

## Preparación y Depuración de Datos

### Reboiler Raw Data with Timestamps

**Elaborado por:**

- Elian Herrera Ravanal
- Cristian Contreras
- Cristhoper Muñoz
- Luis Rodrigo Olivares Andrade

**Asignatura:**

Computación Aplicada

**Profesor:**

Raymi Antonio Vásquez Moreno

**Fecha de Entrega:**

14/06/2025

**Universidad:**

Universidad Federico Santa María

# Introducción

En el presente documento se detalla el proceso de preparación y depuración de datos del archivo "Reboiler Raw Data with Timestamps". Se describen las etapas clave para garantizar la calidad y consistencia de los datos, desde la importación y visualización inicial, pasando por la eliminación de valores nulos y atípicos, hasta el ajuste de formatos y resampleo temporal.

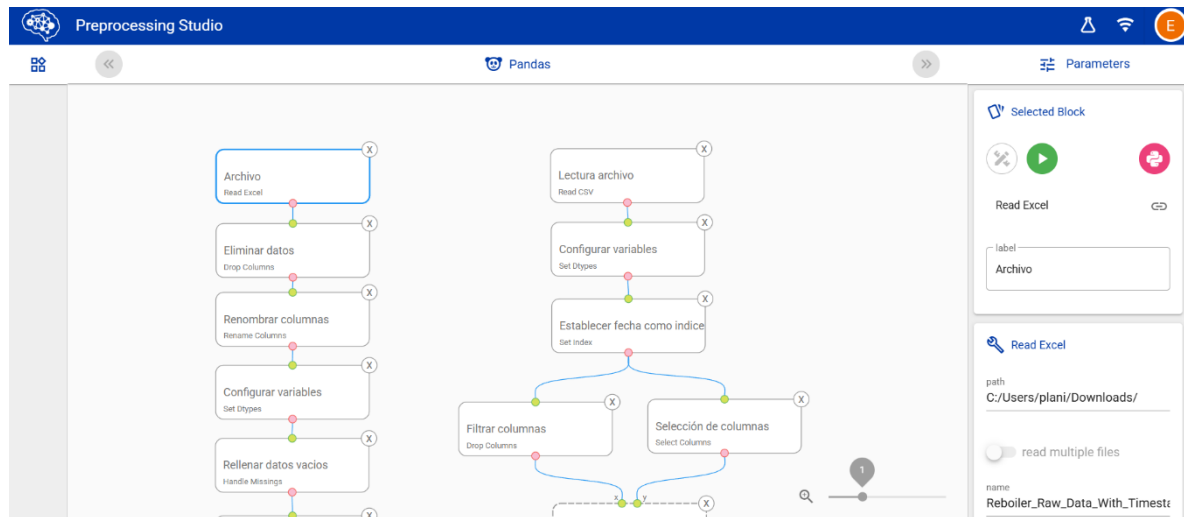
El objetivo principal fue estructurar un conjunto de datos confiable y balanceado, facilitando su análisis posterior. Particular atención se dio al manejo de datos anómalos y al filtrado temporal, asegurando la integridad de las series desde enero de 2016. Finalmente, el resultado se exportó en un archivo CSV depurado, listo para alimentar modelos de predicción o análisis avanzados.

Este enfoque asegura un conjunto de datos robusto y útil para la toma de decisiones basadas en evidencia.

<b>Introducción</b>	<b>2</b>
<b>Procedimiento realizado</b>	<b>3</b>
<b>Conclusión</b>	<b>21</b>

# Procedimiento realizado

## PASO 1

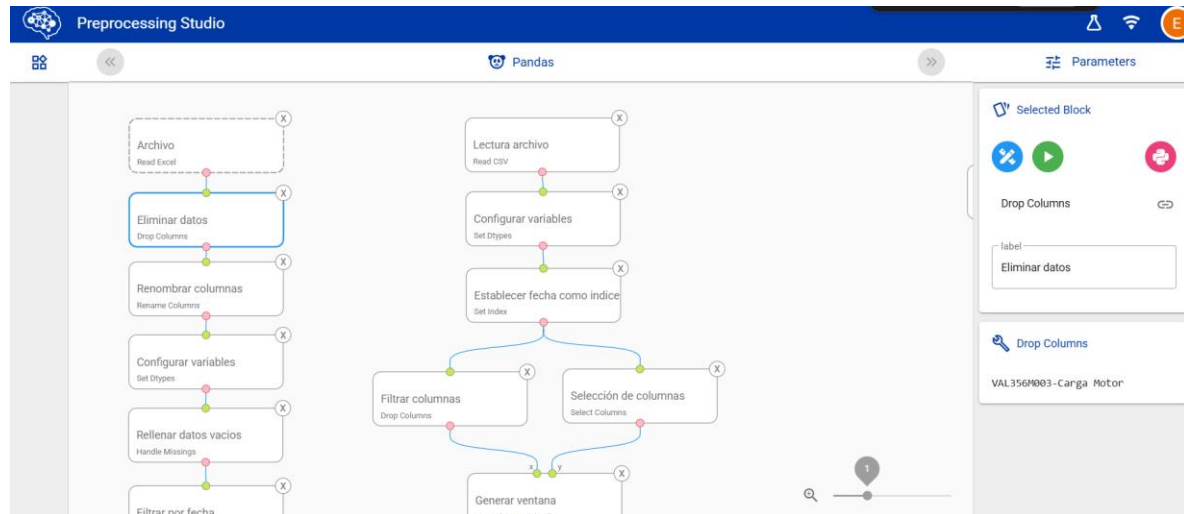


The screenshot shows the 'DataFrame' tab in the Preprocessing Studio interface, displaying a table with the following columns:

	Fecha	VAL356612803- Conductividad	VAL3566963- Carga Motor	VAL3566964- Carga Motor	VAL3566965- Carga Motor	VAL3566966- Ind. Presión	VAL3566967- Ind. Presión	VAL3566968- Ind. Temperatura	VAL3566969- Ind. Temperatura	Estado
0	2006-10-24 23:30:00	12.3	68.826459	45.670056	38.343373	21.482553	28.9	71.883617	183.434846	0
1	2006-10-25 01:30:00	18.9	68.827098	45.924569	38.126654	20.954956	28.6	72.381453	183.482949	0
2	2006-10-25 03:30:00	25.9	68.873664	49.477336	39.781442	21.130845	28.3	71.775936	183.532936	0
3	2006-10-25 05:30:00	27.2	65.658285	51.649763	38.828497	21.021897	28.6	70.589861	183.637808	0
4	2006-10-25 07:30:00	22.0	64.568553	46.851182	27.708380	20.975888	28.4	71.762377	183.688362	0
...	...	...	...	...	...	...	...	...	...	...
42281	2018-09-06 01:30:00	7.4	NaN	72.696518	43.404327	19.695458	19.5	81.372489	99.596179	0
42282	2018-09-06 03:30:00	7.3	NaN	72.989898	44.888929	19.826846	19.5	81.911585	99.633788	0
42283	2018-09-06 05:30:00	6.9	NaN	72.968914	44.584887	19.523726	19.5	86.838454	99.775938	0
42284	2018-09-06 07:30:00	6.9	NaN	72.985579	44.427335	19.388117	19.5	81.587228	100.136855	0
42285	2018-09-06 09:30:00	7.0	NaN	72.956738	44.288788	19.335458	19.5	81.812438	100.482443	0

1) En esta etapa se realiza la lectura del archivo Excel titulado “Reboiler Raw Data with Timestamps” utilizando la biblioteca Pandas. El objetivo es cargar adecuadamente el conjunto de datos para visualizar su estructura —incluyendo filas y columnas— y así confirmar que la importación se ha efectuado correctamente. Esta verificación inicial es clave para asegurar la integridad de los datos antes de cualquier análisis posterior.

## PASO 2

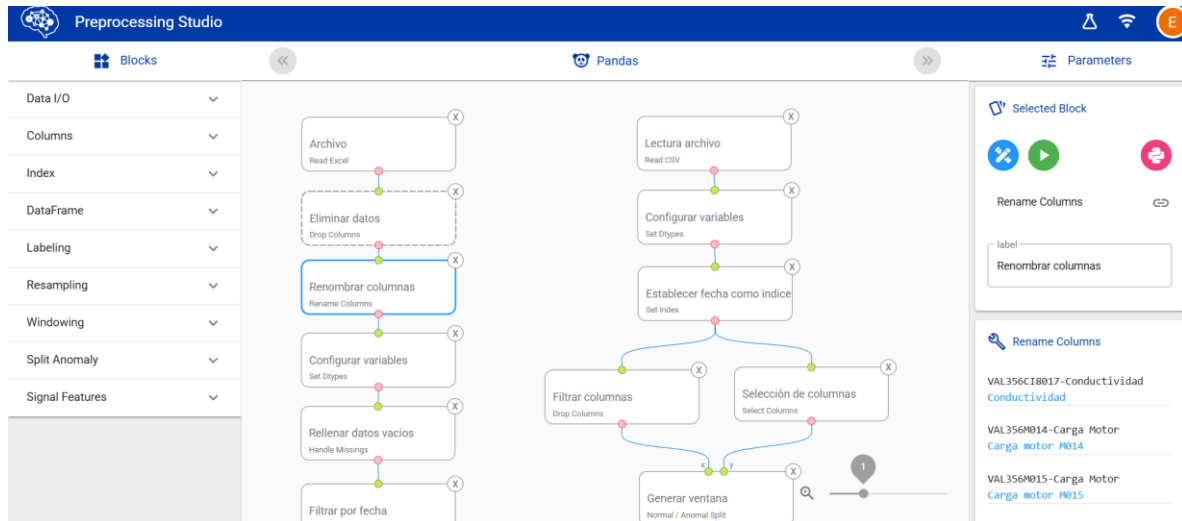


The screenshot shows the Preprocessing Studio interface with the 'RESULTS' tab. A DataFrame is displayed with columns: Fecha, VAL356CT0017-Conductividad, VAL356M014-Carga Motor, VAL356M015-Carga Motor, VAL356P10026-Ind.Presión, VAL356P10025-Ind.Presión, VAL356T10015-Ind.Temperatura, VAL356T10014-Ind.Temperatura, and Estado. The table shows data for various dates and times, with some rows having missing values (indicated by '...').

	Fecha	VAL356CT0017-Conductividad	VAL356M014-Carga Motor	VAL356M015-Carga Motor	VAL356P10026-Ind.Presión	VAL356P10025-Ind.Presión	VAL356T10015-Ind.Temperatura	VAL356T10014-Ind.Temperatura	Estado
0	2006-10-24 23:38:00	12.3	45.670056	30.343173	21.492553	20.9	71.883617	103.414846	0
1	2006-10-25 01:38:00	18.9	45.924569	30.126654	20.954956	20.6	72.301453	103.487449	0
2	2006-10-25 03:38:00	25.9	49.477196	29.701442	21.193045	20.3	71.775536	103.582936	0
3	2006-10-25 05:38:00	27.2	51.649763	28.820497	21.021897	20.6	70.589861	103.697800	0
4	2006-10-25 07:38:00	22.0	46.851182	27.760388	20.975880	20.4	71.762377	103.680362	0
...	...	...	...	...	...	...	...	...	...
42201	2018-09-06 01:38:00	7.4	72.696510	43.464317	19.695450	19.5	81.372489	99.598170	0

2) En este paso se eliminan los datos nulos que podrían generar errores o inconsistencias al momento de analizar la información. Particularmente, la columna “VAL356M003-Carga Motor” de valores faltantes, lo que dificulta su recuperación y puede afectar negativamente el rendimiento de cualquier modelo que se utilice en el estudio. Por esta razón, se opta por descartarla para asegurar la calidad del análisis.

## PASO 3

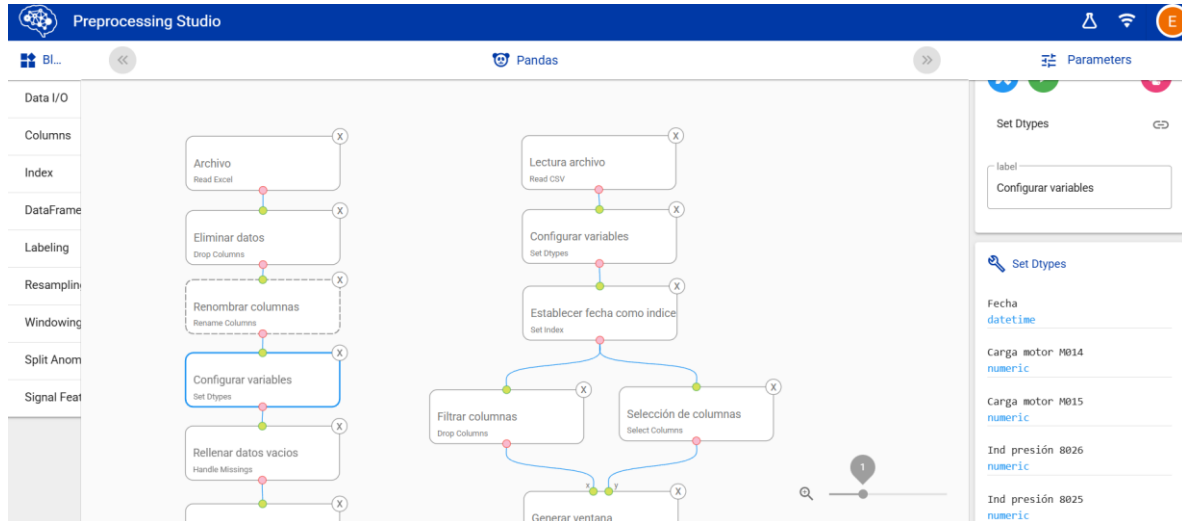


The screenshot shows the 'RESULTS' tab in the Preprocessing Studio interface. The 'DataFrame' is displayed as a table with columns: Fecha, Conductividad, Carga motor M014, Carga motor M015, Ind presión B026, Ind presión B025, Ind temp B015, Ind temp B014, and Estado. The table shows data for various timestamps, including 2006-10-24 and 2018-09-06.

	Fecha	Conductividad	Carga motor M014	Carga motor M015	Ind presión B026	Ind presión B025	Ind temp B015	Ind temp B014	Estado
0	2006-10-24 23:38:00	12.3	45.670056	30.343173	21.492553	20.9	71.883617	103.414846	0
1	2006-10-25 01:38:00	18.9	45.924569	30.126654	20.954956	20.6	72.301453	103.487449	0
2	2006-10-25 03:38:00	25.9	49.477196	29.701442	21.193045	20.3	71.775536	103.582936	0
3	2006-10-25 05:38:00	27.2	51.649763	28.820497	21.021897	20.6	70.589861	103.697800	0
4	2006-10-25 07:38:00	22.0	46.851182	27.760388	20.975880	20.4	71.762377	103.680362	0
...	...	...	...	...	...	...	...	...	...
42201	2018-09-06 01:38:00	7.4	72.696510	43.464317	19.695450	19.5	81.372489	99.598170	0
42202	2018-09-06 03:38:00	7.2	72.905098	44.058929	19.826046	19.5	81.911505	99.633708	0
42203	2018-09-06 05:38:00	6.9	72.968914	44.584087	19.523726	19.5	80.810454	99.775918	0
42204	2018-09-06 07:38:00	6.9	72.905579	44.417335	19.308117	19.5	81.587220	100.136055	0

3) Este paso consiste en renombrar las variables del conjunto de datos a un formato más simple y legible, facilitando su interpretación por parte del usuario. Los nombres originales de los sensores se reemplazan por sus respectivos números de TAG, mientras que la variable correspondiente a la conductividad se etiqueta de forma directa como "Conductividad". Esto mejora la visualización y el manejo de los datos durante el análisis.

## PASO 4

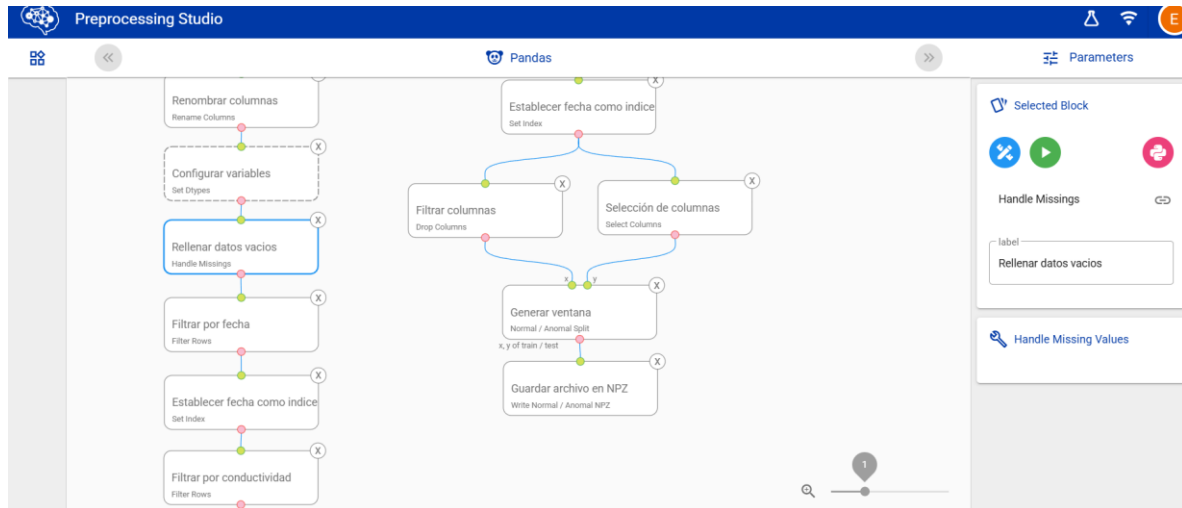


The screenshot shows the 'RESULTS' tab in the Preprocessing Studio interface. The 'DataFrame' is displayed as a table with columns: Fecha, Conductividad, Carga motor M014, Carga motor M015, Ind presión B026, Ind presión B025, Ind temp B015, Ind temp B014, and Estado. The table shows data for various time points, including 2006-10-24 and 2018-09-06.

	Fecha	Conductividad	Carga motor M014	Carga motor M015	Ind presión B026	Ind presión B025	Ind temp B015	Ind temp B014	Estado
0	2006-10-24 23:38:00	12.3	45.670056	30.343173	21.492553	20.9	71.883617	103.414846	0
1	2006-10-25 01:38:00	18.9	45.924569	30.126654	20.954956	20.6	72.301453	103.487449	0
2	2006-10-25 03:38:00	25.9	49.477196	29.701442	21.193045	20.3	71.775536	103.582936	0
3	2006-10-25 05:38:00	27.2	51.649763	28.820497	21.021897	20.6	70.589861	103.697800	0
4	2006-10-25 07:38:00	22.0	46.851182	27.760388	20.975800	20.4	71.762377	103.680362	0
...	...	...	...	...	...	...	...	...	...
42201	2018-09-06 01:38:00	7.4	72.696510	41.464317	19.695450	19.5	81.372489	99.598170	0
42202	2018-09-06 03:38:00	7.2	72.905098	44.058929	19.826046	19.5	81.911505	99.633708	0
42203	2018-09-06 05:38:00	6.9	72.968914	44.584887	19.523726	19.5	80.810454	99.775918	0
42204	2018-09-06 07:38:00	6.9	72.905579	44.417335	19.308117	19.5	81.587220	100.136055	0

4) En este paso se ajusta el formato de las variables según lo requerido para el análisis. Las lecturas de los sensores se convierten a tipo numérico para permitir operaciones matemáticas y estadísticas, mientras que la columna de fecha se transforma al formato estándar “DateTime” (YYYY-MM-DD hh:mm:ss), asegurando su correcta interpretación temporal durante el procesamiento de los datos.

## PASO 5



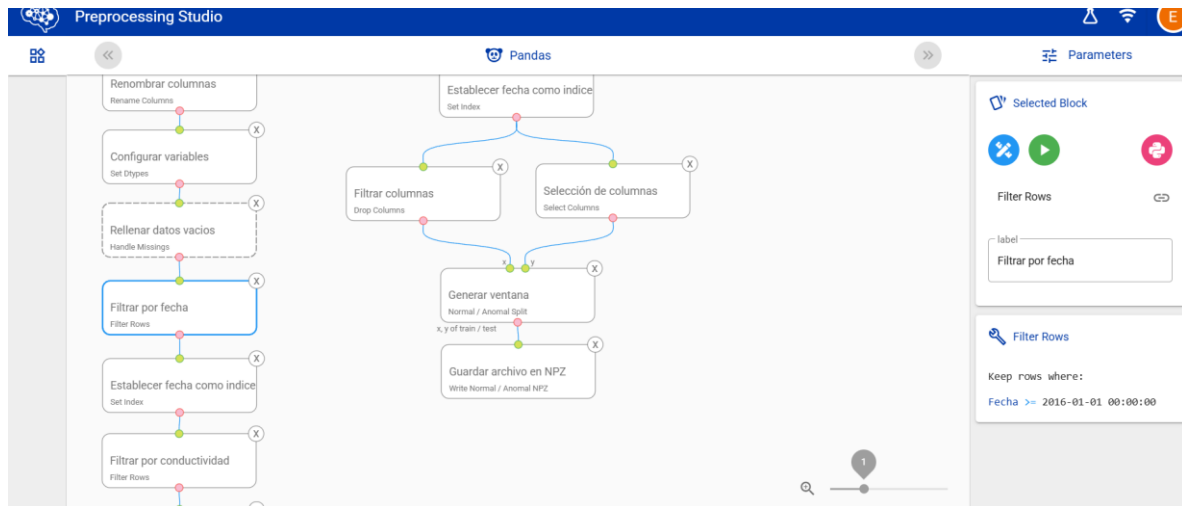
RESULTS EXPLORE

DataFrame

	Fecha	Conductividad	Carga motor PM14	Carga motor PM15	Ind presión 0026	Ind presión 0025	Ind temp 0015	Ind temp 0014	Estado
0	2006-10-24 23:38:00	12.3	45.678056	30.343173	21.492553	20.9	71.883617	103.414846	0
1	2006-10-25 01:38:00	18.9	45.924569	30.126654	20.954956	20.6	72.381453	103.487449	0
2	2006-10-25 03:38:00	25.9	49.477196	29.701442	21.193045	20.3	71.775536	103.582936	0
3	2006-10-25 05:38:00	27.2	51.649763	28.820497	21.021897	20.6	70.589861	103.697800	0
4	2006-10-25 07:38:00	22.0	46.851182	27.760388	20.975880	20.4	71.762377	103.680362	0
...	...	...	...	...	...	...	...	...	...
42201	2018-09-06 01:38:00	7.4	72.696510	43.464317	19.695450	19.5	81.372489	99.598170	0
42202	2018-09-06 03:38:00	7.2	72.905098	44.058929	19.826046	19.5	81.911505	99.633708	0
42203	2018-09-06 05:38:00	6.9	72.968914	44.584087	19.523726	19.5	80.810454	99.775918	0
42204	2018-09-06 07:38:00	6.9	72.905579	44.417335	19.308117	19.5	81.587220	100.136055	0
42205	2018-09-06 09:38:00	7.0	72.956738	44.203780	19.335458	19.5	81.812438	100.482443	0

5) En este paso se rellenan los valores nulos utilizando el método “forward fill” (ffill), que consiste en reemplazar cada dato faltante con el valor válido anterior más cercano. Esta técnica permite mantener la continuidad de las mediciones sin introducir valores artificiales, lo que es útil cuando se trabaja con series temporales o datos de sensores.

## PASO 6



RESULTS EXPLORE

Back To Work

DataFrame

	Fecha	Conductividad	Carga motor 0014	Carga motor 0015	Ind presión 0026	Ind presión 0025	Ind temp 0015	Ind temp 0014	Estado
30676	2016-01-08 11:38:00	4.5	69.658713	42.280192	18.769251	19.0	79.399877	99.160674	0
30677	2016-01-08 13:38:00	4.4	67.954487	42.381221	18.411020	19.0	78.184952	98.775669	0
30678	2016-01-08 15:38:00	4.2	69.688606	41.967134	18.601041	19.0	80.599028	101.165308	0
30679	2016-01-08 17:38:00	4.2	69.274182	42.124866	18.582166	19.0	79.301819	99.934365	0
30680	2016-01-08 19:38:00	4.2	69.095421	41.588124	18.599953	19.0	80.170472	100.804536	0
...	...	...	...	...	...	...	...	...	...
42201	2018-09-06 01:38:00	7.4	72.696510	43.464317	19.695450	19.5	81.372489	99.598170	0
42202	2018-09-06 03:38:00	7.2	72.905098	44.058929	19.826046	19.5	81.911505	99.633708	0
42203	2018-09-06 05:38:00	6.9	72.968914	44.584087	19.523726	19.5	80.810454	99.775918	0
42204	2018-09-06 07:38:00	6.9	72.905579	44.417335	19.308117	19.5	81.587220	100.136855	0
42205	2018-09-06 09:38:00	7.0	72.956738	44.203780	19.335458	19.5	81.812438	100.482443	0

Selected Block

Filter Rows

label

Filtrar por fecha

Filter Rows

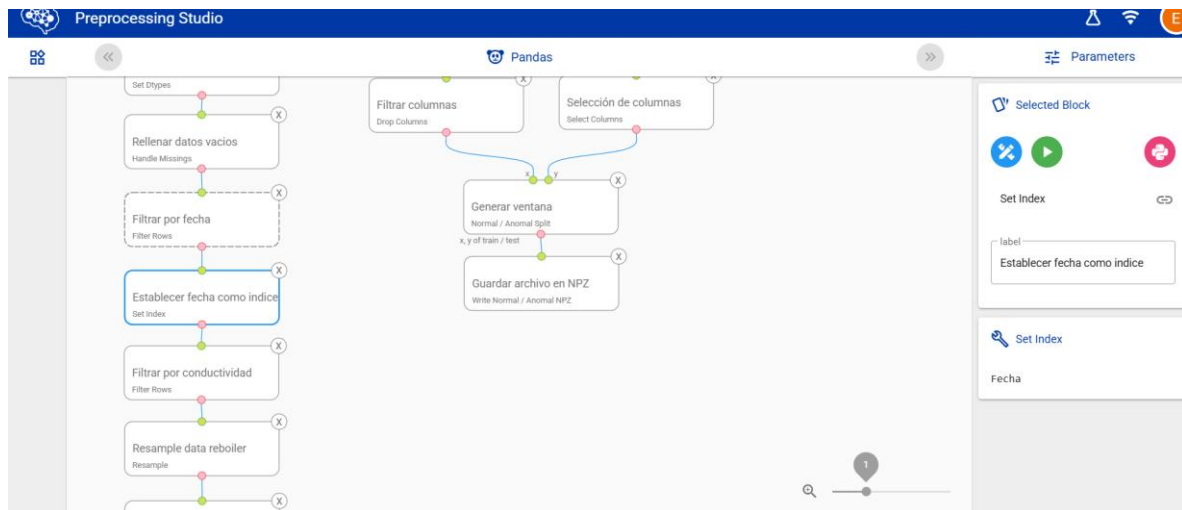
Keep rows where:

Fecha >= 2016-01-01 00:00:00

6) Este paso configura la utilización de los datos disponibles, a solamente considerar a la información más reciente, filtrando a partir del primer día efectivo registrado en la data desde el 2016 (2016-01-01 00:00:00), según lo requerido en el punto (b) de la tarea: Usar solo la data de la planta correspondiente a la etapa más reciente de operación del Reboiler, o sea, a partir del 2016. Para nuestra base de datos, la data disponible es a partir del 08-01-2016.



## PASO 7

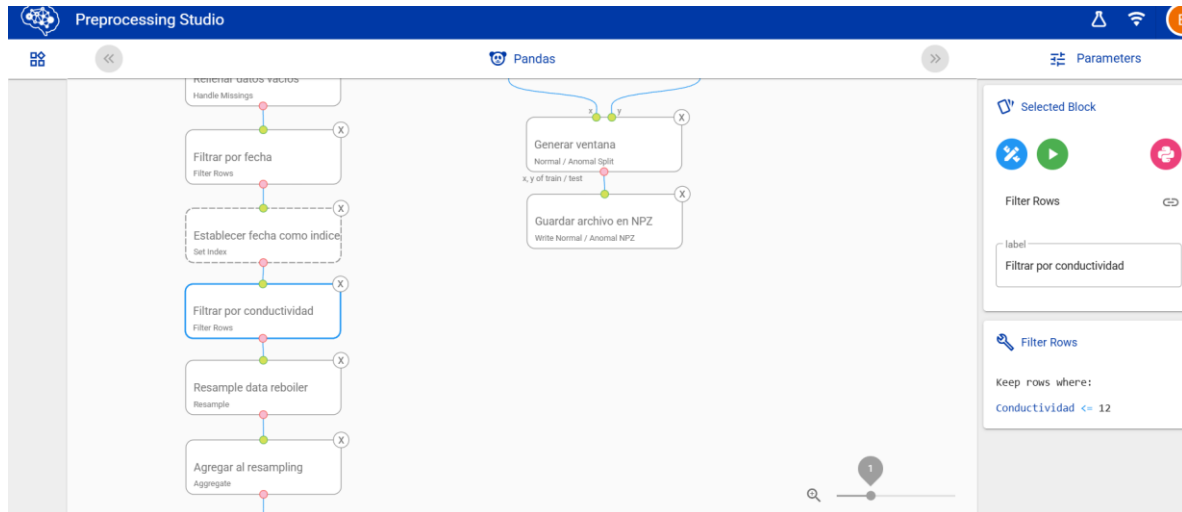


The screenshot shows the 'RESULTS' tab in the Preprocessing Studio interface. A DataFrame table is displayed with the following columns: Fecha, Conductividad, Carga motor MB14, Carga motor MB15, Ind presión 8026, Ind presión 8025, Ind temp 8015, Ind temp 8014, and Estado. The table contains data for various dates and times, starting from 2016-01-08 11:38:00.

Fecha	Conductividad	Carga motor MB14	Carga motor MB15	Ind presión 8026	Ind presión 8025	Ind temp 8015	Ind temp 8014	Estado
2016-01-08 11:38:00	4.5	69.658713	42.280192	18.769251	19.0	79.399877	99.160674	0
2016-01-08 13:38:00	4.4	67.954487	42.381221	18.411020	19.0	78.184952	98.775669	0
2016-01-08 15:38:00	4.2	69.688606	41.967134	18.601041	19.0	80.599028	101.165308	0
2016-01-08 17:38:00	4.2	69.274182	42.124866	18.582166	19.0	79.301819	99.934365	0
2016-01-08 19:38:00	4.2	69.095421	41.588124	18.593953	19.0	80.170472	100.804536	0
...	...	...	...	...	...	...	...	...
2018-09-06 01:38:00	7.4	72.696510	43.464317	19.695450	19.5	81.372489	99.598170	0
2018-09-06 03:38:00	7.2	72.905098	44.058929	19.826046	19.5	81.911505	99.633708	0
2018-09-06 05:38:00	6.9	72.968914	44.584087	19.523726	19.5	80.810454	99.775918	0
2018-09-06 07:38:00	6.9	72.905579	44.417335	19.308117	19.5	81.587220	100.136055	0

7) En este paso se establece la columna “Fecha” como índice principal del conjunto de datos. Esto permite que cada fila quede identificada por su respectiva marca de tiempo (fecha y hora), lo cual es fundamental para el análisis temporal. Como se indicó previamente, los datos considerados comienzan a partir del 8 de enero de 2016, fecha desde la cual se estructurará todo el análisis.

## PASO 8



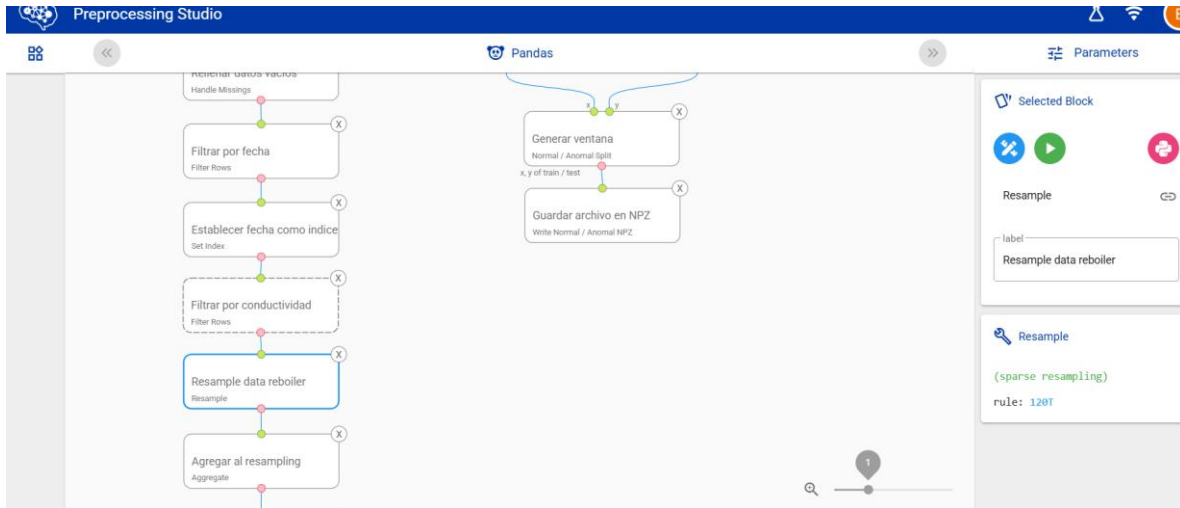
The screenshot shows the 'RESULTS' tab in the Preprocessing Studio interface. The data table has 8 columns and 11166 rows. The first column is 'Fecha'. The table shows data for various dates and times, with values for 8 columns.

Fecha								
2016-01-08 11:38:00	4.5	69.658713	42.280192	18.769251	19.0	79.399877	99.160674	0
2016-01-08 13:38:00	4.4	67.954487	42.381221	18.411020	19.0	78.184952	98.775669	0
2016-01-08 15:38:00	4.2	69.688606	41.967134	18.601041	19.0	80.599028	101.165308	0
2016-01-08 17:38:00	4.2	69.274182	42.124866	18.582166	19.0	79.301819	99.934365	0
2016-01-08 19:38:00	4.2	69.095421	41.588124	18.593953	19.0	80.170472	100.804536	0
...	...	...	...	...	...	...	...	...
2018-09-06 01:38:00	7.4	72.696510	43.464317	19.695450	19.5	81.372489	99.598170	0
2018-09-06 03:38:00	7.2	72.905098	44.058929	19.826046	19.5	81.911505	99.633708	0
2018-09-06 05:38:00	6.9	72.968914	44.584087	19.523726	19.5	80.810454	99.775918	0
2018-09-06 07:38:00	6.9	72.505579	44.417335	19.308117	19.5	81.587220	100.136055	0
2018-09-06 09:38:00	7.0	72.956738	44.203780	19.335458	19.5	81.812438	100.482443	0

11166 rows x 8 columns

8) En este paso se eliminan todas las mediciones consideradas no válidas o fuera de rango (“outliers”). En particular, se filtran aquellas lecturas de conductividad que sean menores o iguales a 12 microSiemens/cm ( $\leq 12$ ), ya que se consideran mediciones erróneas o físicamente imposibles, y podrían afectar negativamente la calidad del análisis.

## PASO 9



The screenshot shows the 'RESULTS' tab in the Preprocessing Studio. It displays two data windows. The first window, titled 'First Window', shows a timestamp of '2016-01-08 10:00:00' and a table with 8 columns: Fecha, Conductividad, Carga motor M014, Carga motor M015, Ind presión 8026, Ind presión 8025, Ind temp 8015, and Estado. The second window, titled 'Second Window', shows a timestamp of '2016-01-08 12:00:00' and a similar table with 8 columns: Fecha, Conductividad, Carga motor M014, Carga motor M015, Ind presión 8026, Ind presión 8025, Ind temp 8015, and Estado.

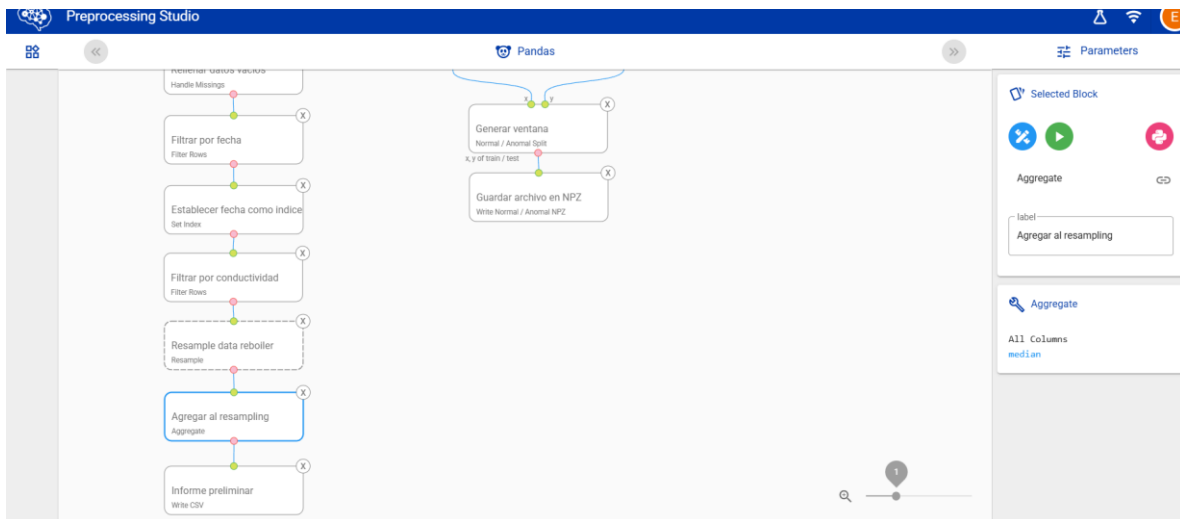
Fecha	Conductividad	Carga motor M014	Carga motor M015	Ind presión 8026	Ind presión 8025	Ind temp 8015	Ind temp 8014	Estado
2016-01-08 11:38:00	4.5	69.658713	42.280192	18.769251	19.0	79.399877	99.168674	0

Fecha	Conductividad	Carga motor M014	Carga motor M015	Ind presión 8026	Ind presión 8025	Ind temp 8015	Ind temp 8014	Estado
2016-01-08 13:38:00	4.4	67.954487	42.381221	18.41182	19.0	78.184952	98.775869	0

9) El método “resample” se utiliza para reagrupar los datos en intervalos de tiempo definidos, permitiendo ajustar la frecuencia de muestreo según el criterio del analista. Esta técnica es útil para homogeneizar la serie temporal y facilitar su análisis. En este caso específico, se decide mantener la frecuencia de muestreo original presente en los datos, establecida en intervalos de 120 minutos (2 horas), respetando así el ritmo de registro definido en la fuente de datos.

## PASO 10



The screenshot shows the Preprocessing Studio interface with the 'RESULTS' tab selected. The table displays data for various parameters over time. The table has 8 columns: Fecha, Conductividad, Carga motor M014, Carga motor M015, Ind presión B026, Ind presión B025, Ind temp B015, Ind temp B014, and Estado. The table shows data for 11166 rows.

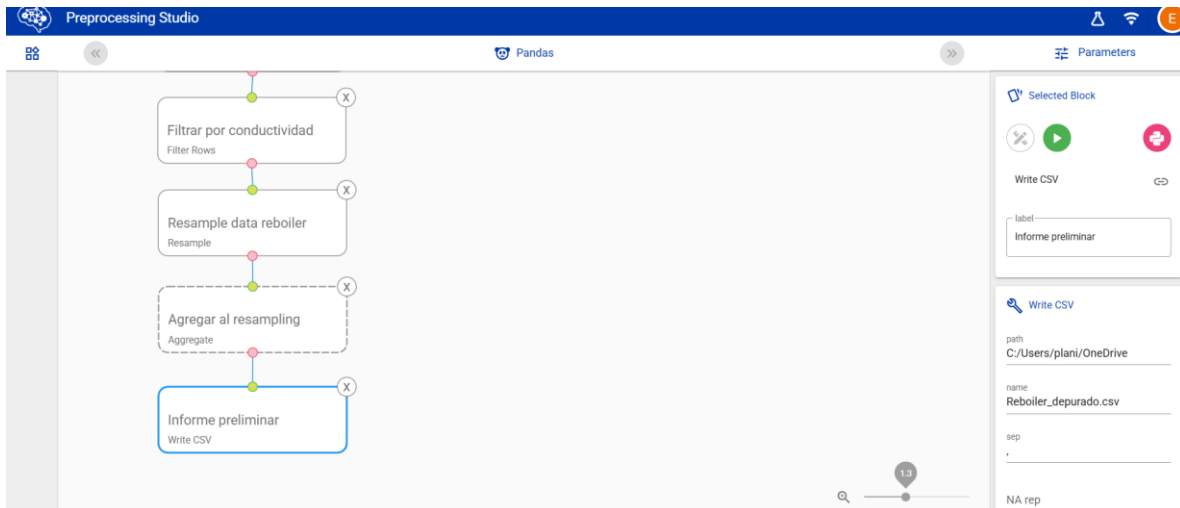
Fecha	Conductividad	Carga motor M014	Carga motor M015	Ind presión B026	Ind presión B025	Ind temp B015	Ind temp B014	Estado
2016-01-08 10:00:00	4.5	69.658713	42.280192	18.769251	19.0	79.399877	99.160674	0.0
2016-01-08 12:00:00	4.4	67.954487	42.381221	18.411808	19.0	78.184952	98.775669	0.0
2016-01-08 14:00:00	4.2	69.688866	41.967134	18.681941	19.0	80.599828	101.165388	0.0
2016-01-08 16:00:00	4.2	69.274182	42.124866	18.582166	19.0	79.301819	99.934365	0.0
2016-01-08 18:00:00	4.2	69.095421	41.588124	18.593953	19.0	80.170472	100.884536	0.0
...	...	...	...	...	...	...	...	...
2018-09-06 00:00:00	7.4	72.696510	43.464317	19.695450	19.5	81.372489	99.598170	0.0
2018-09-06 02:00:00	7.2	72.985898	44.058929	19.820846	19.5	81.911585	99.633788	0.0
2018-09-06 04:00:00	6.9	72.968914	44.584087	19.523726	19.5	80.818454	99.775918	0.0
2018-09-06 06:00:00	6.9	72.985579	44.417335	19.388117	19.5	81.587220	100.136855	0.0
2018-09-06 08:00:00	7.0	72.956738	44.203780	19.335458	19.5	81.812438	100.482443	0.0

11166 rows x 8 columns

10) El método “aggregate” se aplica inmediatamente después de realizar un “resample” y permite resumir los datos agrupados según una función matemática que represente adecuadamente su comportamiento. En este caso, se opta por utilizar la *mediana* como función de agregación, ya que ofrece una representación robusta y menos

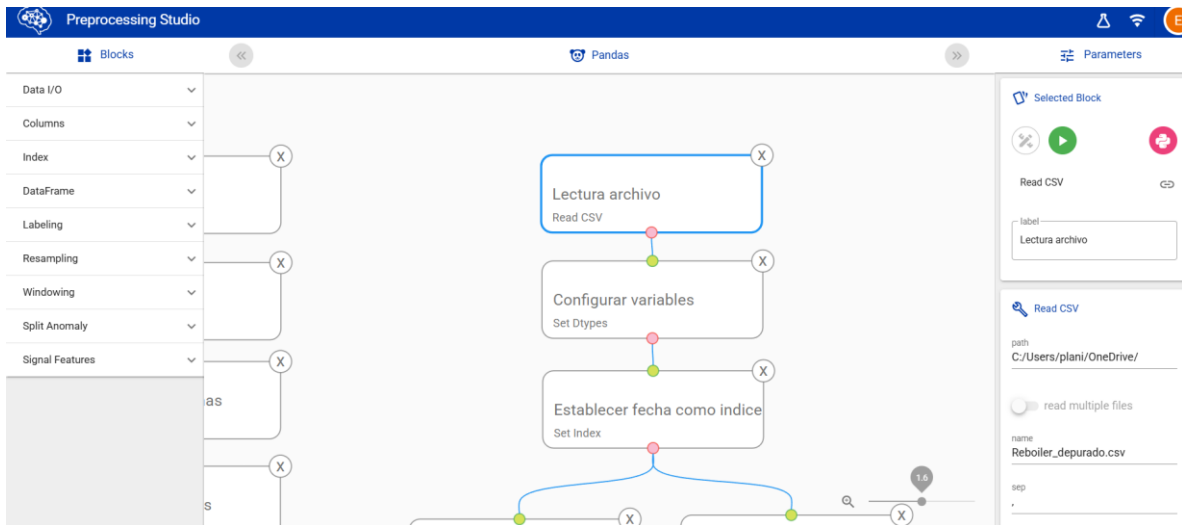
sensible a valores extremos o atípicos (outliers). Esta elección mejora la estabilidad del análisis posterior y mantiene la coherencia de los datos ajustados en el DataFrame.

## PASO 11



11) En este paso, se guarda el conjunto de datos procesado en un archivo CSV limpio y depurado, en formato plano (Flat File Format). Este archivo representa una versión preliminar lista para análisis posteriores o para alimentar modelos de predicción. El nombre asignado al archivo es: Reboiler\_depurado.csv

## PASO 12



RESULTS EXPLORE

Back To Work

DataFrame

	Fecha	Conductividad	Carga motor M014	Carga motor M015	Ind presión 8026	Ind presión 8025	Ind temp 8015	Ind temp 8014	Estado
0	2016-01-08 10:00:00	4.5	69.658713	42.280192	18.769251	19.0	79.399877	99.160674	0.0
1	2016-01-08 12:00:00	4.4	67.954487	42.381221	18.411820	19.0	78.184952	98.775669	0.0
2	2016-01-08 14:00:00	4.2	69.688606	41.967134	18.681841	19.0	80.599828	101.165388	0.0
3	2016-01-08 16:00:00	4.2	69.274182	42.124866	18.582166	19.0	79.301819	99.934365	0.0
4	2016-01-08 18:00:00	4.2	69.095421	41.588124	18.593953	19.0	80.170472	100.804536	0.0
...	...	...	...	...	...	...	...	...	...
11161	2018-09-06 00:00:00	7.4	72.696510	43.464317	19.695458	19.5	81.372489	99.598170	0.0
11162	2018-09-06 02:00:00	7.2	72.905898	44.858929	19.828046	19.5	81.911585	99.633788	0.0
11163	2018-09-06 04:00:00	6.9	72.968914	44.584087	19.523726	19.5	80.818454	99.775918	0.0
11164	2018-09-06 06:00:00	6.9	72.905579	44.417335	19.308117	19.5	81.587220	100.136955	0.0
11165	2018-09-06 08:00:00	7.0	72.956738	44.203788	19.335458	19.5	81.812438	100.482443	0.0

11166 rows x 9 columns

Selected Block

Read CSV

label

Lectura archivo

Read CSV

path

C:/Users/plani/OneDrive/

read multiple files

name

Reboiler\_depurado.csv

sep

12) En esta etapa, se lee el archivo previamente guardado en el paso 11.

## PASO 13

**Preprocessing Studio**

**Workflow:**

- Lectura archivo (Read CSV)
- Configurar variables (Set Dtypes)
- Establecer fecha como índice (Set Index)
- Filtrar columnas (Drop Columns)
- Selección de columnas (Select Columns)

**Set Dtypes Configuration:**

- Fecha: datetime
- Carga motor M014: numeric
- Carga motor M015: numeric
- Ind presión B026: numeric

**DataFrame Preview:**

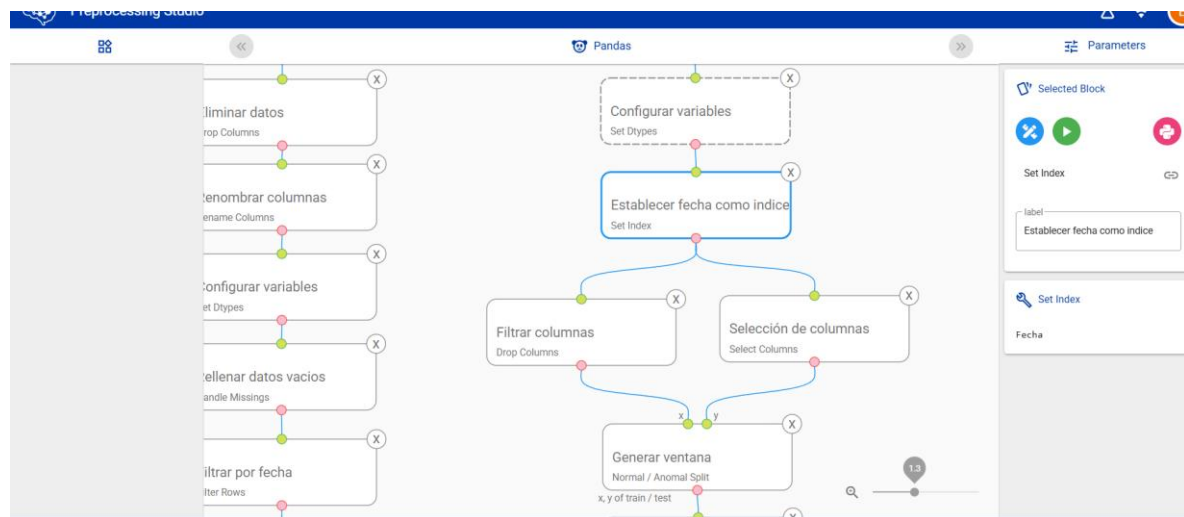
	Fecha	Conductividad	Carga motor M014	Carga motor M015	Ind presión B026	Ind presión B025	Ind temp B015	Ind temp B014	Estado
0	2016-01-08 10:00:00	4.5	69.658713	42.280192	18.769251	19.0	79.399877	99.160674	0.0
1	2016-01-08 12:00:00	4.4	67.954487	42.381221	18.413020	19.0	78.184952	98.775669	0.0
2	2016-01-08 14:00:00	4.2	69.688606	41.967134	18.601041	19.0	80.599028	101.165300	0.0
3	2016-01-08 16:00:00	4.2	69.274382	42.124866	18.582166	19.0	79.381839	99.934365	0.0
4	2016-01-08 18:00:00	4.2	69.805421	41.588124	18.593953	19.0	80.170472	100.804536	0.0
...	...	...	...	...	...	...	...	...	...
11161	2018-09-06 00:00:00	7.4	72.696510	43.464317	19.695450	19.5	81.372489	99.598170	0.0
11162	2018-09-06 02:00:00	7.2	72.905098	44.858929	19.826046	19.5	81.911595	99.631700	0.0
11163	2018-09-06 04:00:00	6.9	72.968914	44.584087	19.523726	19.5	80.810454	99.775918	0.0
11164	2018-09-06 06:00:00	6.9	72.905579	44.417335	19.380117	19.5	81.587220	100.136055	0.0
11165	2018-09-06 08:00:00	7.0	72.956738	44.283780	19.335458	19.5	81.812438	100.402443	0.0

11166 rows x 9 columns

13) En este paso se ajusta el formato de las variables según lo requerido para el análisis. Las lecturas de los sensores se convierten a tipo numérico para permitir operaciones matemáticas y estadísticas, mientras que la columna de fecha se

transforma al formato estándar “DateTime” (YYYY-MM-DD hh:mm:ss), asegurando su correcta interpretación temporal durante el procesamiento de los datos.

## PASO 14



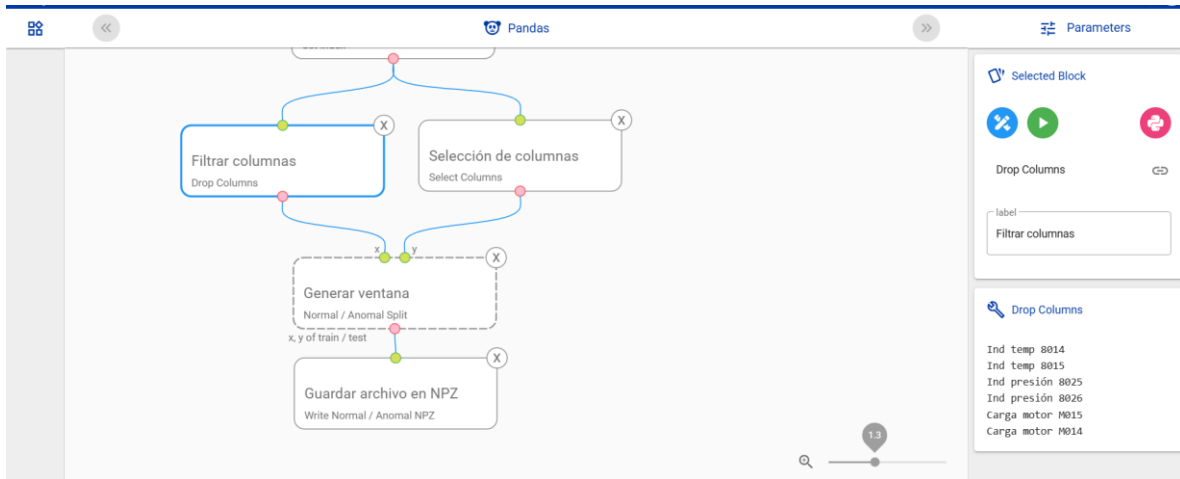
The screenshot shows the 'RESULTS' tab in the Anaconda Data Science Platform. A DataFrame is displayed with the following columns: Fecha, Conductividad, Carga motor M014, Carga motor M015, Ind presión B026, Ind presión B025, Ind temp B015, Ind temp B014, and Estado. The data is filtered to show rows from 2016-01-08 to 2018-09-06.

Fecha	Conductividad	Carga motor M014	Carga motor M015	Ind presión B026	Ind presión B025	Ind temp B015	Ind temp B014	Estado
2016-01-08 10:00:00	4.5	69.658713	42.288192	18.769251	19.0	79.399877	99.160674	0.0
2016-01-08 12:00:00	4.4	67.954487	42.381221	18.411820	19.0	78.184952	98.775669	0.0
2016-01-08 14:00:00	4.2	69.688686	41.967134	18.681041	19.0	80.599628	101.165388	0.0
2016-01-08 16:00:00	4.2	69.274182	42.124866	18.582166	19.0	79.381815	99.934365	0.0
2016-01-08 18:00:00	4.2	69.895421	41.588124	18.593953	19.0	80.178472	100.884536	0.0
...	...	...	...	...	...	...	...	...
2018-09-06 00:00:00	7.4	72.696510	43.464317	19.695450	19.5	81.373489	99.598170	0.0
2018-09-06 02:00:00	7.2	72.985808	44.058929	19.828086	19.5	81.911585	99.633788	0.0
2018-09-06 04:00:00	6.9	72.968914	44.584087	19.523726	19.5	80.818454	99.775918	0.0
2018-09-06 06:00:00	6.9	72.985579	44.417335	19.388117	19.5	81.587220	100.136855	0.0
2018-09-06 08:00:00	7.0	72.956738	44.203780	19.335458	19.5	81.812438	100.482443	0.0

14) En este paso se establece la columna “Fecha” como índice principal del conjunto de datos. Esto permite que cada fila quede identificada por su respectiva marca de tiempo (fecha y hora), lo cual es fundamental para el análisis temporal. Como se indicó previamente, los datos considerados comienzan a partir del 8 de enero de 2016, fecha desde la cual se estructurará todo el análisis.



## PASO 15

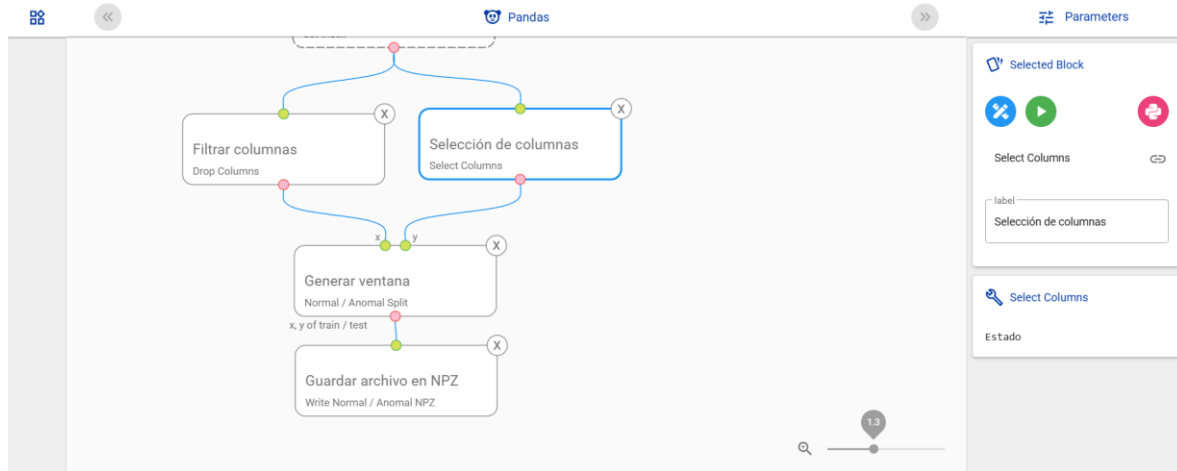


The 'RESULTS' tab displays the following data:

Fecha	Conductividad	Estado
2016-01-08 10:00:00	4.5	0.0
2016-01-08 12:00:00	4.4	0.0
2016-01-08 14:00:00	4.2	0.0
2016-01-08 16:00:00	4.2	0.0
2016-01-08 18:00:00	4.2	0.0
...	...	...
2018-09-06 00:00:00	7.4	0.0
2018-09-06 02:00:00	7.2	0.0
2018-09-06 04:00:00	6.9	0.0
2018-09-06 06:00:00	6.9	0.0
2018-09-06 08:00:00	7.0	0.0

15) En este paso se realiza un filtrado de datos para generar un normal Split, dejando los datos de conductividad y estado como eje X.

## PASO 16

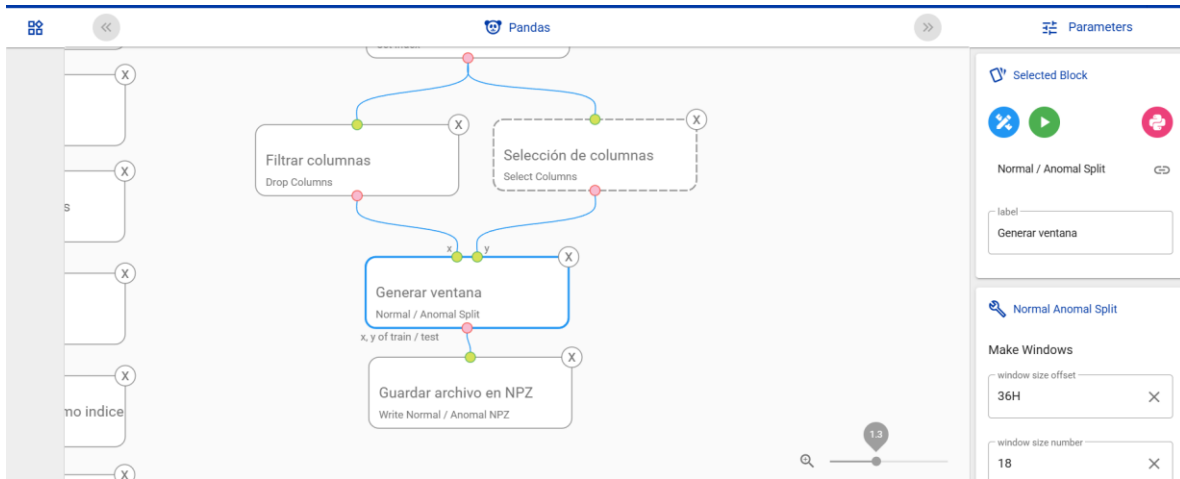


The screenshot shows the Anaconda workflow editor interface with the **RESULTS** tab selected. The table displays the output of the workflow, showing a list of dates and their corresponding 'Estado' (Status) values.

Fecha	Estado
2016-01-08 18:00:00	0.0
2016-01-08 12:00:00	0.0
2016-01-08 14:00:00	0.0
2016-01-08 16:00:00	0.0
2016-01-08 18:00:00	0.0
...	...
2018-09-06 00:00:00	0.0
2018-09-06 02:00:00	0.0
2018-09-06 04:00:00	0.0
2018-09-06 06:00:00	0.0
2018-09-06 08:00:00	0.0

16) En este paso se selecciona la columna de datos “Estado”, con el fin de dejarlo como eje Y.

## PASO 17



The screenshot shows the **RESULTS** tab in the Anaconda interface. It displays the output of the **Generar ventana** block, which is a dictionary containing training and testing data shapes:

```
{'x_train': '7105 x (18, 2)',
 'x_test': '3130 x (18, 2)',
 'y_test': '3130 x ()'}
```

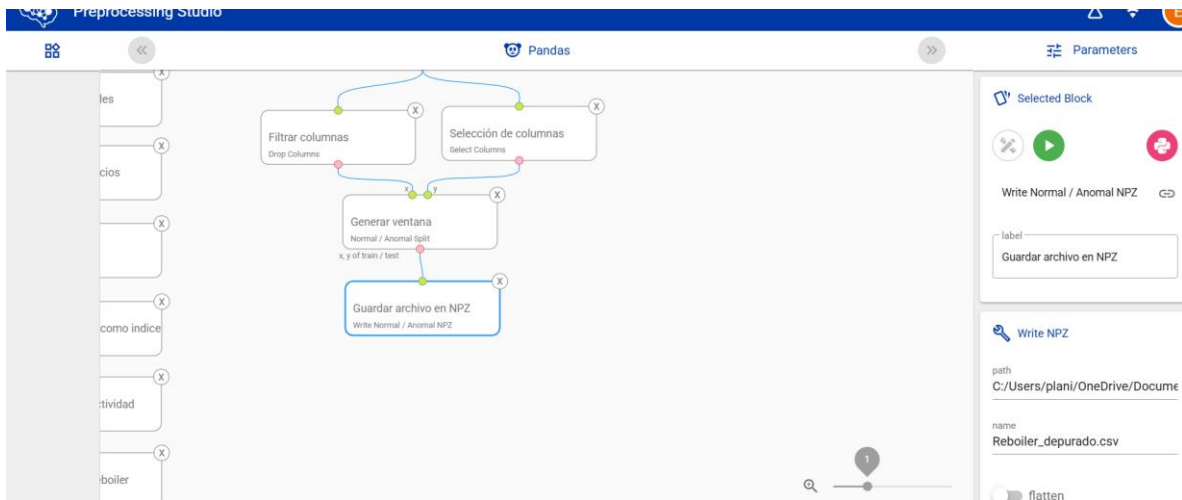
Below this, the **x train** data is shown as a NumPy array:

```
array([[0.77310924, 0.],
       [0.80672269, 0.],
       [0.79831933, 0.],
       ...,
       [0.71428571, 0.],
       [0.70588235, 0.],
       [0.70588235, 0.]],
       [[0.35294118, 0.],
       [0.35294118, 0.],
       [0.36134454, 0.],
       ...,
       [0.36134454, 0.],
       [0.36134454, 0.],
       [0.3697479, 0.]],
       [[0.40336134, 0.],
       [0.38655462, 0.],
       ...,
       [0.40336134, 0.],
       [0.38655462, 0.],
       [0.40336134, 0.]])
```

The right sidebar shows the **Parameters** panel for the **Normal / Anomal Split** block, which is identical to the one in the previous screenshot, with **label** set to **Generar ventana** and **Make Windows** parameters set to **36H** and **18**.

17) Se realiza una ventana de 36 Horas, con un número de tamaño de ventana de 18.

## PASO 18



18) En este último paso, se guarda el archivo con formato NPZ, generando una ruta de acceso que es: "C:\Users\plani\OneDrive\Documentos\Reboiler\_depurado.csv.npz".

## Conclusión

El procesamiento y depuración del conjunto de datos del Reboiler ha permitido obtener una versión limpia, estructurada y adecuada para su análisis. A través de la eliminación de valores nulos y atípicos, la estandarización de formatos, y la reorganización temporal de las mediciones, se ha logrado garantizar la calidad e integridad de la información. Además, el etiquetado de las condiciones operacionales y la verificación del balance entre clases aseguran que el conjunto de datos esté listo para ser utilizado en tareas de modelado predictivo o diagnóstico de fallas. respecto al total es particularmente valiosa, ya que proporciona una base sólida para entrenar modelos de balanceo adicionales. En definitiva, el archivo resultante representa un recurso confiable y robusto para estudios futuros relacionados con la operación y mantenimiento del Reboiler.