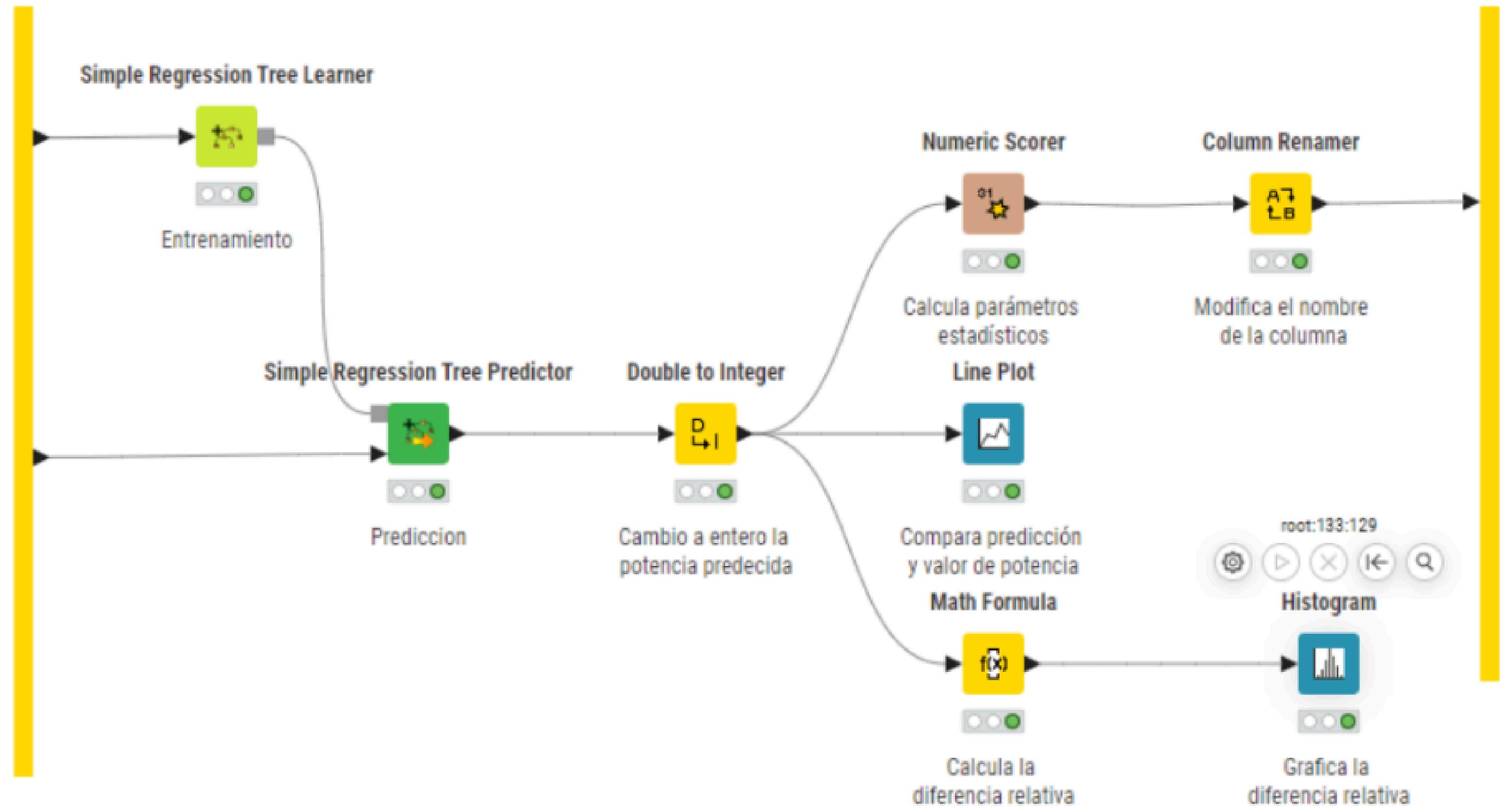
Histogram Grado 4



- Modelos grado 4 y 10 ↓ diferencia relativa.
- Grado 10 muestra riesgo de sobreajuste (overfitting).
- Errores entre -9% y $+11,5\%$.
- Máximas diferencias ocurren en potencias medias, no extremas.

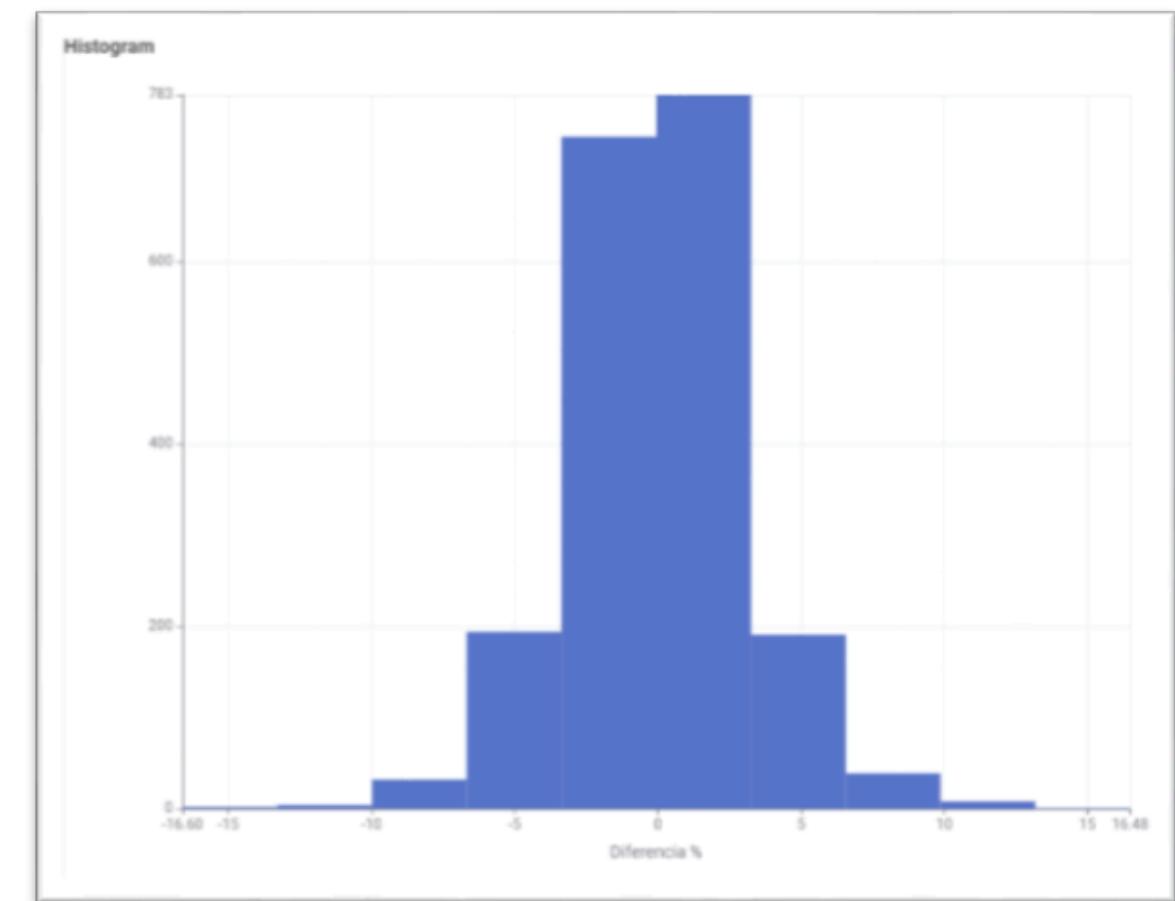


Árbol de Regresión



Árbol de Regresión

- Se limitó la profundidad a 10 → evita overfitting.
- Error relativo entre -7% y +7% → más concentrado que en regresión polinómica.
- Mayor precisión en valores extremos: error ~2% en potencias máximas.

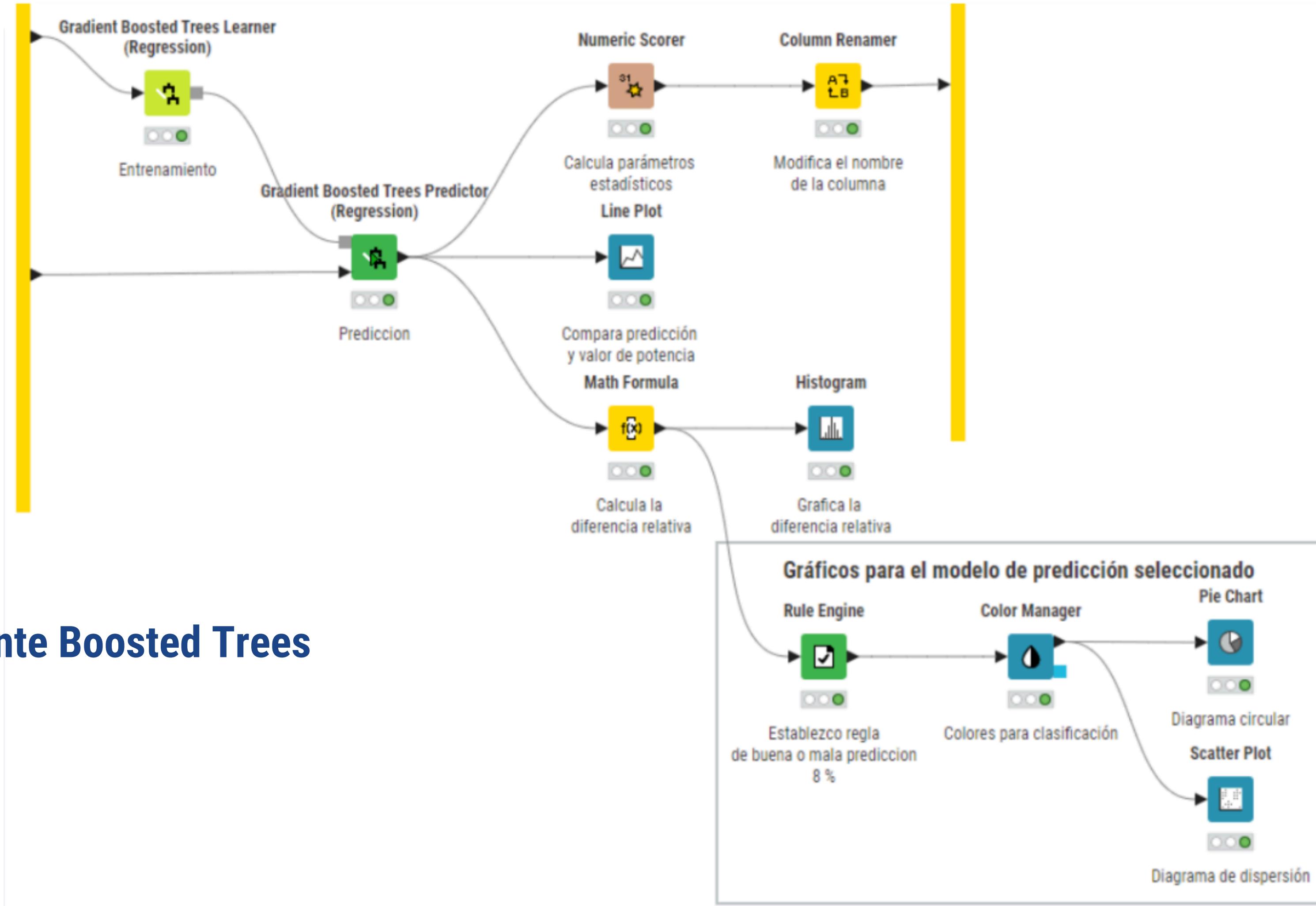


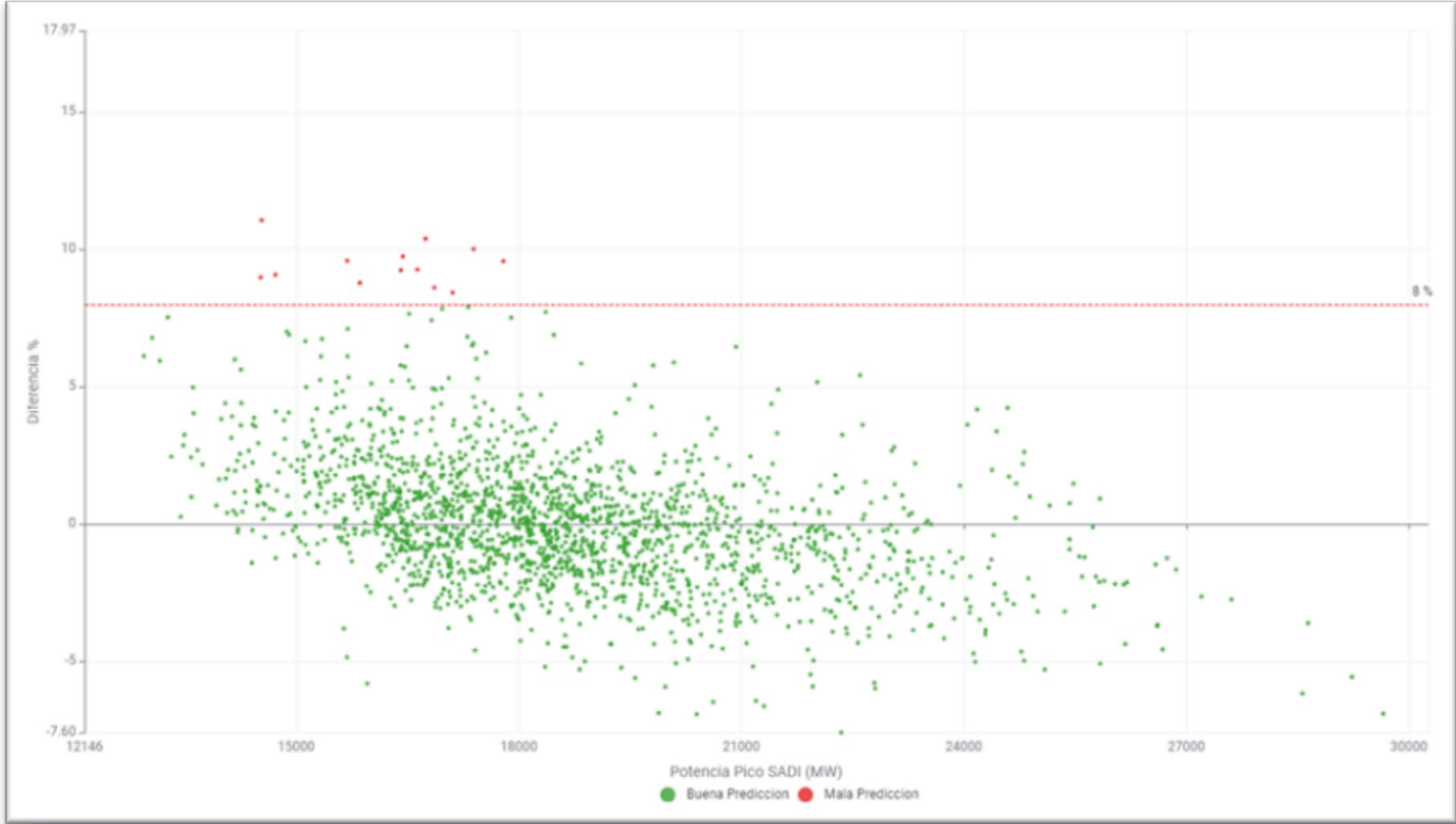
#	RowID	AÑO Number (inte...)	Nº MES Number (inte...)	Energía S... Number (dou...)	Potencia ... Number (inte...)	Temperat... Number (dou...)	VERANO Number (inte...)	INVIERNO Number (inte...)	FERIADO Number (inte...)	HÁBIL Number (inte...)	sabado Number (inte...)	domingo Number (inte...)	Predictio... Number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

#	RowID	AÑO Number (inte...)	Nº 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 (inte...)	Diferencia % Number (double)
110	Row36	2007	12	291.841	15956	27.8	1	0	0	0	0	1	13667	-14.346

#	RowID	AÑO Number (inte...)	Nº MES Number (inte...)	Energía S... Number (dou...)	Poten... 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 (inte...)	Diferencia % 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

Gradiente Boosted Trees





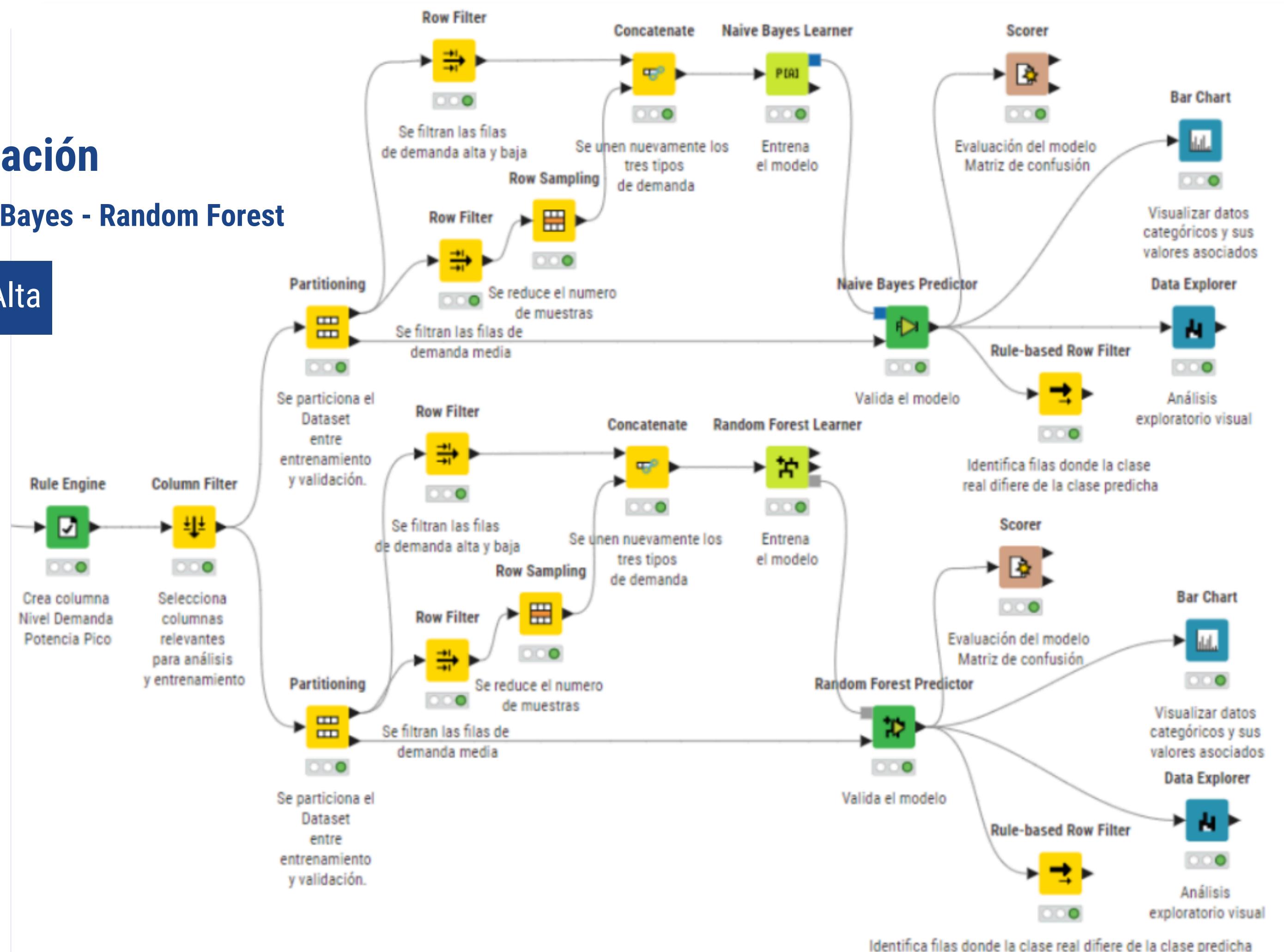
- Error relativo concentrado en $\pm 5\%$ → similar a regresión polinómica.
- Buen equilibrio técnico y económico: evita generación innecesaria.
- Mayor error ($\sim -6.9\%$) en consumos bajos → sin riesgo de saturación.
- Ventaja operativa: permite ajustar despacho con anticipación.
- Apoyo a decisiones de reserva energética o importación si es necesario.

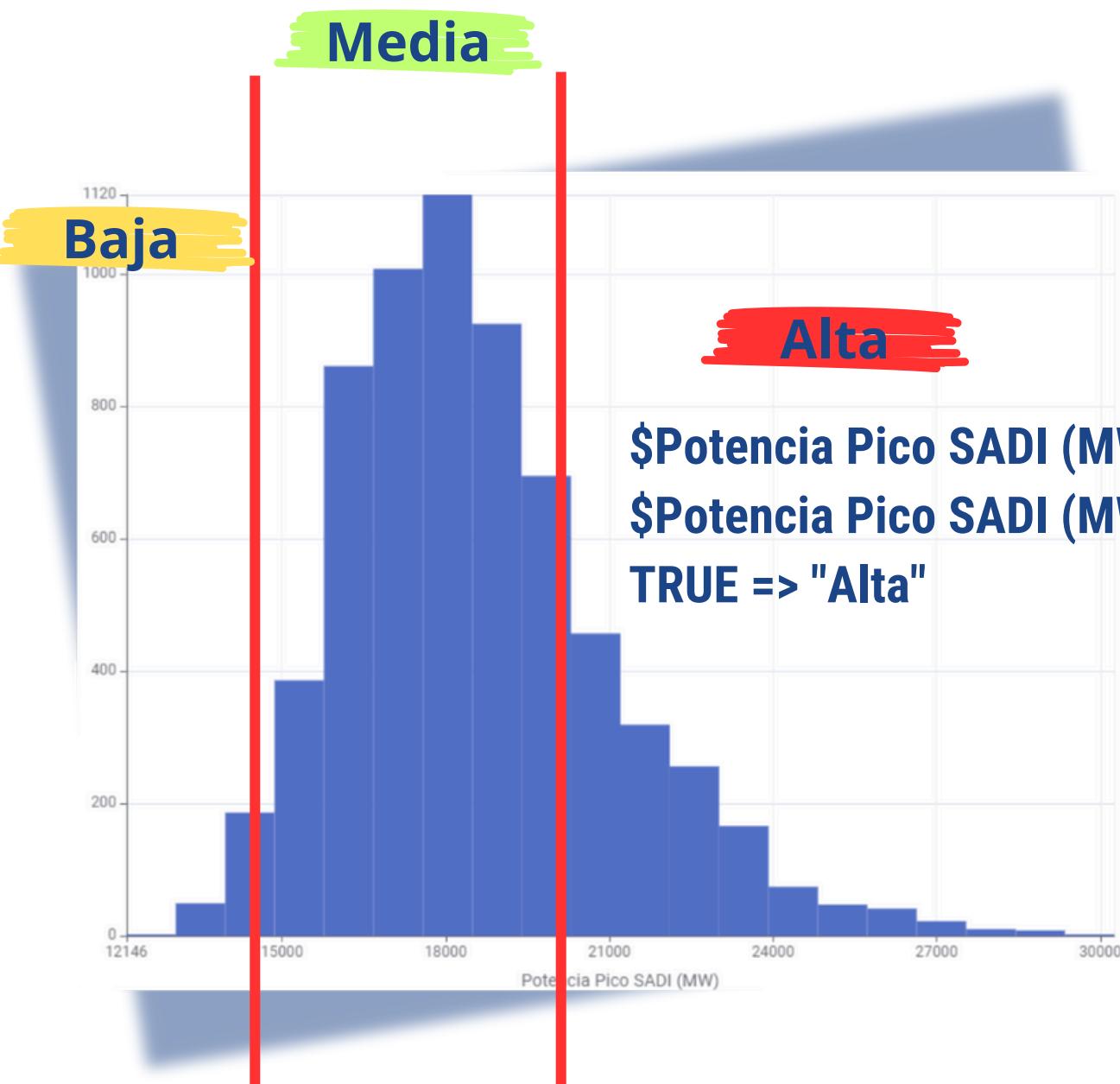
RowID ↓	Regr_Pol_2º - Prediction (Number (double))	Regr_Pol_3º- Prediction (Number (double))	Regr_Pol_4º - Prediction (Number (double))	Regr_Pol_10º - Prediction (Number (double))	Arbol_Regresión - Prediction (Number (double))	Gradient Boosted Trees - Prediction (potencia) (Number (double))
root mean 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

Modelos de Clasificación

Comparación modelos Naive Bayes - Random Forest

Demanda: Baja - Media - Alta

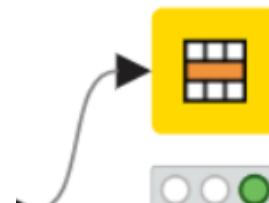




Distribucion de las clases

	Media	Baja	Alta
Entrenamiento	2400; 48.24%	1369; 27.52%	1206; 24.24%
Validación	800; 48.22%	457; 27.55%	402; 24.23%

Row Sampling



Reduce el numero de muestras
de la Demanda Media

1.400; 35.22%

Matriz de Confusión

Random Forest

Nivel Dema...	Baja	Media	Alta
Baja	419	38	0
Media	76	653	71
Alta	0	51	351

Correct classified: 1.423

Accuracy: 85,775%

Cohen's kappa (κ): 0,778%

Wrong classified: 236

Error: 14,225%

Naives Bayes

Nivel Dema...	Baja	Media	Alta
Baja	376	69	12
Media	164	460	176
Alta	16	93	293

Correct classified: 1.129

Accuracy: 68,053%

Cohen's kappa (κ): 0,513%

Wrong classified: 530

Error: 31,947%

- La Clasificación es un herramienta operacional útil para decisiones rápidas de despacho.
- Mejor capacidad de generalización sin necesidad de utilizar SMOTE, optándose por una reducción mediante muestreo.

01

Regresión Lineal: $R^2 > 0,96$. La normalización mejora MAE y RMSE, pero sin normalizar da menor MAPE.

04

Complejidad vs Beneficio:
Modelos más complejos (grado >4) no mejoran métricas. Grado 4 es suficiente.

06

Variables clave:
Temperatura media diaria,
Energía diaria SADI, Tipo de día,
Estación, Semana/Año.

02

Modelos complejos:
Polinómica grado 4 y Gradient Boosted Trees (GBT) logran buen balance sesgo-varianza.
GBT fue el mejor modelo: mayor R^2 y menores errores (MAE, RMSE, MAPE), con error máximo de -6,91%.

05

Impacto operativo:
GBT predice demanda con error $<\pm 7\%$, optimizando el despacho y reduciendo costos.

03

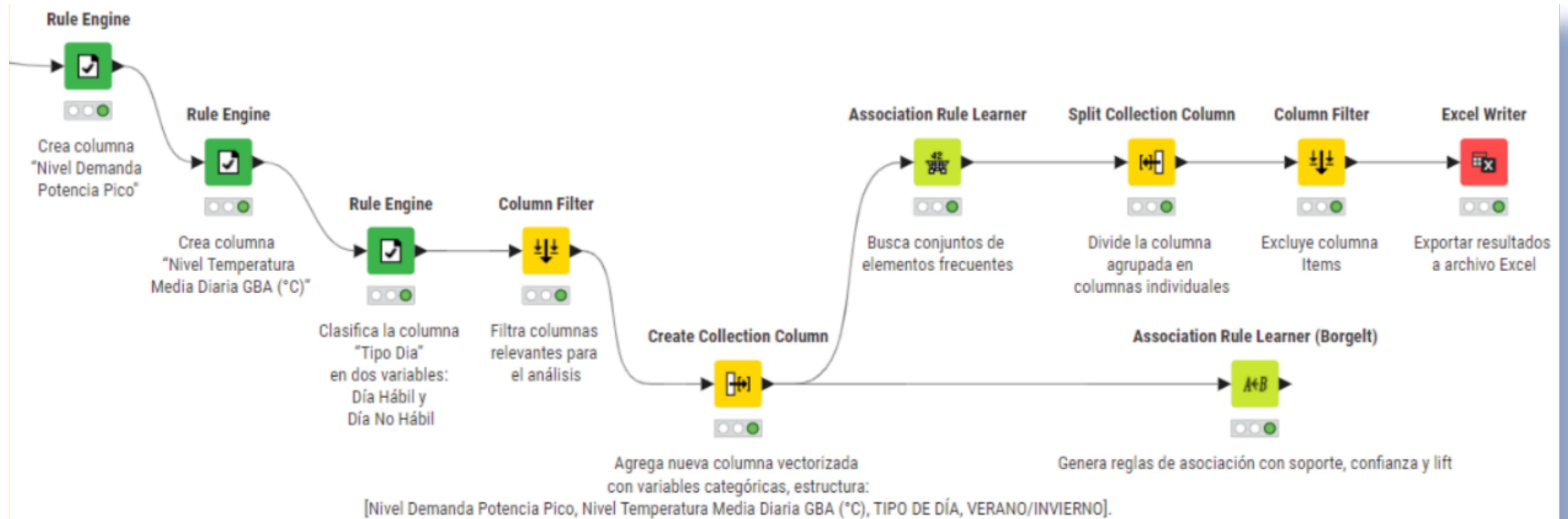
Regresión vs Clasificación:
Regresión: ideal para estimar potencia pico de forma continua.
Clasificación: útil para decisiones rápidas; Random Forest superó a Naive Bayes (85,8% vs. 66,9%).

Conclusiones de los Modelos Predictivos

Gradient Boosted Trees es el modelo más preciso y confiable para predecir la demanda energética del SADI.

Modelos Descriptivos y Técnicas de Evaluación

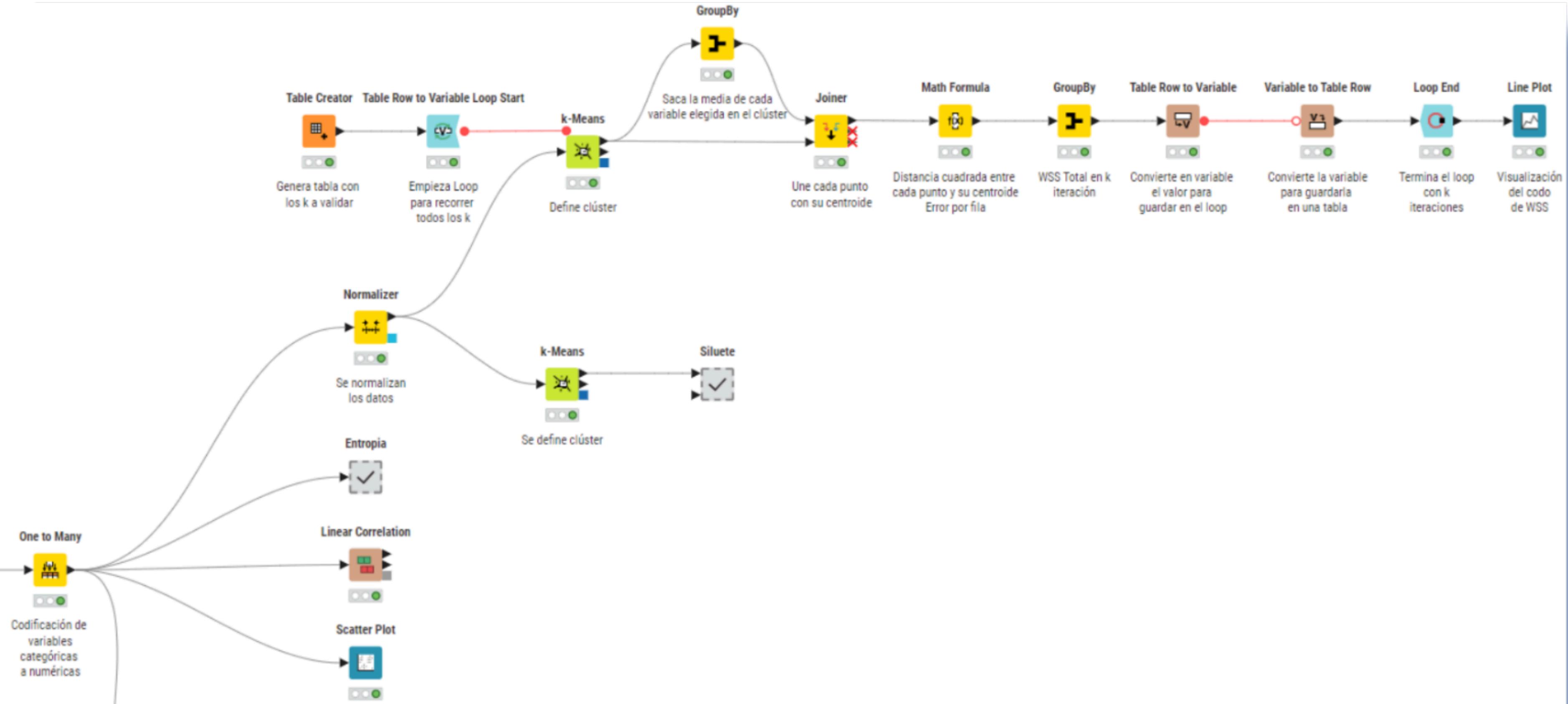
Reglas de Asociación



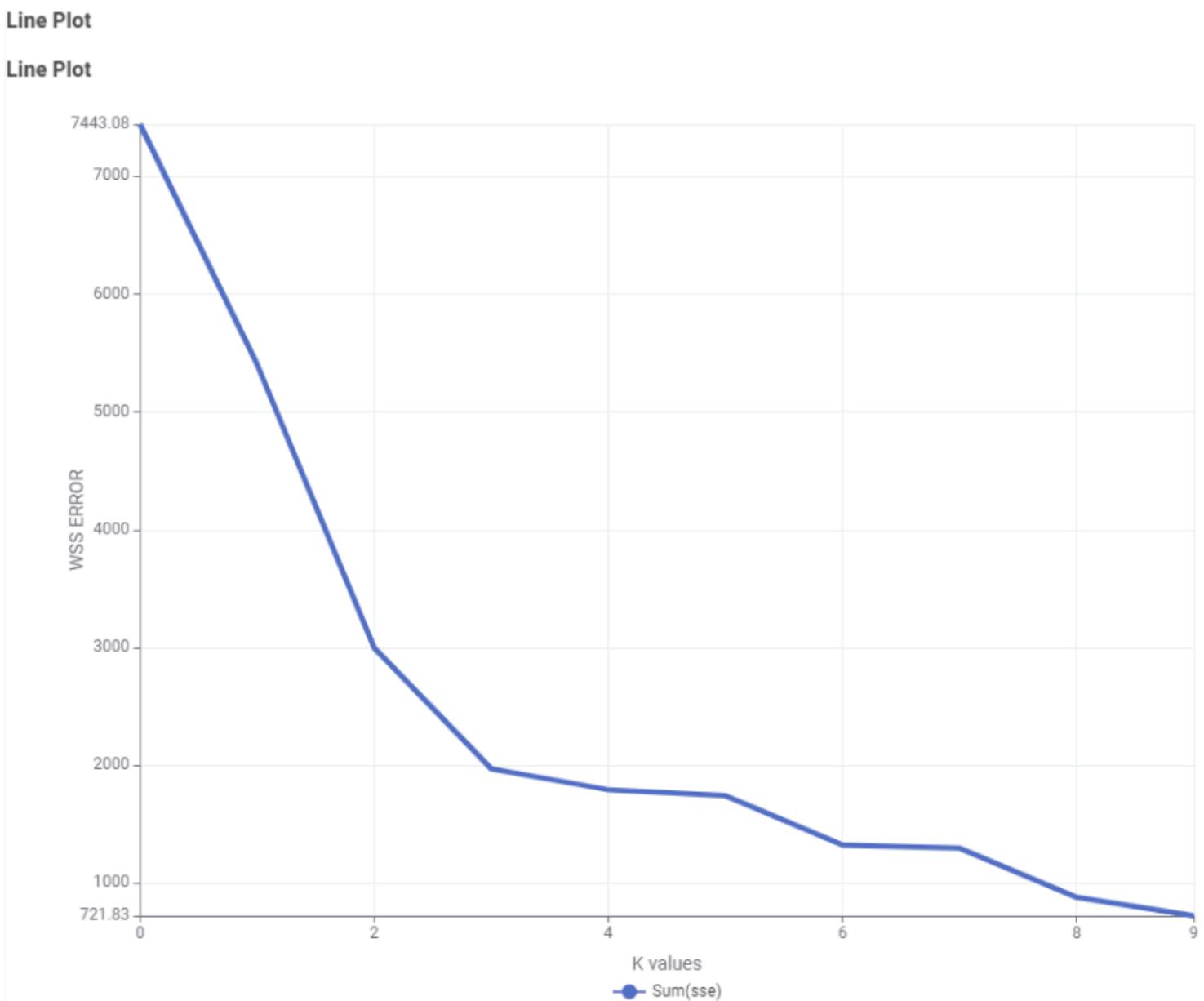
Consequent ↑ String	Antecedent List	ItemSetSupport Number (integer)	RuleConfidence% Number (double)	RuleLift Number (double)
1 selected	Antecedent	ItemSetSupport	RuleConfidence%	RuleLift
Potencia Pico Alta	[HÁBIL,Temperatura Alta,VERANO]	494	67.8	2.796
Potencia Pico Alta	[Temperatura Alta,VERANO]	622	58.7	2.421
Potencia Pico Alta	[HÁBIL,Temperatura Alta]	499	67.3	2.775
Potencia Pico Alta	[Temperatura Alta]	627	58.1	2.397
Potencia Pico Alta	[HÁBIL,INVIERNO,Temperatura Baja]	649	50.4	2.08
Potencia Pico Alta	[INVIERNO,Temperatura Baja]	758	40	1.65

- Filtro en función del **Consecuente**, objetivo principal del análisis → identificar condiciones bajo las cuales se produce una Potencia Pico Alta, Medio o Baja.
- Foco especial en **Potencia Pico Alta** → posibles consecuencias operativas (sobrecarga del sistema y potenciales cortes del suministro).
- Regla relevante, combinación:
 - día Habil - Temperatura Alta - Verano
- Permite predecir potencia Pico Alta con un:
 - **67,8% de confianza.**
 - un **Lift del 2.796** → esta relación ocurre casi 3 veces más de lo que se esperaría si las variables fueran independientes.
 - Presenta un **Soporte del 7,45%** (494 ocurrencias sobre el total de días analizados) → se trata de un patrón con una frecuencia razonable.

Clustering



Tecnica del Codo



¿Cómo obtenemos el número adecuado de k?

- Calculamos el WSS (Within-Cluster Sum of Squares) total para cada valor de k dentro del rango [1;10].
- Se busca minimizar el error cuadrático.
- Se busca equilibrio entre precisión y simplicidad. (Mejor agrupamiento pero mayor interpretabilidad).

Pasos:

1. Obtener la media de cada atributo por cluster
2. Se asigna cada punto a su centroide
3. Calcular la distancia cuadrada entre cada punto y su centroide.
4. Sumatoria de errores para calcular el WSS según K valor.
5. Se grafica los valores de WSS obtenidos según el valor de k.

4 Clusters	
cluster_0 (coverage: 1092)	<ul style="list-style-type: none"> ● AÑO = 0.4855514855514808 ● Energía SADI (GWh) = 0.3005810620823225 ● Potencia Pico SADI (MW) = 0.27720039609222996 ● Temperatura Media Diaria GBA (°C) = 0.6261527343795721 ● VERANO = 1.0 ● INVIERNO = 0.0 ● FERIADO = 0.1446886446886447 ● HÁBIL = 0.0 ● SÁBADO = 0.42765567765567764 ● DOMINGO = 0.42765567765567764
cluster_1 (coverage: 2248)	<ul style="list-style-type: none"> ● AÑO = 0.4796609331751647 ● Energía SADI (GWh) = 0.4305776573269734 ● Potencia Pico SADI (MW) = 0.3913344478523207 ● Temperatura Media Diaria GBA (°C) = 0.6317217718554218 ● VERANO = 1.0 ● INVIERNO = 0.0 ● FERIADO = 0.0 ● HÁBIL = 1.0 ● SÁBADO = 0.0 ● DOMINGO = 0.0
cluster_2 (coverage: 1071)	<ul style="list-style-type: none"> ● AÑO = 0.4732600456215867 ● Energía SADI (GWh) = 0.2808192384515555 ● Potencia Pico SADI (MW) = 0.28658898371774877 ● Temperatura Media Diaria GBA (°C) = 0.369399241587292 ● VERANO = 0.0 ● INVIERNO = 1.0 ● FERIADO = 0.13271028037383178 ● HÁBIL = 0.0 ● SÁBADO = 0.43457943925233644 ● DOMINGO = 0.4327102803738318
cluster_3 (coverage: 2224)	<ul style="list-style-type: none"> ● AÑO = 0.471722621902474 ● Energía SADI (GWh) = 0.4045685215390509 ● Potencia Pico SADI (MW) = 0.3914517797721403 ● Temperatura Media Diaria GBA (°C) = 0.3632820580596226 ● VERANO = 0.0 ● INVIERNO = 1.0 ● FERIADO = 0.0 ● HÁBIL = 1.0 ● SÁBADO = 0.0 ● DOMINGO = 0.0

Descripción de los clústeres

Clúster 0: días no laborables donde el consumo eléctrico sube mayormente por la refrigeración en hogares y comercios (gastronomía, centros comerciales, etc)

- Consumo energía: Medio/Alto - Temperatura: Alta - Días no hábiles

Clúster 1: contempla días hábiles con altas temperaturas donde el alto consumo eléctrico se da por la actividad en oficinas, comercios, industrias y hogares.

- Consumo energía: Alto - Temperatura: Alta - Días hábiles

Clúster 2: representa días no laborales con baja temperatura y demanda de energía. Donde la calefacción en hogares no siempre es eléctrica.

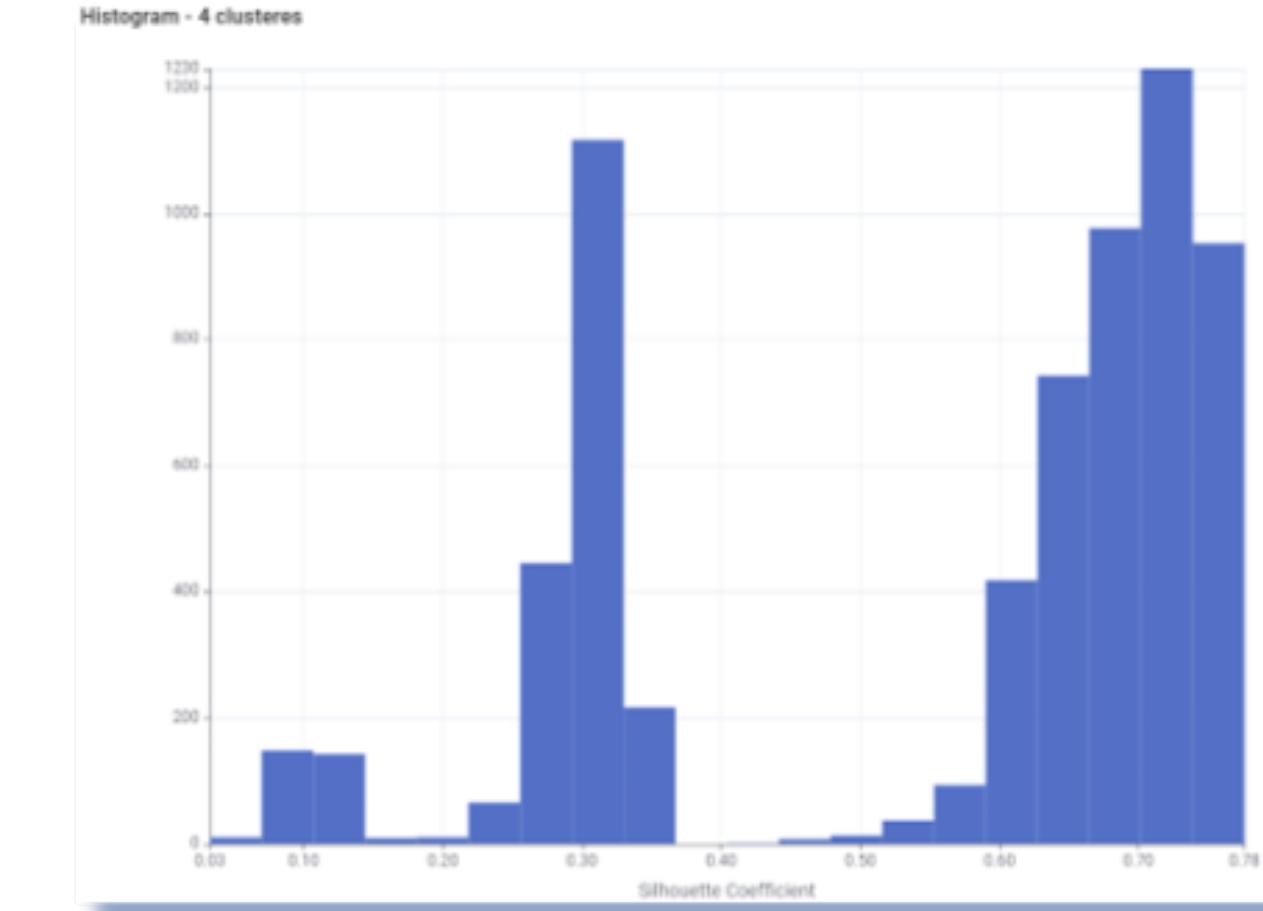
- Consumo energía: Bajo - Temperatura: Baja - Días no hábiles

Clúster 3: días laborales de baja temperatura donde el alto consumo de energía está dado por la actividad en industrias, comercios, oficinas y hogares en menor medida.

- Consumo energía: Alto - Temperatura: Baja - Días hábiles

Coeficiente de Silhouette

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



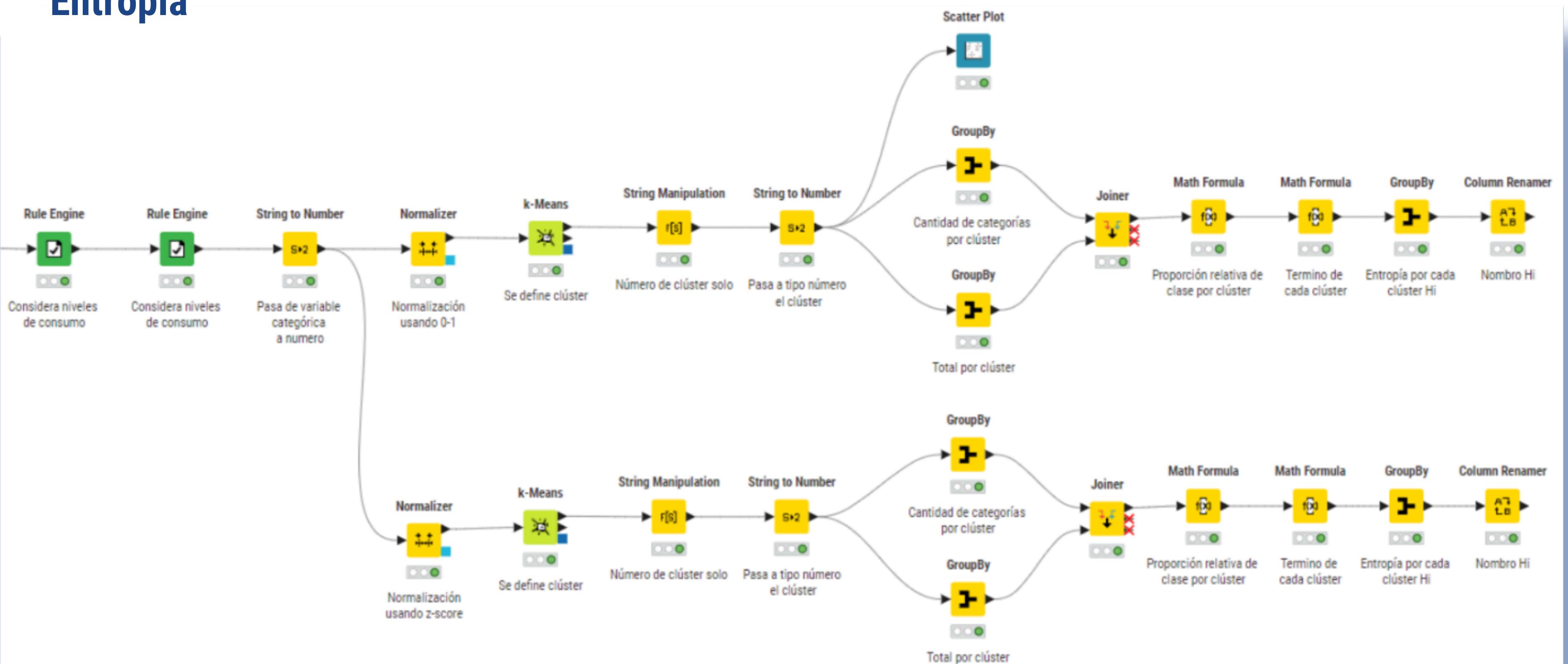
- A mayor cantidad de clústeres → coeficientes negativos
- Indica sobreajuste y pérdida de cohesión.

Dilema

¿Simplificar con menos clústeres o complejizar arriesgando calidad?



Entropía

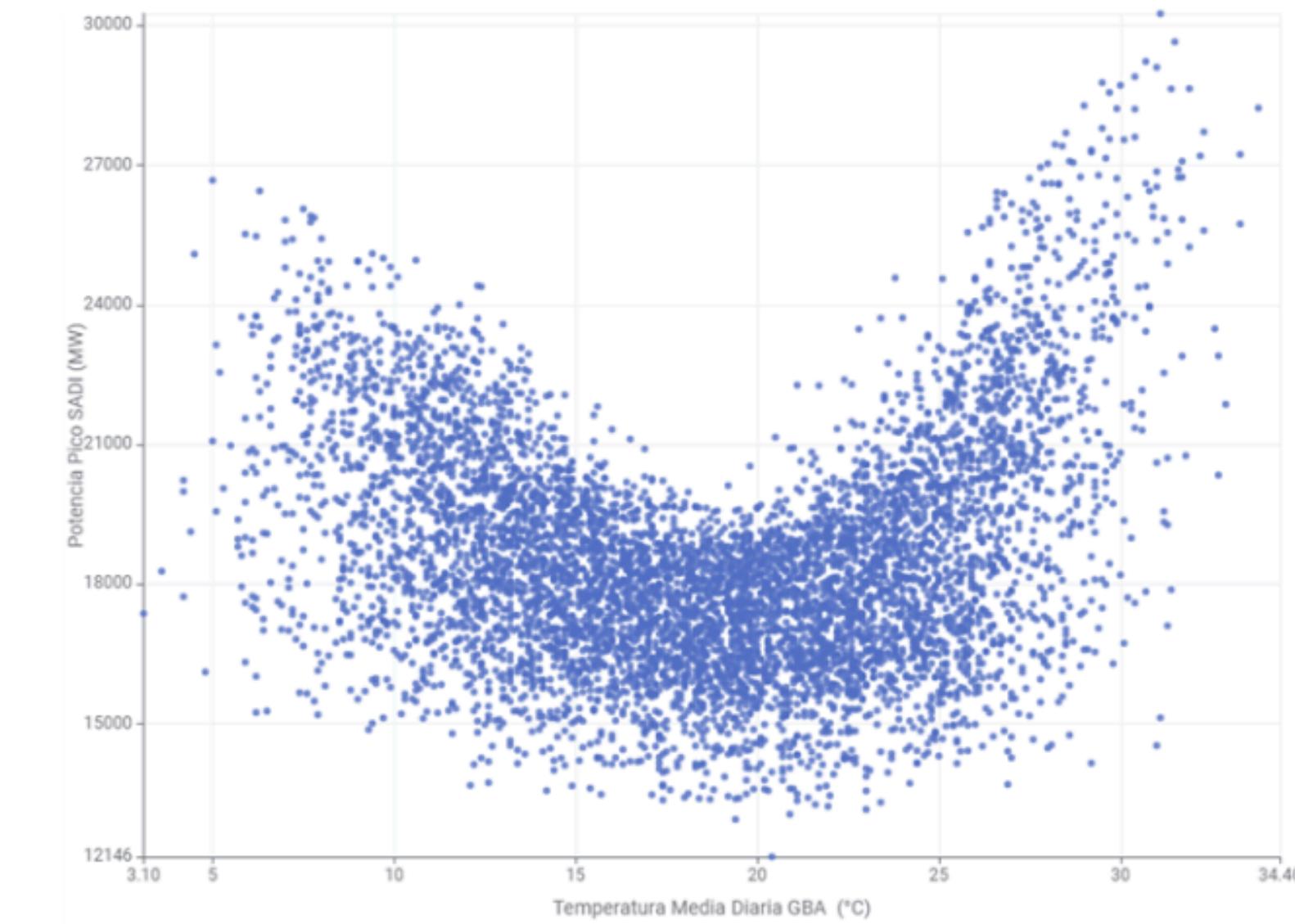


Complejidad del problema: hay relaciones no lineales y comportamientos inesperados entre variables como Temperatura y Potencia

RowID	New Cluster Number (integer)	Entropia de Cluster Hi Number (double)
Row0	0	1.535
Row1	1	1.367
Row2	2	1.372

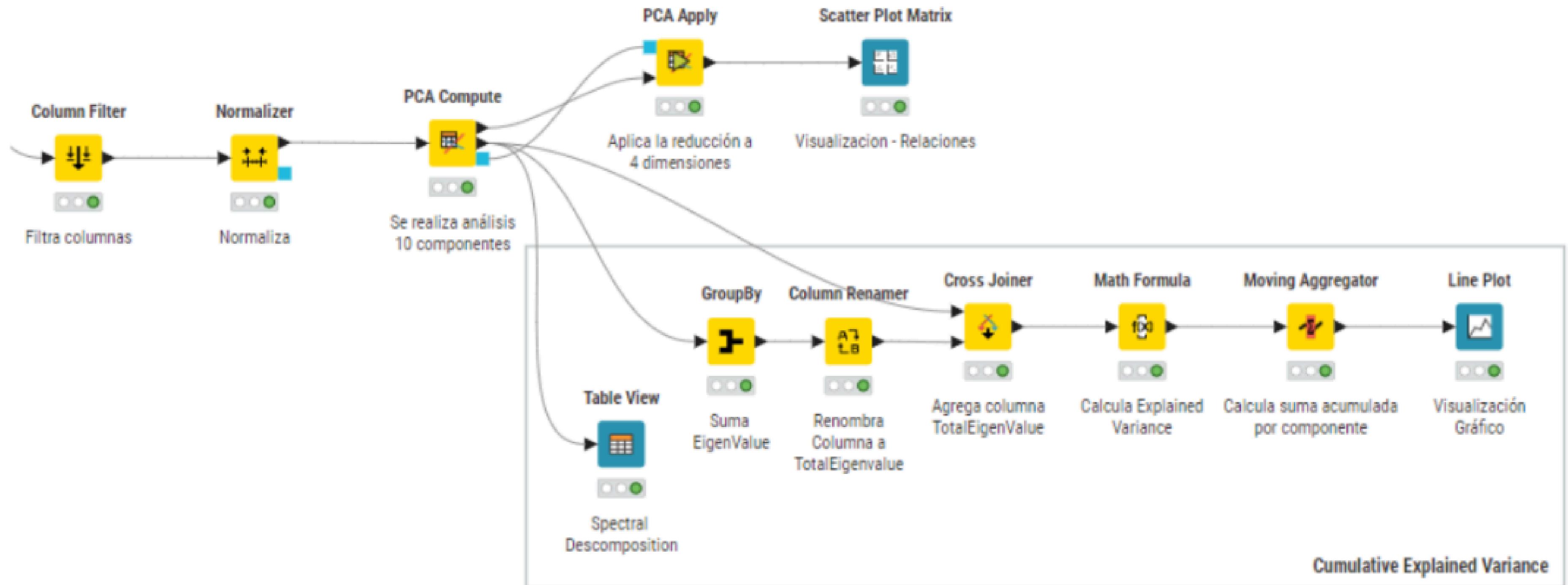
New Cluster Number (integer)	Entropia de Cluster Hi Number (double)
0	1.366
1	1.248
2	0.745
3	0.021
4	1.367
5	1.372

Valores de entropía por cluster



- Variable categórica analizada: Potencia (baja, media, alta)
- Resultados esperables por la complejidad del problema:
 - Casos donde Potencia y Energía no evolucionan igual
 - Relación no lineal entre Temperatura y Potencia

PCA

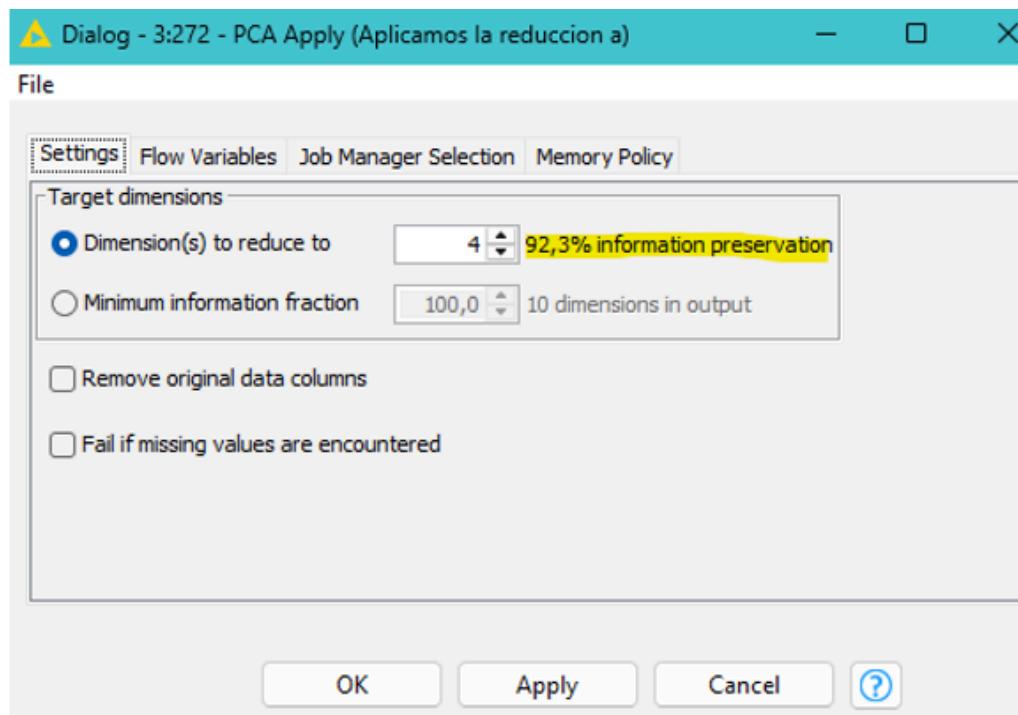


PCA

- Reducción de dimensionalidad
- Preservación de información

¿Qué buscamos?

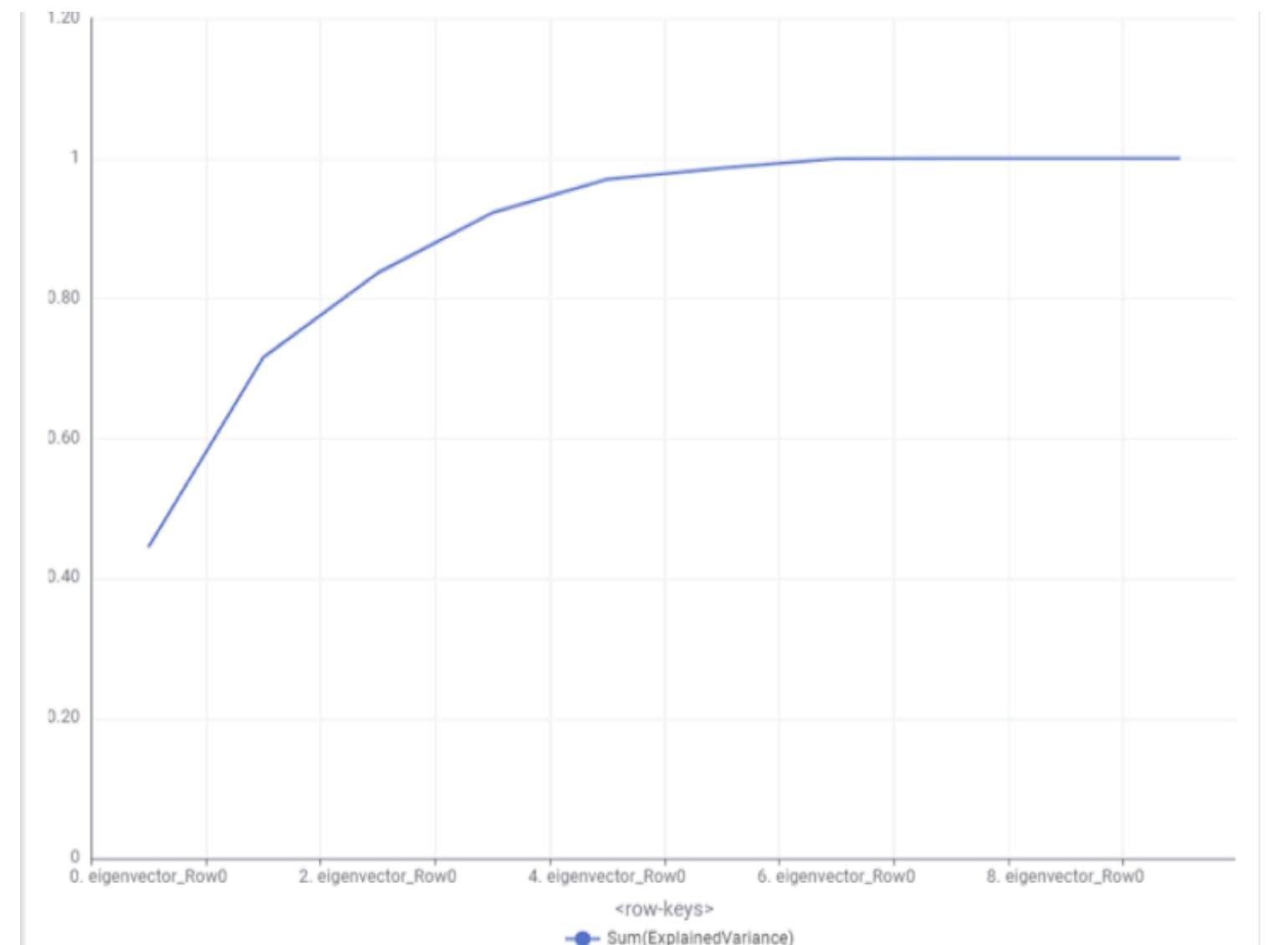
- Preservar al menos un 90% de información reduciendo la dimensionalidad de 10 atributos.



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

Metrica

- Se utiliza la métrica Varianza explicada acumulada.
- Se obtiene que con 4 dimensiones se logra un 92,3% de información preservada.
- A partir del séptimo componente, agregar más componentes no aumenta el %.



Clustering y K óptimo

Se evaluaron distintos valores de k usando Silhouette y entropía. Se eligió k = 4 por ofrecer un buen equilibrio entre separación y homogeneidad de los clústeres.

Patrones frecuentes con Reglas de Asociación

Se identificaron 113 reglas, destacándose una con 67,8% de confianza y lift 2,796, útil para anticipar situaciones de alta demanda.

Relaciones no lineales

Se observaron relaciones complejas entre variables como temperatura y potencia, lo que refuerza la necesidad de segmentar adecuadamente los datos.

Reducción de Dimensionalidad

Con PCA se logró reducir a 4 componentes manteniendo 92,3% de la varianza, simplificando el análisis sin perder información relevante.

Agrupamientos estacionales

El PCA reveló agrupaciones claras entre días hábiles/no hábiles y estaciones del año, lo que ayuda a interpretar comportamientos de demanda.

Conclusiones de los Modelos Descriptivos

Para CAMMESA, esto facilita la toma de decisiones estratégicas y la optimización de la gestión de la energía en el SADI.