



UNIVERSIDAD TÉCNICA  
FEDERICO SANTA MARÍA

# BIGDATA

**ALUMNO: Abraham Mosqueira**

**Bastián Varas**

**Deivit Vega**

**PROFESOR GUÍA: Raymi Vázquez**

**ASIGNATURA: Computación aplicada**

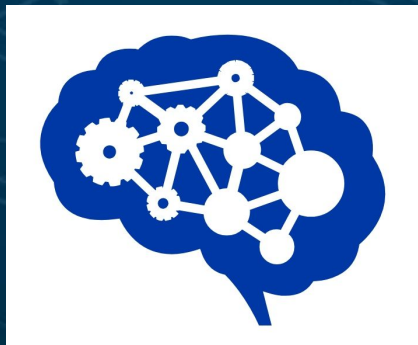
**usm.cl**



UNIVERSIDAD TECNICA  
FEDERICO SANTA MARIA

Para iniciar el procesamiento de datos recopilados por diferentes sensores de un reboiler ubicado en una planta generadora de dióxido de cloro, se utilizará la herramienta de análisis DataBruin.

**DataBruin ProProcessing Studio:** es una herramienta visual e interactiva que permite procesar y analizar datos sin necesidad de programar directamente, facilitando tareas como cargar archivos, filtrar, agrupar o resumir información mediante bloques visuales, siendo ideal para estudiantes que trabajan con datos de sensores, series temporales o procesos industriales para realizar análisis de sus datos y variaciones







UNIVERSIDAD TECNICA  
FEDERICO SANTA MARIA

- Los cuadrados representan bloques de procesamiento que ejecutan una acción específica mediante programaciones predeterminadas, pudiendo leer y filtrar datos, modificar tipos o eliminar columnas. Estos bloques se conectan entre sí para transformar los datos de forma progresiva y ordenada
- El flujo completo abarca desde la importación del archivo Excel hasta la exportación final de el archivo csv con la tasa de muestreo necesario para el análisis.





UNIVERSIDAD TECNICA  
FEDERICO SANTA MARIA

lectura del excel

Read Excel

**Read Excel:** Se comenzó el procesamiento de datos con el bloque de comandos Read Excel, utilizado para importar los datos desde un Excel al entorno de trabajo DataBruin ProProcessing Studio con el fin de trabajar con ellos de forma visual. Este paso permite acceder a las mediciones registradas y utilizarlas en distintos tipos de análisis.

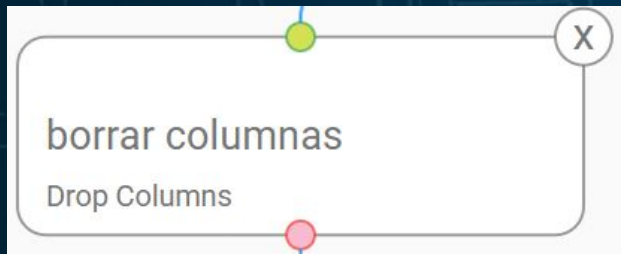
	Fecha	VAL356CI8017- Conductividad	VAL356M003- Carga Motor	VAL356M014- Carga Motor	VAL356M015- Carga Motor	VAL356PI8026- Ind.Presión	VAL356PIC8025- Ind.Presión	VAL356TI8015- Ind.Temperatura	VAL356TIC8014- Ind.Temperatura	Estado
30676	2016-01-08 11:38:00	4.5	NaN	69.658713	42.280192	18.769251	19.0	79.399877	99.160674	0
30677	2016-01-08 13:38:00	4.4	NaN	67.954487	42.381221	18.411020	19.0	78.184952	98.775669	0
30678	2016-01-08 15:38:00	4.2	NaN	69.688606	41.967134	18.601041	19.0	80.599028	101.165308	0
30679	2016-01-08 17:38:00	4.2	NaN	69.274182	42.124866	18.582166	19.0	79.301819	99.934365	0
30680	2016-01-08 19:38:00	4.2	NaN	69.095421	41.588124	18.593953	19.0	80.170472	100.804536	0
...	...	...	...	...	...	...	...	...	...	...






UNIVERSIDAD TECNICA  
FEDERICO SANTA MARIA

**Drop Columns:** se utiliza para eliminar columnas seleccionadas de un conjunto de datos. En este caso, se utilizó para eliminar los sensores de carga de motor de la bomba M003, debido a mediciones inexistentes durante un largo período en los datos capturados, evitando resultados anómalos con respecto a la realidad y permitiendo un procesamiento más preciso.



DESIGN EXPLORE		
DataFrame Columns	Dropped Columns	Drop
Fecha	VAL356M003-Carga Motor	2 - VAL356M003-Carga Motor
VAL356CI8017-Conductividad		
VAL356M014-Carga Motor		
VAL356M015-Carga Motor		
VAL356PI8026-Ind.Presión		
VAL356PIC8025-Ind.Presión		
VAL356TI8015-Ind.Temperatura		
VAL356TIC8014-Ind.Temperatu...		
Estado		



Drop Columns

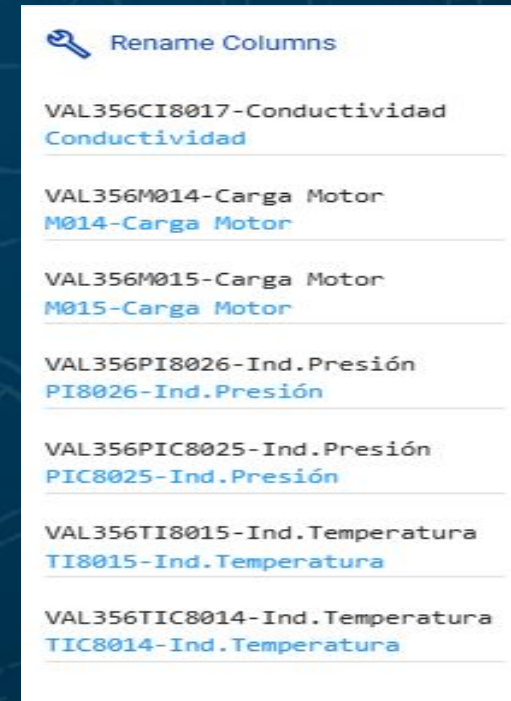
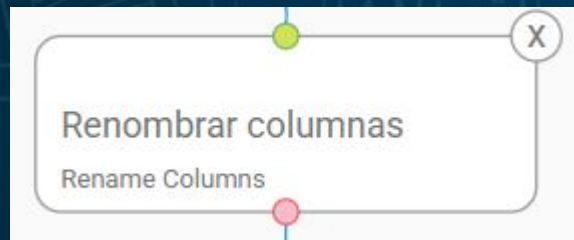
VAL356M003-Carga Motor



UNIVERSIDAD TECNICA  
FEDERICO SANTA MARIA

**Rename Columns:** permite cambiar el nombre de una o más columnas del conjunto de datos de esa manera asignar nombres más claros o corregir errores

se realizaron cambios en Los nombres de las columnas se han actualizado para que sean más intuitivos. Los originales (en negro) han sido reemplazados por versiones abreviadas y mas claras (en azul).







Definir tipo de datos  
Set Dtypes

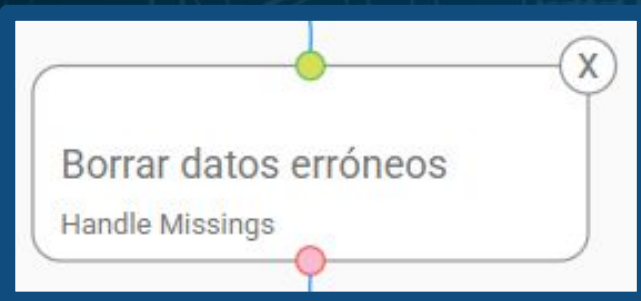
Selected Block	Set Dtypes	
	Fecha <a href="#">datetime</a>	PIC8025-Ind.Presión <a href="#">numeric</a>
	Conductividad <a href="#">numeric</a>	TI8015-Ind.Temperatura <a href="#">numeric</a>
	M014-Carga Motor <a href="#">numeric</a>	TIC8014-Ind.Temperatura <a href="#">numeric</a>
	M015-Carga Motor <a href="#">numeric</a>	Estado <a href="#">numeric</a>

label  
configurar variables



UNIVERSIDAD TECNICA  
FEDERICO SANTA MARIA

**Handle missing:** Bloque que permite borrar datos erróneos del archivo, datos que no son número por ejemplo: espacios en blanco.



DESIGN EXPLORE

☒ All Columns

Handle Missing Values

function  
fillna

type  
method

method  
ffill

axis  
None

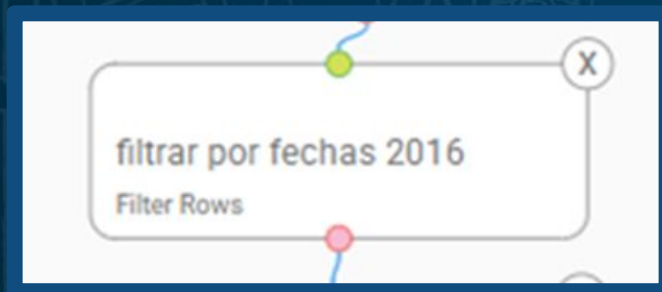
limit









UNIVERSIDAD TECNICA  
FEDERICO SANTA MARIA

**Filter Rows:** Se utiliza para filtrar filas de una tabla de datos según las condiciones aplicadas. En este caso, según el planteamiento se utilizó para filtrar los datos capturados a partir del año 2016 en adelante y eliminar la información de los años anteriores. De esta manera, se trabaja únicamente con la data correspondiente a la etapa más reciente de operación del Reboiler.




Selected Block

Filter Rows 

label  
filtrar en funcion de la fecha

---

 Filter Rows

Keep rows where:

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



UNIVERSIDAD TECNICA  
FEDERICO SANTA MARIA

**Set Index:** Se utiliza para establecer la columna de fecha como base temporal del análisis DataFrame, permitiendo ordenar y trabajar los datos según esa columna

establecer fecha como indice

Set Index

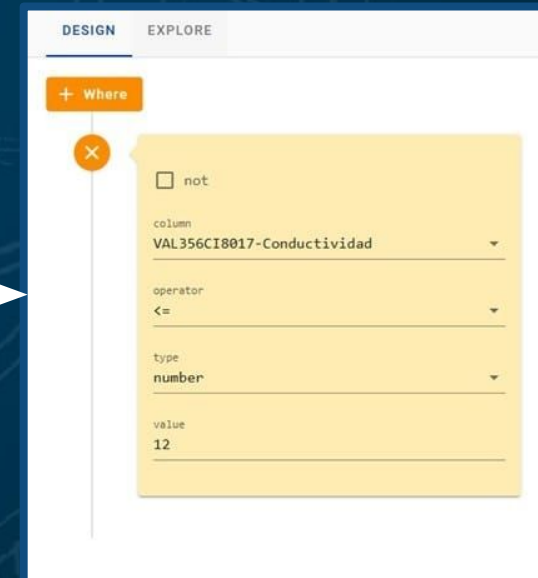
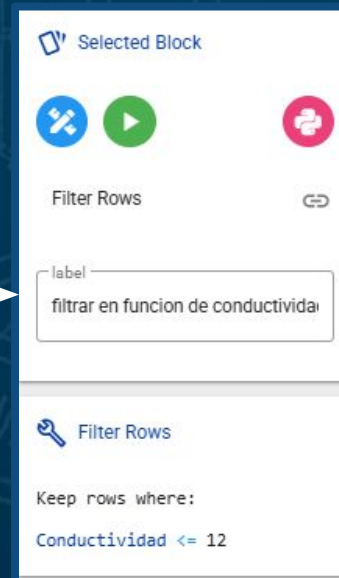
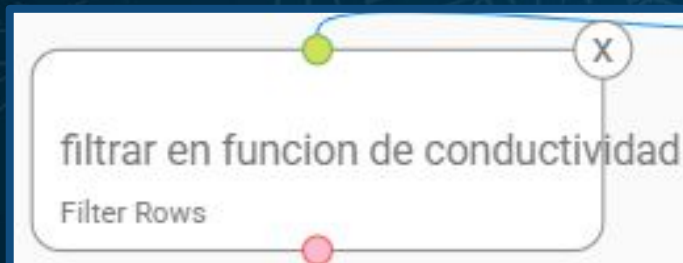
	VAL356CI8017- Conductividad	VAL356M014- Carga Motor	VAL356M015- Carga Motor	VAL356PI8026- Ind.Presión	VAL356PIC8025- Ind.Presión	VAL356TI8015- Ind.Temperatura	VAL356TIC8014- Ind.Temperatura	Estado
Fecha								
2016-01-08 11:38:00	4.5	69.658713	42.280192	18.769251	19.0	79.399877	99.160674	0
2016-01-08 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
...	...	...	...	...	...	...	...	...





UNIVERSIDAD TECNICA  
FEDERICO SANTA MARIA

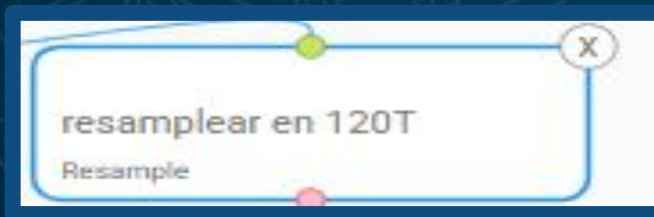
**Filter rows:** se utiliza para filtrar datos con ciertas características. En este caso, se filtraron los datos no posibles (outliers), los cuales no se consideran en el análisis por ser valores poco realistas. descartando la data del sensor de conductividad (CI8017) superior a 12 [ $\mu\text{S}/\text{cm}$ ], ubicado entre el tanque de condensado y la bomba, para asegurar mediciones representativas en el procesamiento.









UNIVERSIDAD TECNICA  
FEDERICO SANTA MARIA

La función resample permite reagrupar y ajustar los datos en intervalos o ventanas de tiempo definidos, según el criterio del analista. Esta herramienta resulta útil tanto para modificar la granularidad temporal de la información como para establecer una nueva tasa de muestreo, dependiendo del análisis a realizar. En este caso particular, se opta por conservar la tasa de muestreo original de la base de datos, la cual está establecida en intervalos de 120 minutos (2 horas).




Selected Block

Resample 

label

---

 Resample

(sparse resampling)

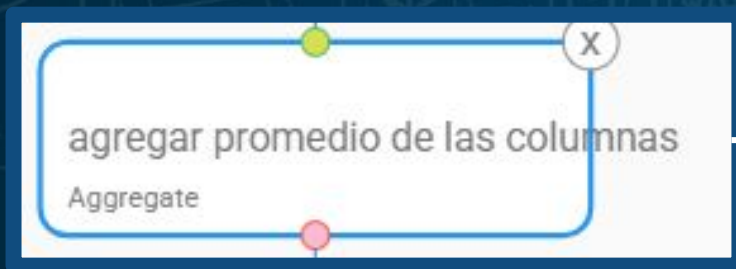
rule: 120T









UNIVERSIDAD TECNICA  
FEDERICO SANTA MARIA

Aggregate: Este bloque nos ayuda a agregar funciones, en este caso particular se añadió el promedio de las columnas.




Selected Block

Aggregate 

label

---

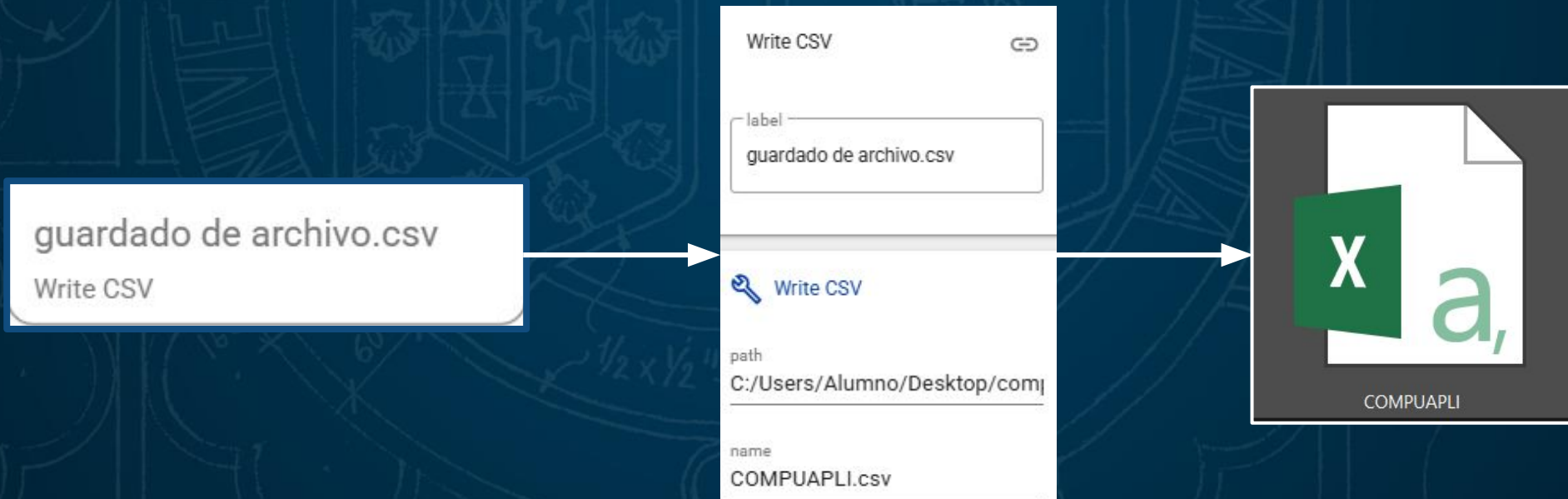
 Aggregate

All Columns  
median



UNIVERSIDAD TECNICA  
FEDERICO SANTA MARIA

**Write CSV:** Se utiliza para exportar los datos procesados a un archivo en formato .csv (valores separados por comas). Este bloque marca el paso final del flujo de trabajo, permitiendo guardar los resultados obtenidos tras la limpieza, filtrado y transformación de los datos.

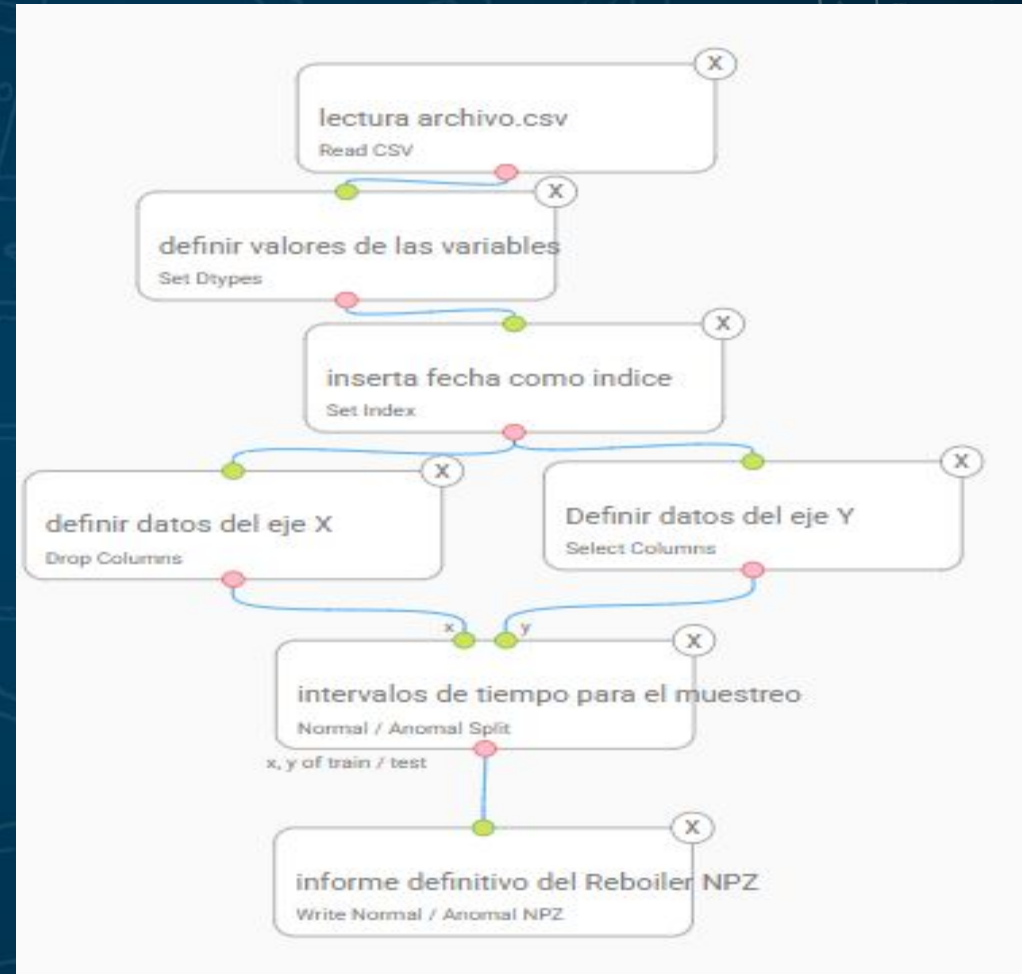
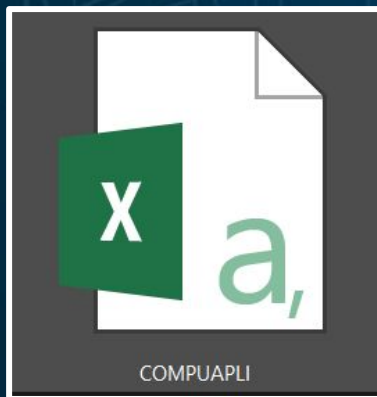






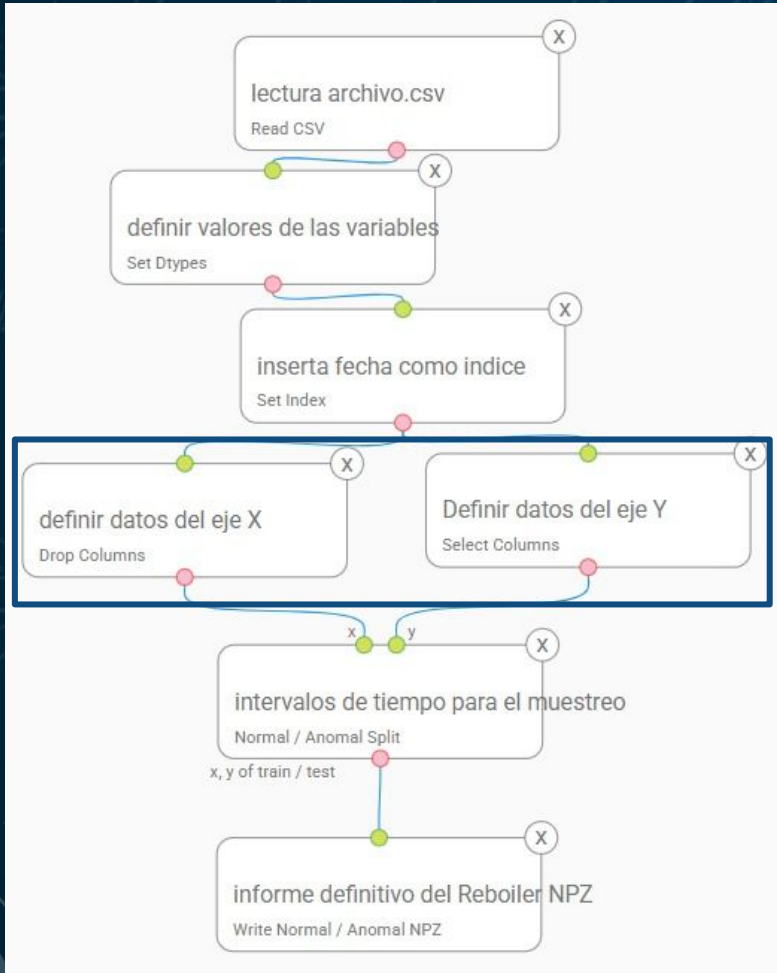
UNIVERSIDAD TECNICA  
FEDERICO SANTA MARIA

Ahora se debe leer el archivo.csv que creamos anteriormente, esto se hace con el bloque que ya conocemos Read CSV, luego redefinir los valores, posteriormente volver a insertar la fecha como índice de la columna.





UNIVERSIDAD TECNICA  
FEDERICO SANTA MARIA



definir datos del eje X  
Drop Columns

Definir datos del eje Y  
Select Columns

EJE X : SENSORES

EJE Y : ESTADO





UNIVERSIDAD TECNICA  
FEDERICO SANTA MARIA

### Make Windows

window size offset

600T



window size number

5



label to choose in window

last



Normal Data Split

☒ Custom Split

Train Ratio: 0.8

Test Ratio: 0.2

0.2

```
{'x_train': '6742 x (5,)', 'x_test': '4115 x (5,)', 'y_test': '4115 x ()'}
```

x train

```
array([[0., 0., 0., 0., 0.],  
       [0., 0., 0., 0., 0.],  
       [0., 0., 0., 0., 0.],  
       ...,  
       [0., 0., 0., 0., 0.],  
       [0., 0., 0., 0., 0.],  
       [0., 0., 0., 0., 0.]])
```

x test

```
array([[0., 0., 0., 0., 0.],  
       [1., 0., 1., 1., 0.],  
       [0., 0., 0., 0., 0.],  
       ...,  
       [1., 0., 1., 1., 1.],  
       [0., 1., 1., 1., 1.],  
       [1., 1., 1., 1., 1.]])
```

y test

```
array([0, 0, 0, ..., 1, 1, 1])
```

X\_TRAIN

NORMALES

X\_TEST

FALLA &  
NORMAL

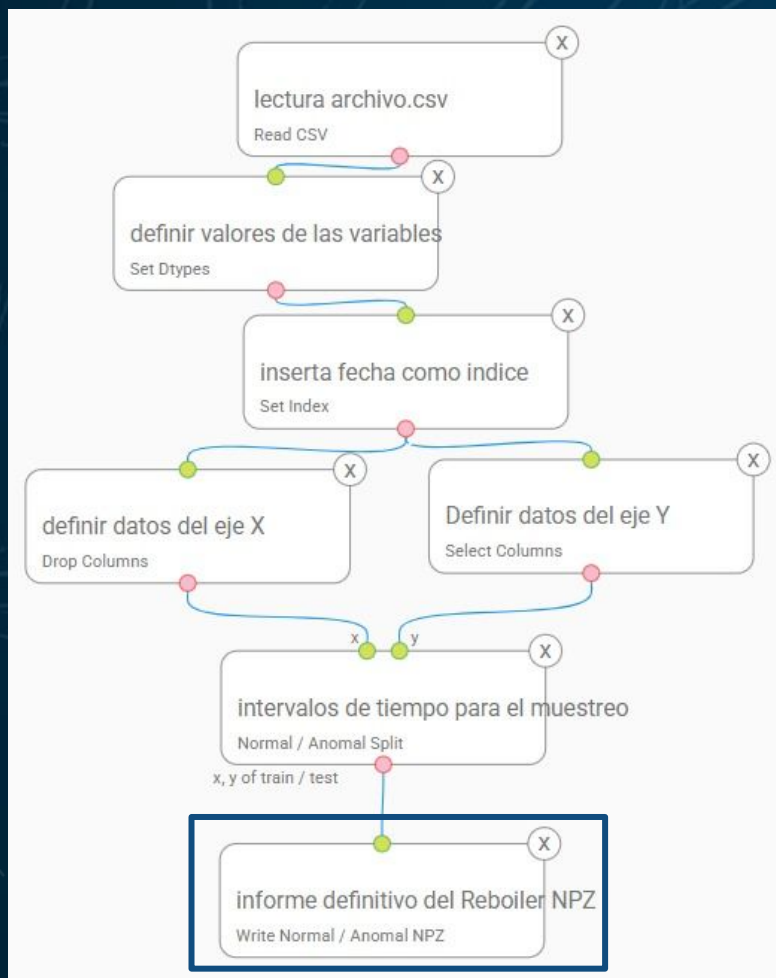
Y\_TEST

FALLA &  
NORMAL

usm.cl



UNIVERSIDAD TECNICA  
FEDERICO SANTA MARIA



Selected Block

Write Normal / Anomal NPZ

label

informe definitivo del Reboiler NF

Write NPZ

path

C:/Users/Alumno/Desktop/comj

name

informe final Reboiler NPZ

flatten

scaler name

scaler.joblib

