An introduction to the 3W Dataset

Andre Paulo Ferreira Machado
Celso Jose Munaro

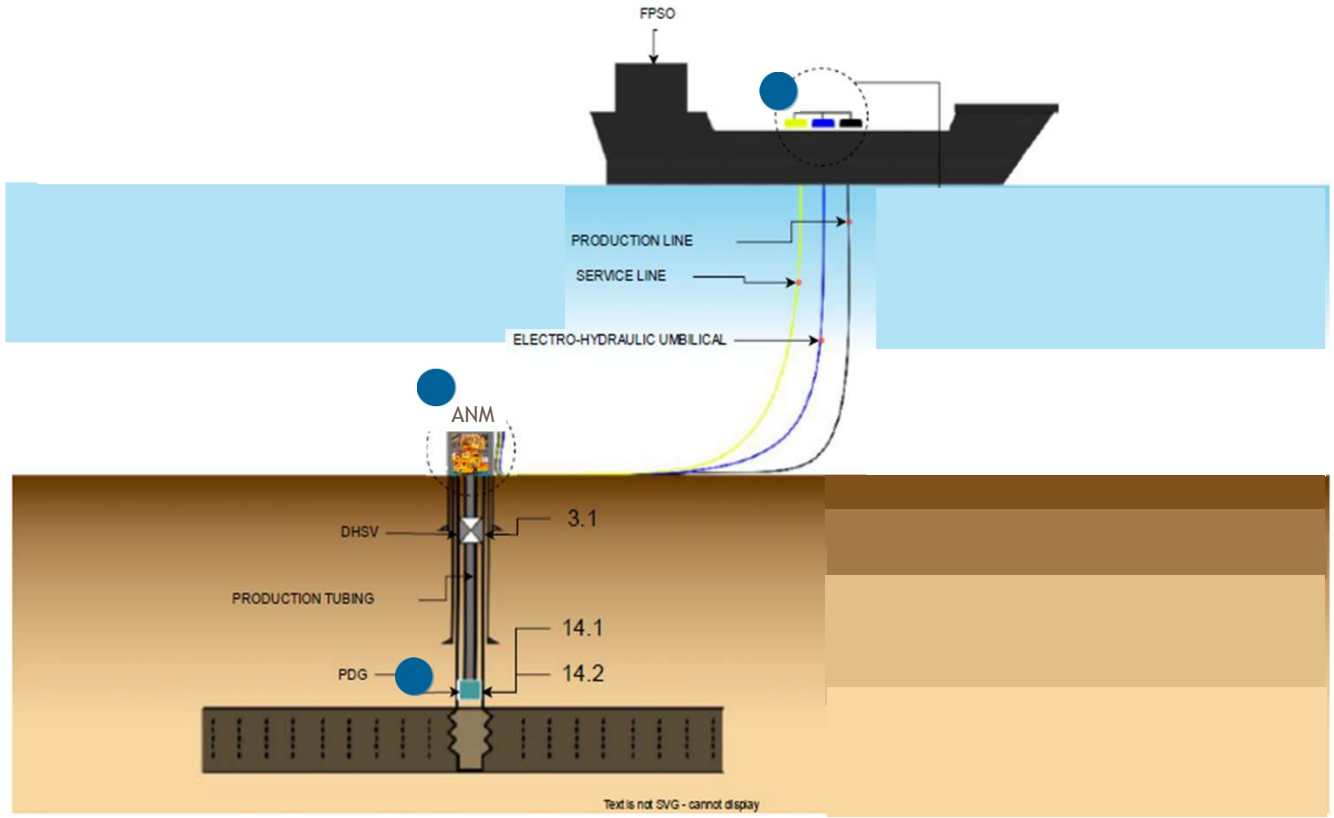
October 20, 2025

Agenda

1. Oil and gas production - overview
2. What's inside the 3W dataset
3. Types of events
4. Statistics from the 3W dataset
5. Examples of events and labels
6. 3W Dataset 2.0.0
7. Downloading parquet files on Matlab
8. Working with Python
9. Challenges

Oil and gas production - overview

Production platform, the well, the subsea Christmas tree, the production and service lines, the electro-hydraulic umbilical, as well as sensors and valves.




FMC
FC268ME

Oil and gas production - overview

- All instances, regardless of their type, are related to satellite-type offshore oil-producing wells that operate without manifold.
- The natural method is employed when the reservoir pressure is sufficient to produce hydrocarbons at a commercially viable rate without the need for additional energy.
- When reservoir pressure is insufficient, an artificial lifting method is required to introduce extra energy into the system and maintain production.

What's inside 3W dataset

PETROBRAS

Important definitions:

Term	Meaning
Samples	measured values from the sensors, one each second.
Time series	a collection of samples of a given variable
Instances	a collection of time series. All instances have the same number of time series.
Class label	characterizes normality or an unwanted event (nine in all)
State label	Operational status of the well

What's inside 3W dataset

Class labels and their codes

Class label	Steady State Code	Transient code
Normality	0	-
Abrupt Increase of BSW	1	101
Spurious Closure of DHSV	2	102
Severe Slugging	3	-
Flow Instability	4	-
Rapid Productivity Loss	5	105
Quick Restriction in PCK	6	106
Scaling in PCK	7	107
Hydrate in Production Line	8	108
Hydrate in Service Line	9	109

What's inside 3W dataset

State labels and their codes

State label	Code
Open	0
Shut-In	1
Flushing diesel	2
Flushing gas	3
Bullheading	4
Closed with diesel	5
Close with gas	6
Restart	7
Depressurization	8

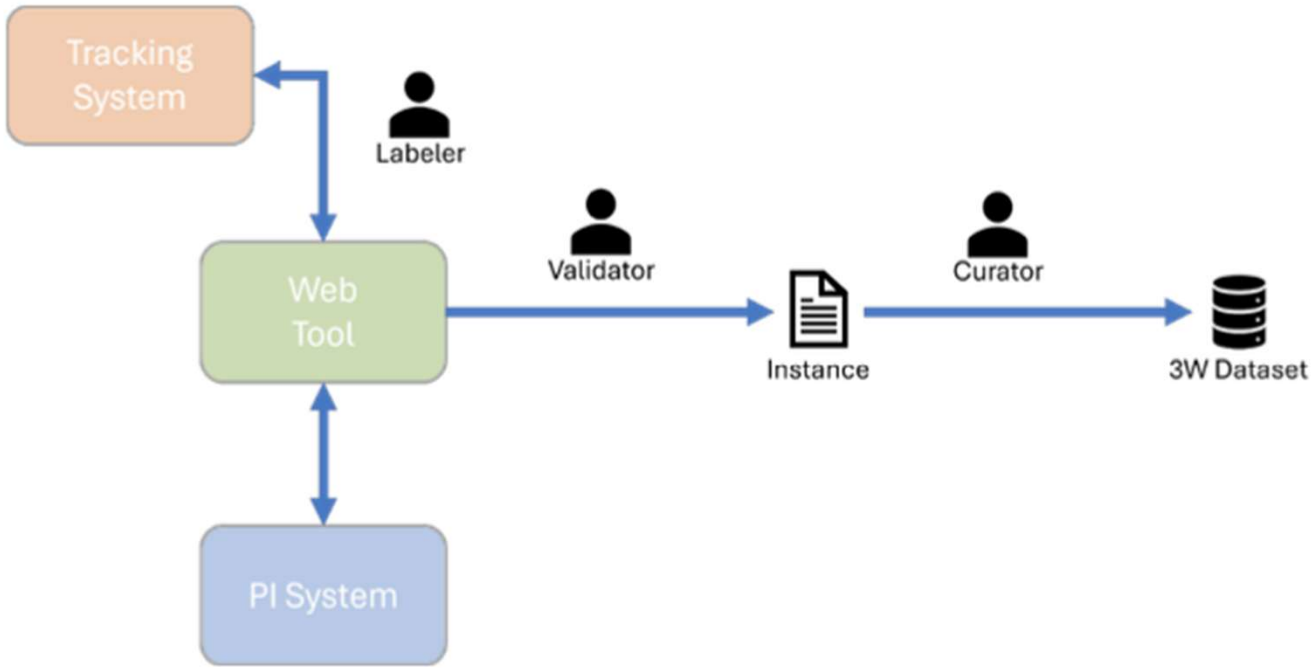
What's inside 3W dataset

- The objective of the dataset is to characterize normal behavior and unwanted events using available sensor measurements of temperature and pressure, installed in the lines and equipment.

- These signals are measured continuously and are stored in databases. A team of experts analyzed this data, relating unwanted events to data segments.

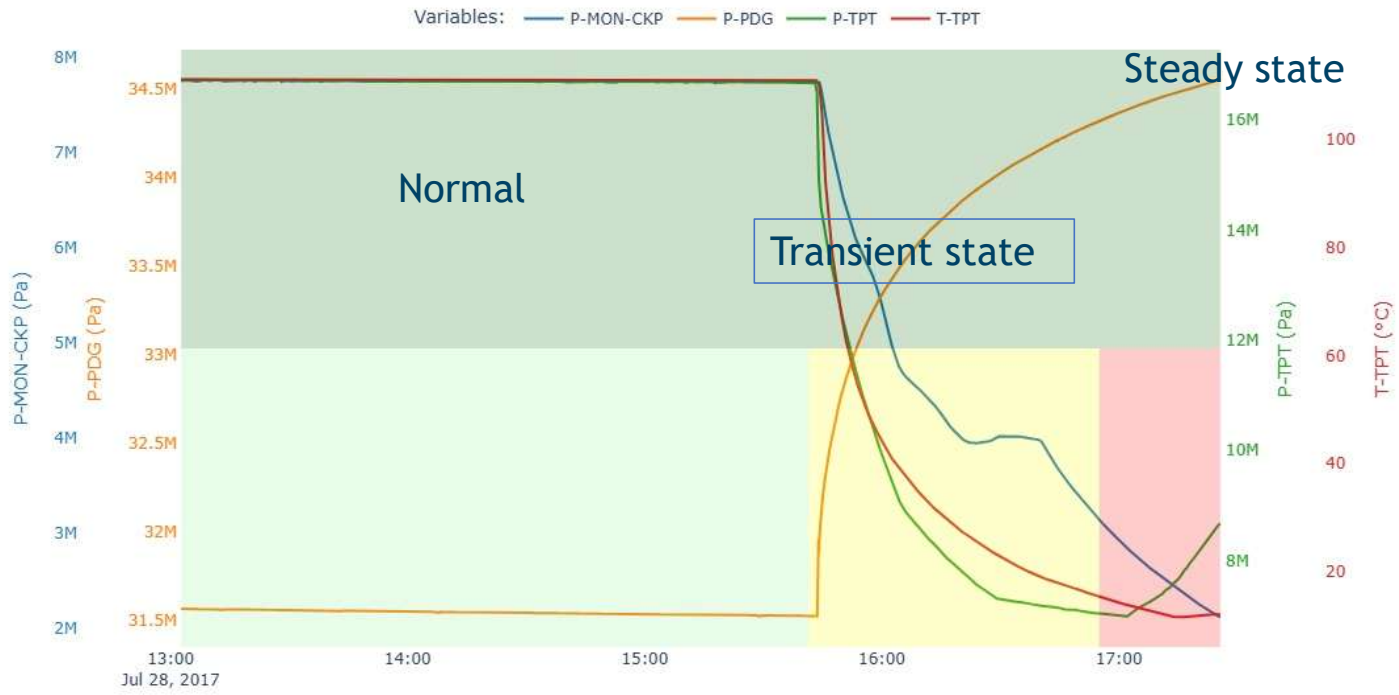
What's inside 3W dataset

The procedure to add new instances in the 3W dataset is shown below.



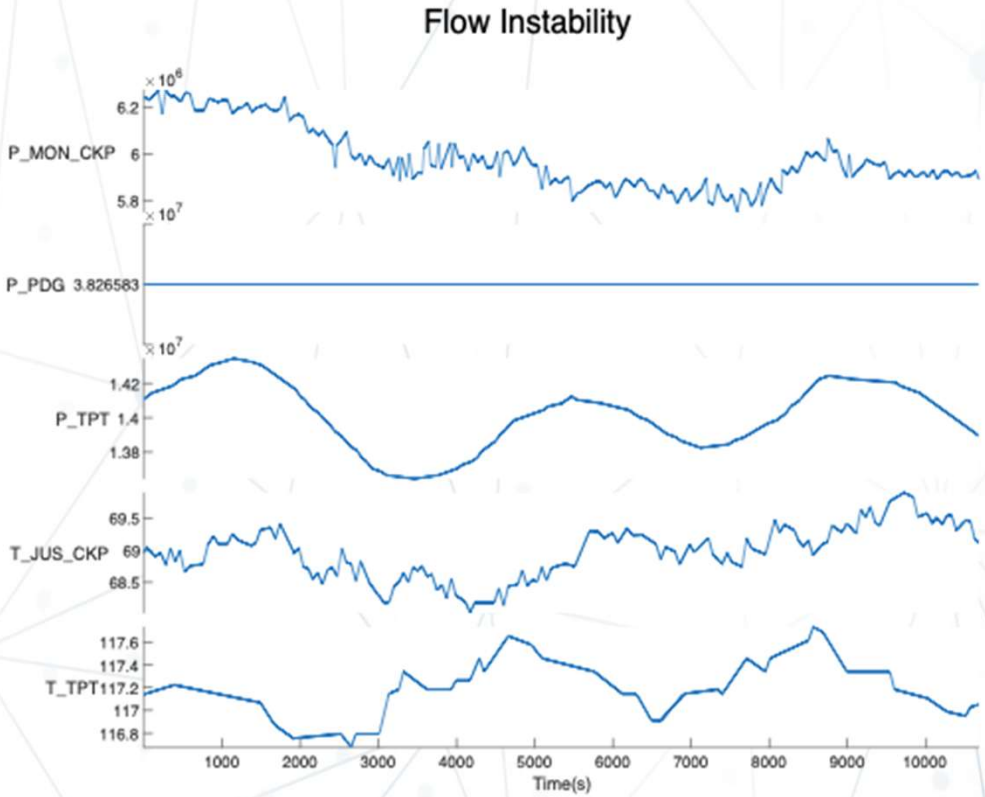
What's inside 3W dataset

This figure shows two measurements of pressure and one measurement of temperature when the status changes from normality to an unwanted event.



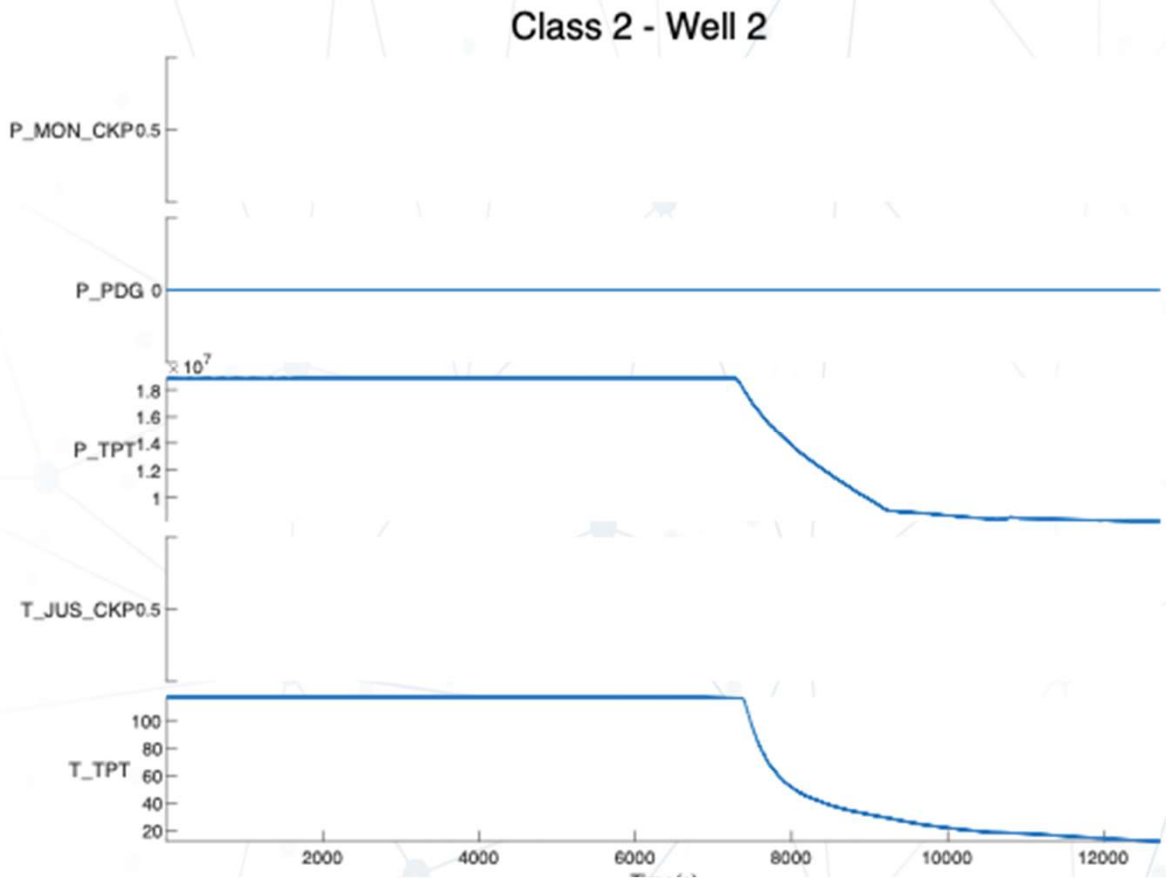
What's inside 3W dataset

The classes flow instability (4) and severe slugging (3) do not start with normal operation, as shown below.



What's inside 3W dataset

The number of available time series in the instances varies, since the sensors are often unavailable.



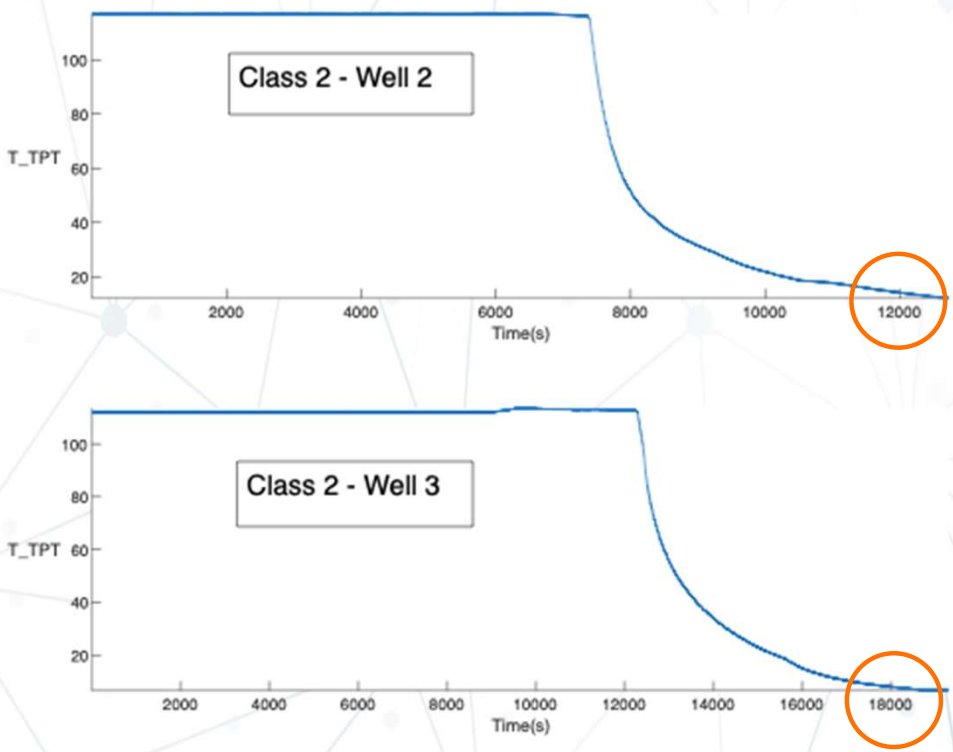
NaN (unavailable)

Frozen

NaN (unavailable)

What's inside 3W dataset

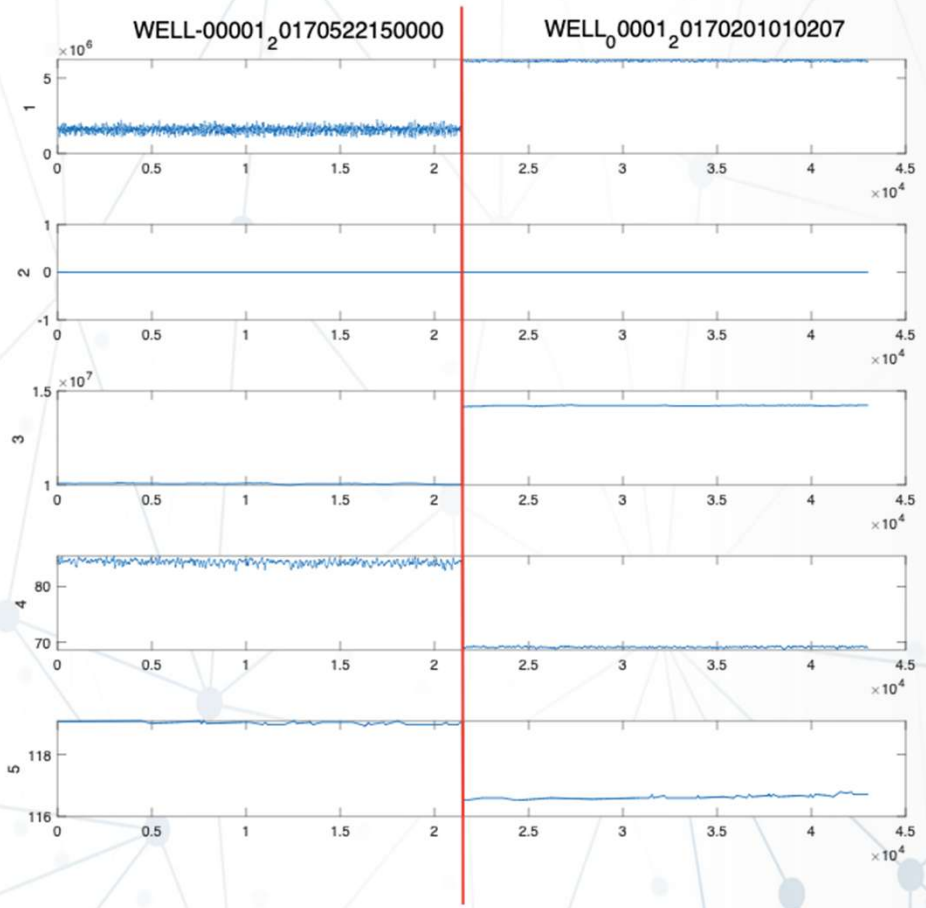
The time series have different length in different instances



What's inside 3W dataset

Two instances of class 0 (normality) for the same well.

The point of operation changes!



What's inside 3W dataset

What’s inside "3W dataset"?

- It contains 3 different sources (**real, simulated, and hand-drawn**) representing undesirable events that occur in oil **Wells**.
- Some of these events are (fortunately) very rare, with few instances available.
- This fact motivated the generation of simulation data.
- However, some events require a lot of computational effort and for this reason they were drawn by hand.

What's inside 3W dataset

Available instances:

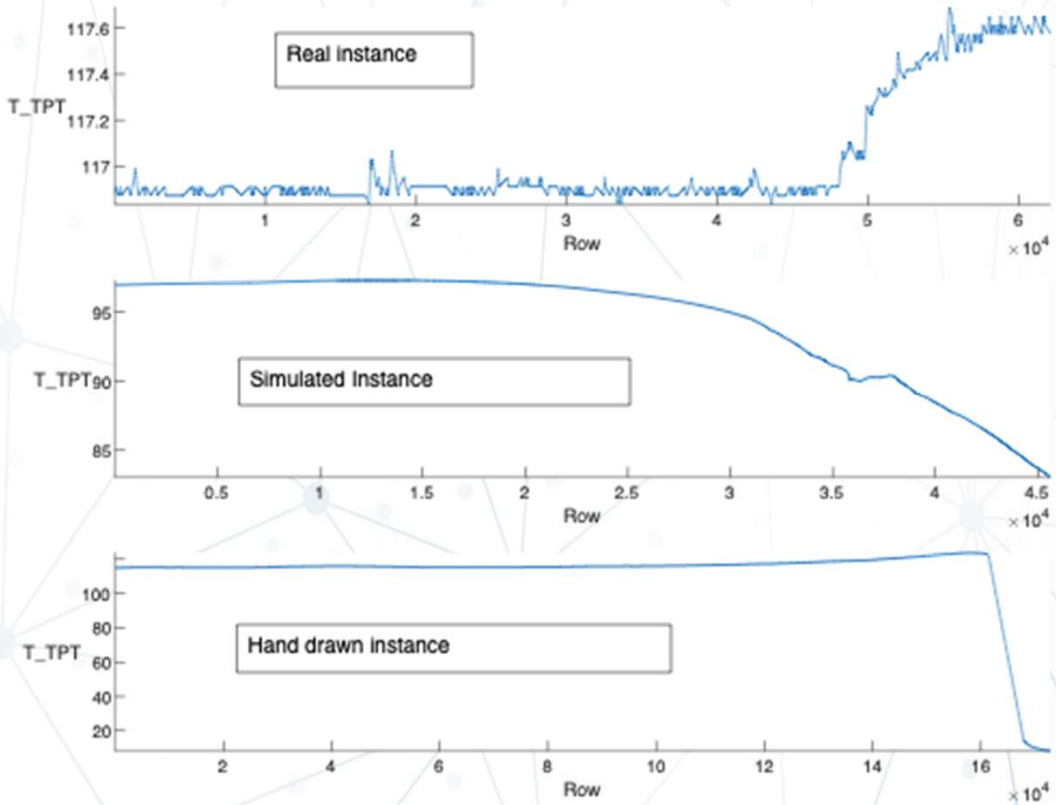
Real instances: preservation of real data characteristics, labeling by experts, and validation by expert committee;

Simulated instances: simulation models calibrated by experts, and systematized labeling;

Hand-drawn instances: hand-drawn graphs by experts, and systematized labeling.

What's inside 3W dataset

The figure shows the same variable for the event of class 1 in a real, simulated and hand drawn instance.



Noisy

Noiseless

Types of Events

What's inside 3W dataset

- **Abrupt Increase of BSW**: a sudden rise in Basic Sediment and Water ratio, potentially leading to flow assurance issues, reduced oil production, and operational challenges;
- **Spurious Closure of DHSV**: an unexpected closure of Downhole Safety Valve, risking production losses;
- **Severe Slugging**: periodic and intense flow instability that can damage equipment and disrupt operations;
- **Flow Instability**: non-periodic flow disturbances that, if unaddressed, may escalate into severe slugging;
- **Rapid Productivity Loss**: a sudden decrease in well productivity, often driven by changes in system properties;

Types of Events

What's inside 3W dataset

- **Quick Restriction in PCK**: a rapid and significant restriction in the production choke valve, typically caused by operational challenges;
- **Scaling in PCK**: formation of inorganic deposits in the production choke valve, reducing oil and gas production;
- **Hydrate in Production Line**: formation of crystalline compounds (hydrates) that can block pipelines, leading to significant production losses and high unblocking costs;
- **Hydrate in Service Line**: similar to Hydrate in Production Line, but occurring in the service line, leading to distinct signature patterns in operational data.

Statistics from the 3W dataset

Type of Event	Real	Simulated	Hand-Drawn	Total
0 - Normal Operation	594 (597)	0	0	**594 (597)**
1 - Abrupt Increase of BSW	4 (5)	114	10	**128 (129)**
2 - Spurious Closure of DHSV	22	16	0	**38**
3 - Severe Slugging	32	74	0	**106**
4 - Flow Instability	343 (344)	0	0	**343 (344)**
5 - Rapid Productivity Loss	11 (12)	439	0	**450 (451)**
6 - Quick Restriction in PCK	6	215	0	**221**
7 - Scaling in PCK	36 (4)	0	10	**46 (14)**
8 - Hydrate in Production Line	14 (3)	81	0	**95 (84)**
9 - Hydrate in Service Line	57 (0)	150 (0)	0	**207 (0)**
Total	**1119 (1025)**	**1089 (939)**	**20**	**2228 (1984)**

Inside parentheses, the number of instances in the previous versions.

Statistics from the 3W dataset

The 3W dataset's main statistics related to inherent difficulties of real data.

Statistic	Amount	Percentage
Missing Variables	41109	65.90% of 62384
Frozen Variables	6095	9.77% of 62384
Unlabeled Observations	4028400	5.26% of 76587318

Statistics from the 3W dataset

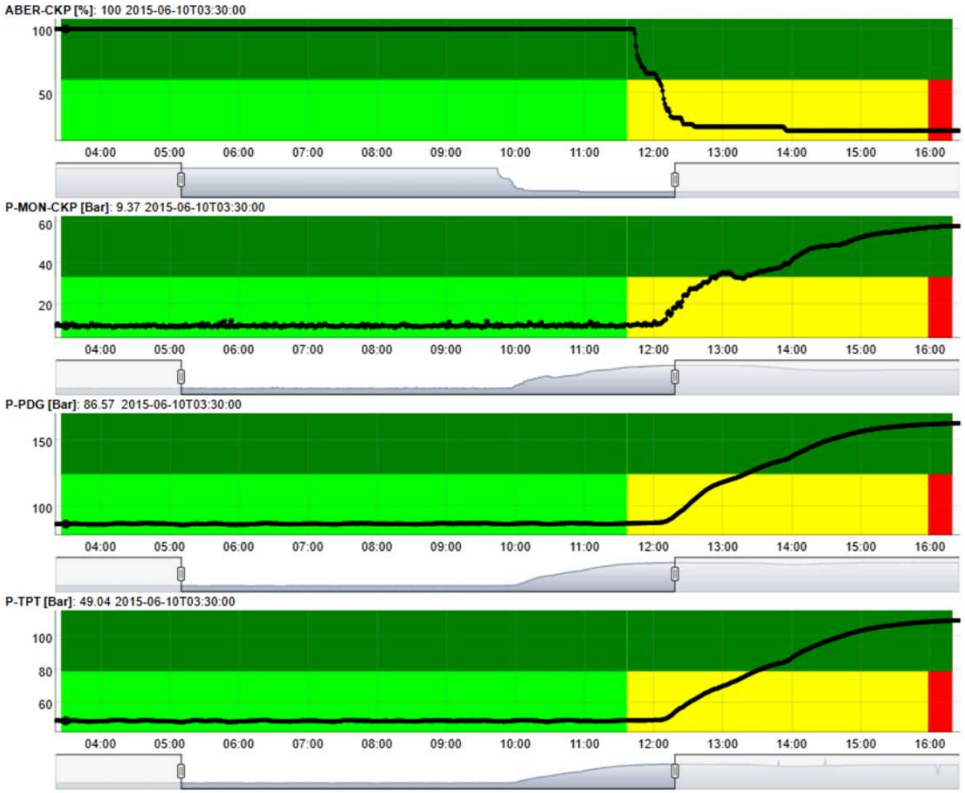
Frozen and missing values in 5 process variables of the 8 classes

Variable	Class	Frozen	Missing Values
0	1	3.10%	0.00%
	2	15.79%	0.00%
	8	1.19%	0.00%
1	1	0.00%	0.00%
	2	0.00%	0.00%
	8	1.19%	0.00%
2	1	0.00%	0.00%
	2	0.00%	0.00%
	8	1.19%	0.00%
3	1	0.00%	0.00%
	2	0.00%	36.84%
	8	0.00%	0.00%
4	1	0.78%	0.00%
	2	2.63%	50.00%
	8	0.00%	3.57%

Examples of events and labels

Real Quick Restriction in PCK

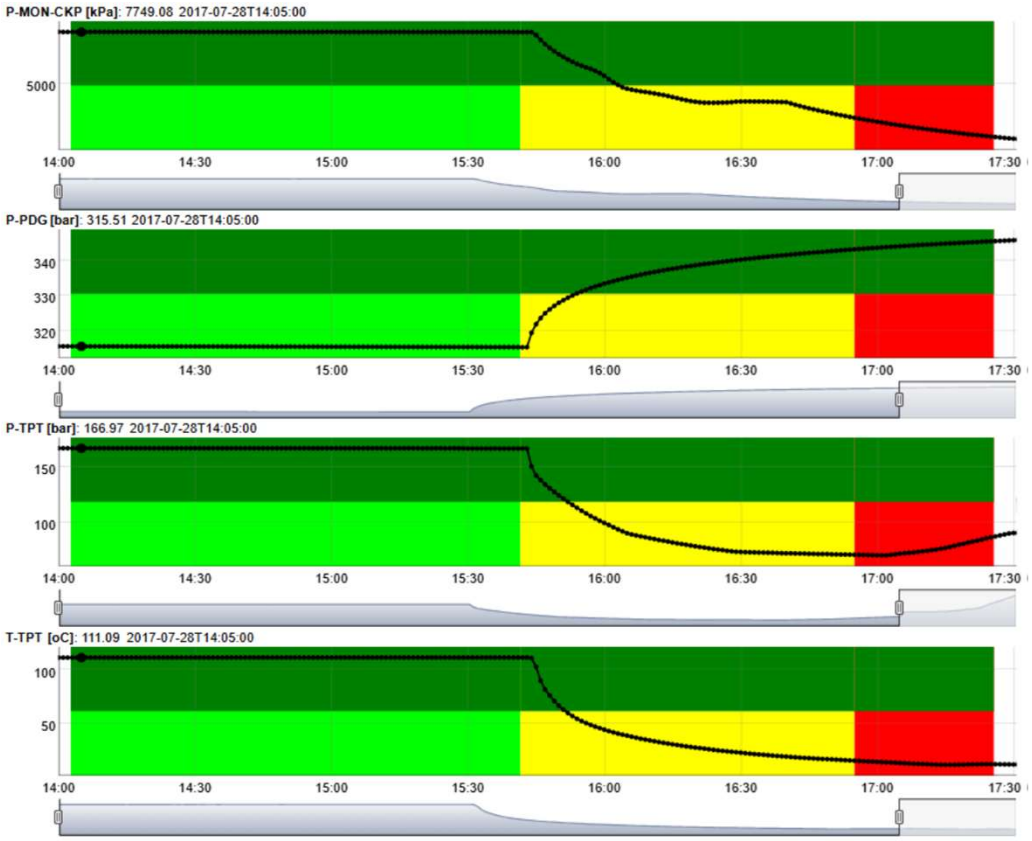
- Pressure increases are observed in sensors located in TPT, PDG, and MON-CKP.
- This example has 3 periods which were labeled with the Petrobras' Web Tool as follows:
  - light green (1st period = Normal Operation): class label = 0;
  - yellow (2nd period = Transient Condition): class label = 106;
  - red (3rd period = Steady State): class label = 6; dark green (Open): state label = 0.



Examples of events and labels

Real Spurious Closure of DHSV

- Pressure increase is observed in sensors located upstream of the DHSV (PDG) and pressure decreases are observed in sensors located downstream of the DHSV (TPT and MON-CKP).
- This example has 3 periods which were labeled with the Petrobras' Web Tool as follows:
  - light green (1st period = Normal Operation): class label = 0;
  - yellow (2nd period = Transient Condition): class label = 102;
  - red (3rd period = Steady State): class label = 2; dark green (Open): state label = 0.



3W Dataset 2.0.0.

Diagram representing the considered scenario

Name	Description	Position
ABER-CKGL	Opening of the GLCK (gas lift choke)	1.1
ABER-CKP	Opening of the PCK (production choke)	2.1
ESTADO-DHSV	State of the DHSV (downhole safety valve)	3.1
ESTADO-M1	State of the PMV (production master valve)	4.1
ESTADO-M2	State of the AMV (annulus master valve)	5.1
ESTADO-PXO	State of the PXO (pig-crossover) valve	6.1
ESTADO-SDV-GL	State of the gas lift SDV (shutdown valve)	7.1
ESTADO-SDV-P	State of the production SDV	8.1
ESTADO-W1	State of the PWV (production wing valve)	9.1
ESTADO-W2	State of the AWV (annulus wing valve)	10.1
ESTADO-XO	State of the XO (crossover) valve	11.1
P-ANULAR	Pressure in the well annulus	12.1
P-JUS-BS	Downstream pressure of the SP (service pump)	13.1
P-JUS-CKGL	Downstream pressure of the GLCK	1.2
P-JUS-CKP	Downstream pressure of the PCK	2.2
P-MON-CKGL	Upstream pressure of the GLCK	1.3

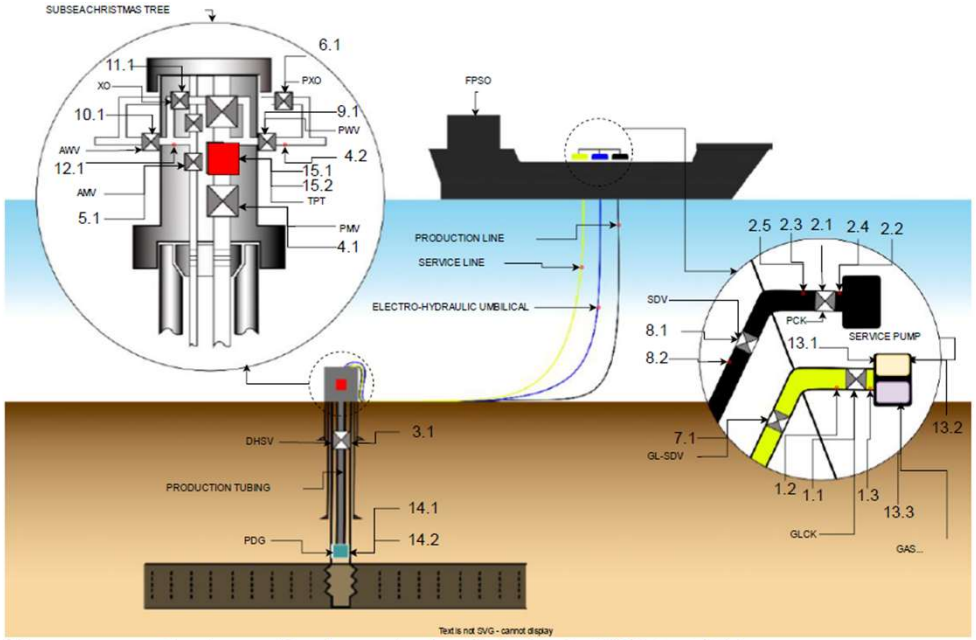


Diagram representing the considered scenario when designing the 3W Dataset 2.0.0.

3W Dataset 2.0.0

- The data itself is organized into subdirectories dedicated to each type of event. The name of each directory corresponds to the code associated with each type of event.
- Each instance is stored in its own Apache Parquet file, or simply Parquet file. Parquet is an open source, column-oriented data file format designed for efficient data storage and retrieval.
- The logic used to formulate file names depends on the type of instance.
- Real Instances: File names follow the format WELL-[incremental ID]_[timestamp of oldest observation].parquet. Example: WELL-00014_20170917140000.parquet;
- Each real well is associated with a unique ID, regardless of the type of event (subdirectory). As each real well can give rise to one or multiple instances, the [timestamp of oldest observation] ensures unique identification.
- Simulated Instances: File names follow the format SIMULATED_[incremental ID].parquet. Example: SIMULATED_00072.parquet;

3W Dataset 2.0.0

- Hand-Drawn Instances: File names follow the format DRAWN_[incremental ID].parquet. Example: DRAWN_00007.parquet;

- The incremental ID starts at 1 for each type of event (subdirectory) and is sufficient to uniquely identify all hand-drawn instances.

- The timestamp vector of each instance is used as the index in the corresponding Parquet file. All timestamps are represented in the format 'YYYY-MM-DD HH:MM:SS'.

3W Dataset 2.0.0

- In the root of the directory containing the dataset, there is a file called dataset.ini, which specifies properties of the 3W Dataset 2.0.0. The proposal is that all users concentrate their searches for these properties in this file.

- The data itself is organized into subdirectories dedicated to each type of event. The name of each directory corresponds to the code associated with each type of event

3W Dataset 2.0.0

- All instances are now saved in Parquet files (created with the `pyarrow` engine and `brotli` compression);
- Reduction in disk space occupied by the 3W Dataset of 3.15 GB (from 4.89 GB to 1.74 GB);
- Real and simulated instances of type 9 were added;
- Several instances of types 0, 3, 4, 5, 6 and 8 were added;
- Another 24 real wells were covered with new real instances (now 42 real wells are covered);
- Some real instances, mainly of type 1, were removed;
- 1 variable was removed (`T-JUS-CKGL`);
- Another 20 variables were added (there are now 27 variables);
- Another label referring to well operational status was added;

3W Dataset 2.0.0

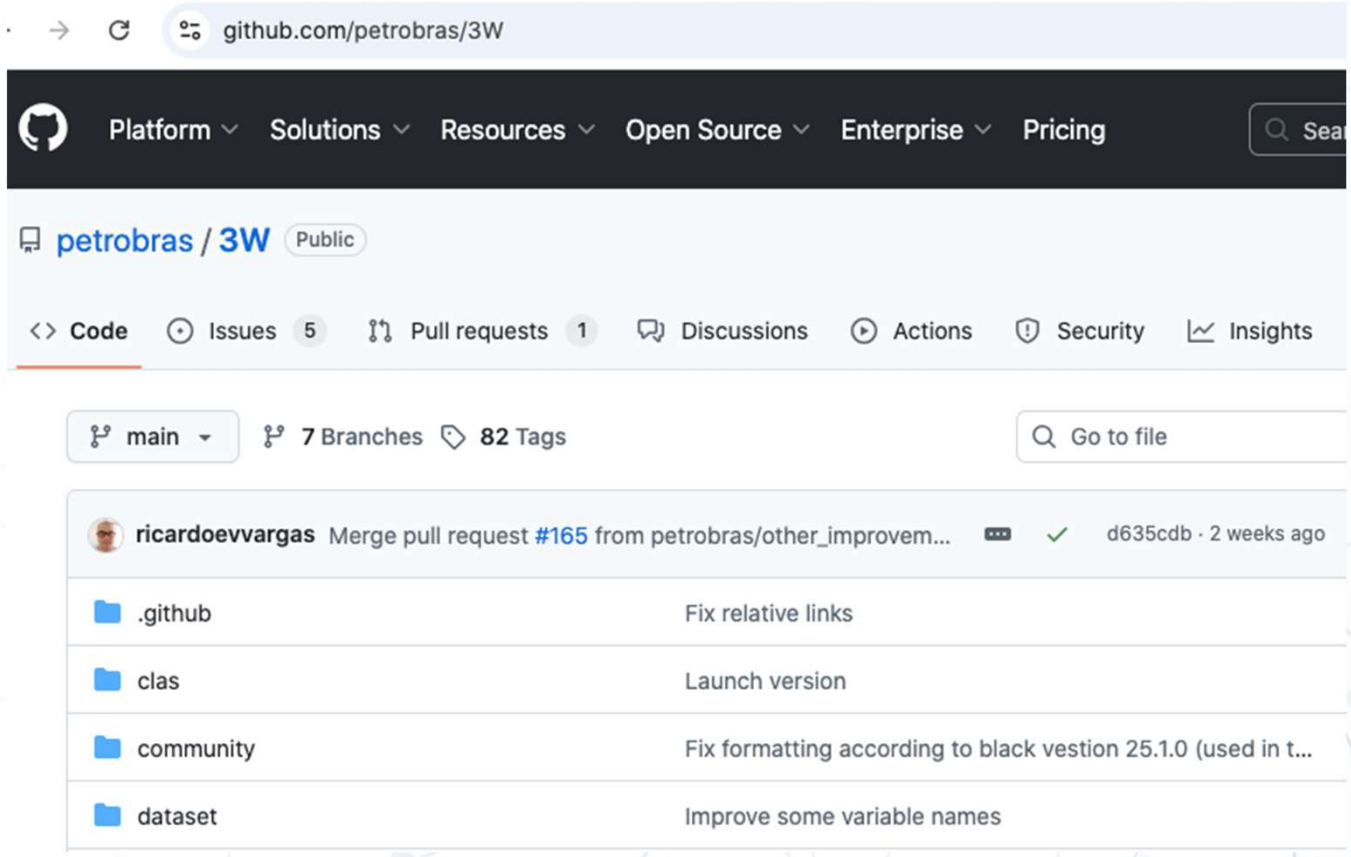
PETROBRAS

- Normal periods in several real instances with unwanted events were extended;
- All labeling gaps in real instances were eliminated (all observations were labeled);
- Conversions between measurement units in several instances were corrected;
- Labels in several real instances were adjusted by experts;
- All values of some variables in some real instances were corrected due to corrections in historian systems' tag configurations;
- Certain variable values have undergone minimal change due to different rounding;
- The 3W Dataset's main configuration file (dataset.ini) was updated.

Important: for the 20 added variables, NaN values were assigned for the instances already defined in the previous versions.

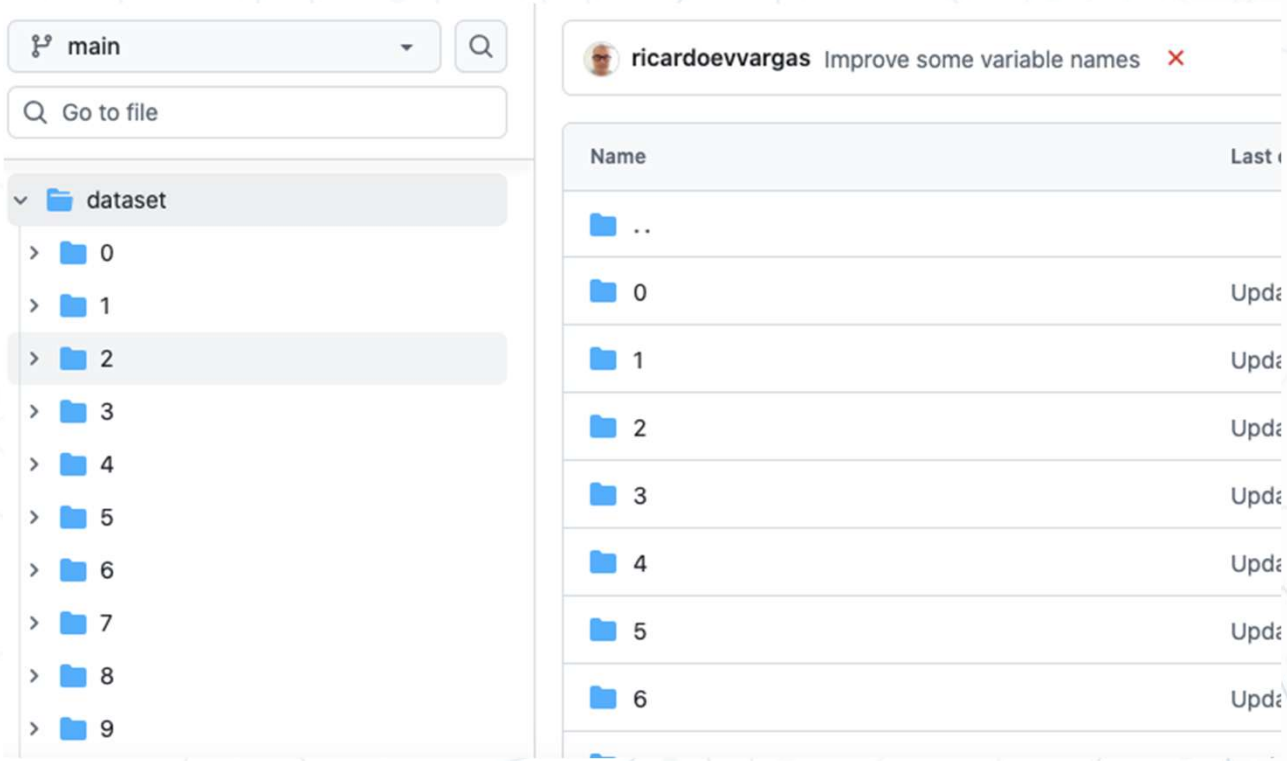
Downloading parquet files in Matlab

Access the dataset on Github



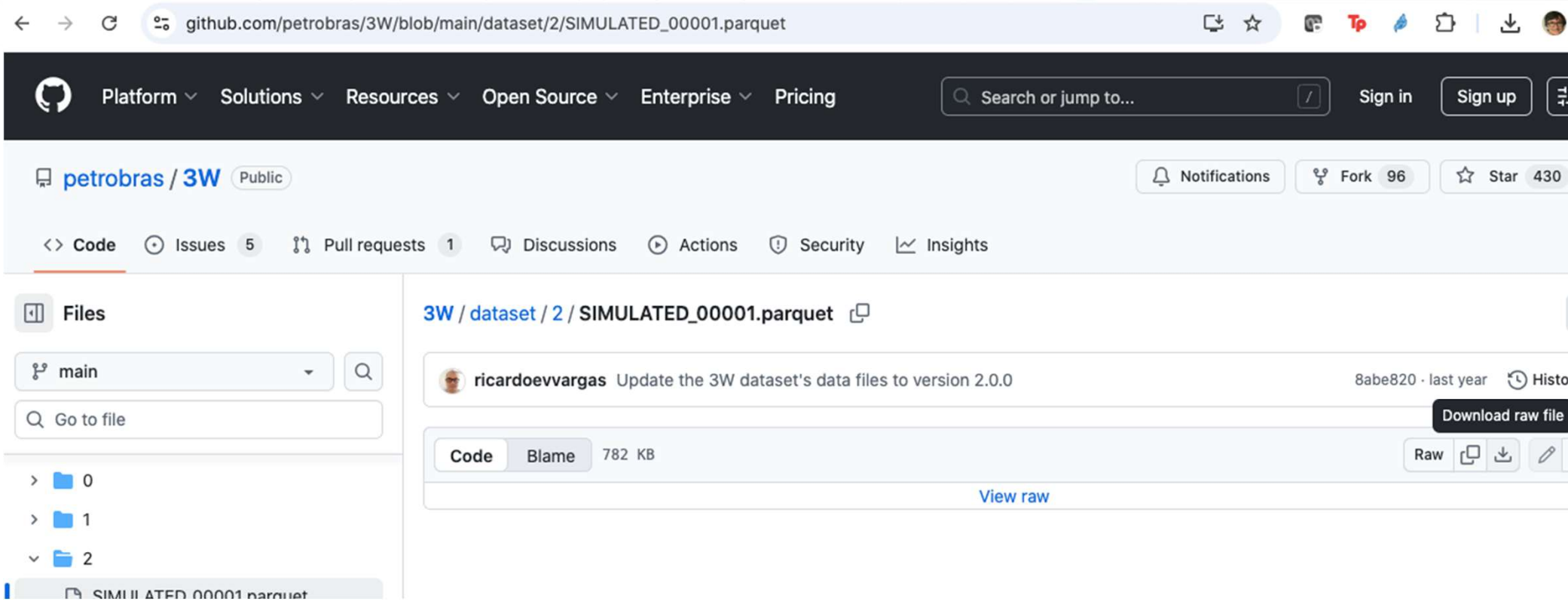
Downloading parquet files in Matlab

Access the desired classes and instances



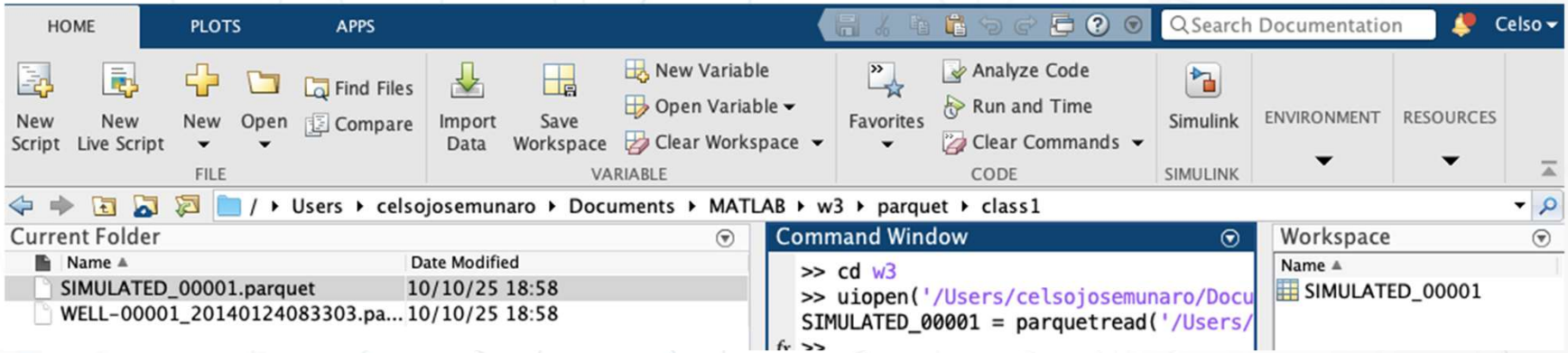
Downloading parquet files in Matlab

Select the desired parquet file and click download



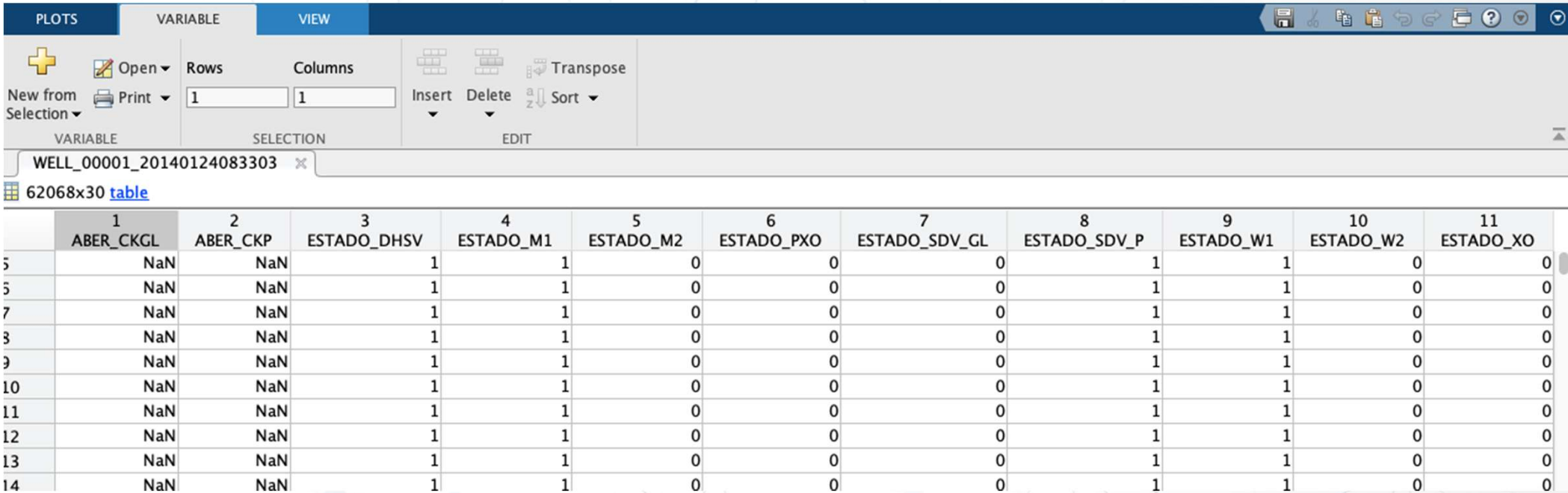
Downloading parquet files in Matlab

Drag the parquet file to command window and a table is produced in the Workspace.



Downloading parquet files in Matlab

Click in the variable to access all columns of the table



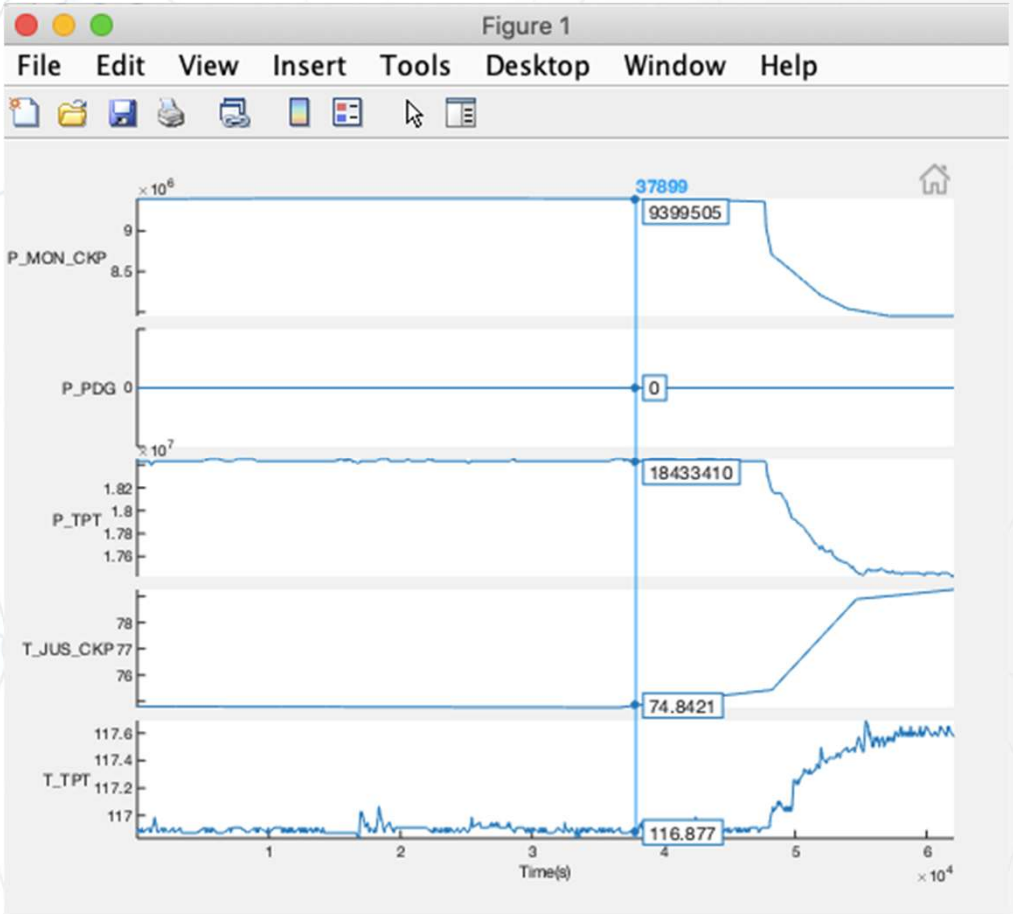
	1 ABER_CKGL	2 ABER_CKP	3 ESTADO_DHSV	4 ESTADO_M1	5 ESTADO_M2	6 ESTADO_PXO	7 ESTADO_SDV_GL	8 ESTADO_SDV_P	9 ESTADO_W1	10 ESTADO_W2	11 ESTADO_XO
5	NaN	NaN	1	1	0	0	0	1	1	0	0
6	NaN	NaN	1	1	0	0	0	1	1	0	0
7	NaN	NaN	1	1	0	0	0	1	1	0	0
8	NaN	NaN	1	1	0	0	0	1	1	0	0
9	NaN	NaN	1	1	0	0	0	1	1	0	0
10	NaN	NaN	1	1	0	0	0	1	1	0	0
11	NaN	NaN	1	1	0	0	0	1	1	0	0
12	NaN	NaN	1	1	0	0	0	1	1	0	0
13	NaN	NaN	1	1	0	0	0	1	1	0	0
14	NaN	NaN	1	1	0	0	0	1	1	0	0

Downloading parquet files in Matlab

Use the command stackedplot to plot selected columns

```
T=WELL_00001_20140124083303;
stackedplot(T(:,[17 21 24 27]));
```

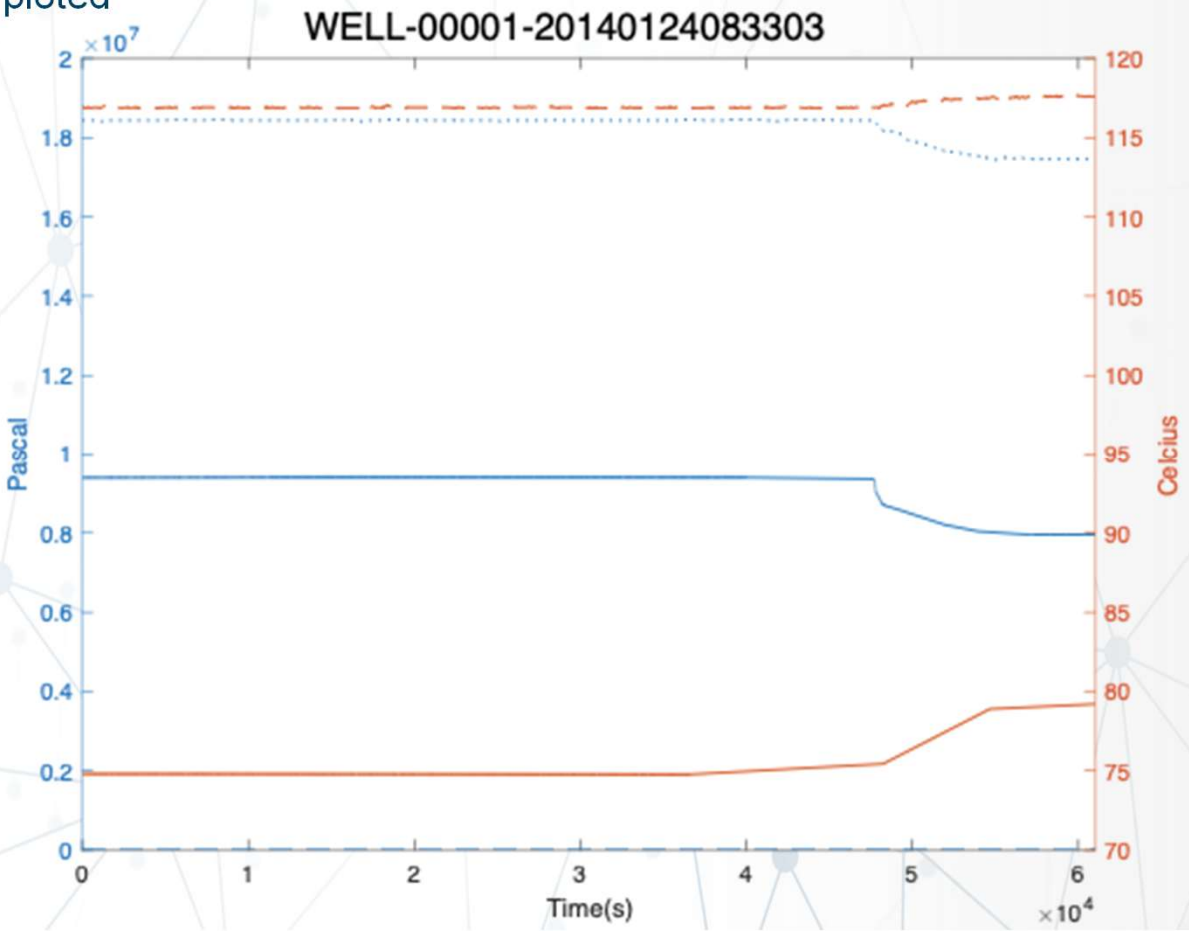
Ploting data from a real instance (class 1).



Downloading parquet files in Matlab

The table can be converted to array and ploted

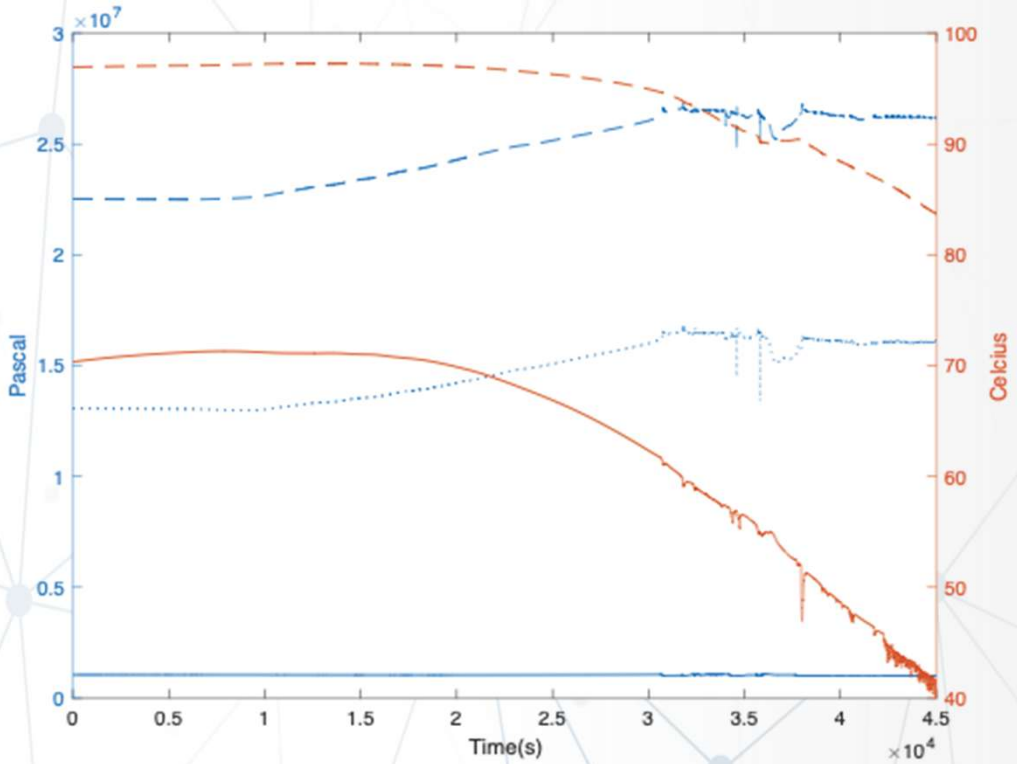
```
x=table2array(x);
yyaxis left
plot(x(:,[1 2 3 ]));
ylabel('Pascal');
yyaxis right
plot(x(:,[4 5]));
ylabel('Celcius');
```



Downloading parquet files in Matlab

Ploting data from a simulated instance (class 1)

```
T=SIMULATED_00001;
stackedplot(T(:,[17 21 24 27]));
x=table2array(x);
yyaxis left
plot(x(:,[1 2 3 ]));
ylabel('Pascal');
yyaxis right
plot(x(:,[4 5]));
ylabel('Celcius');
```



Working with Python

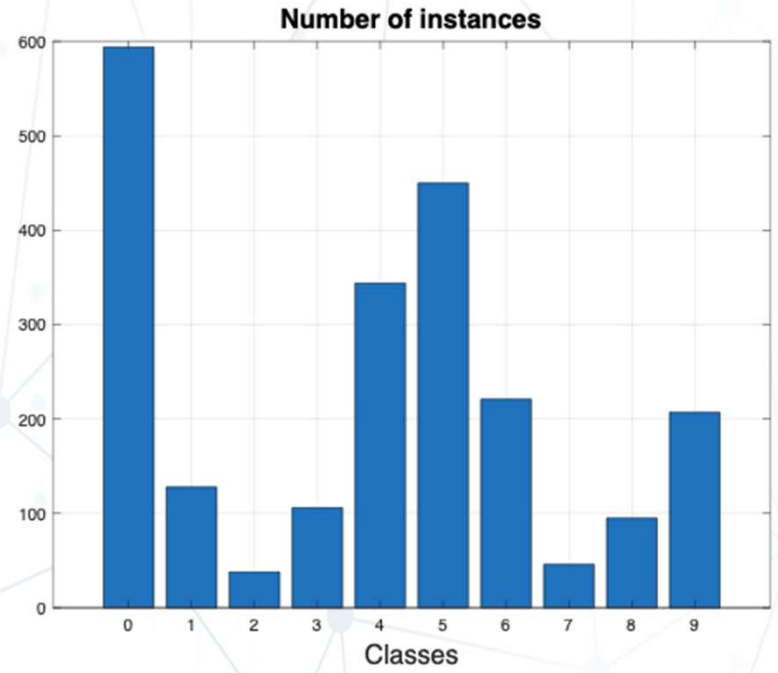
PETROBRAS

Python - Notebook

Challenges

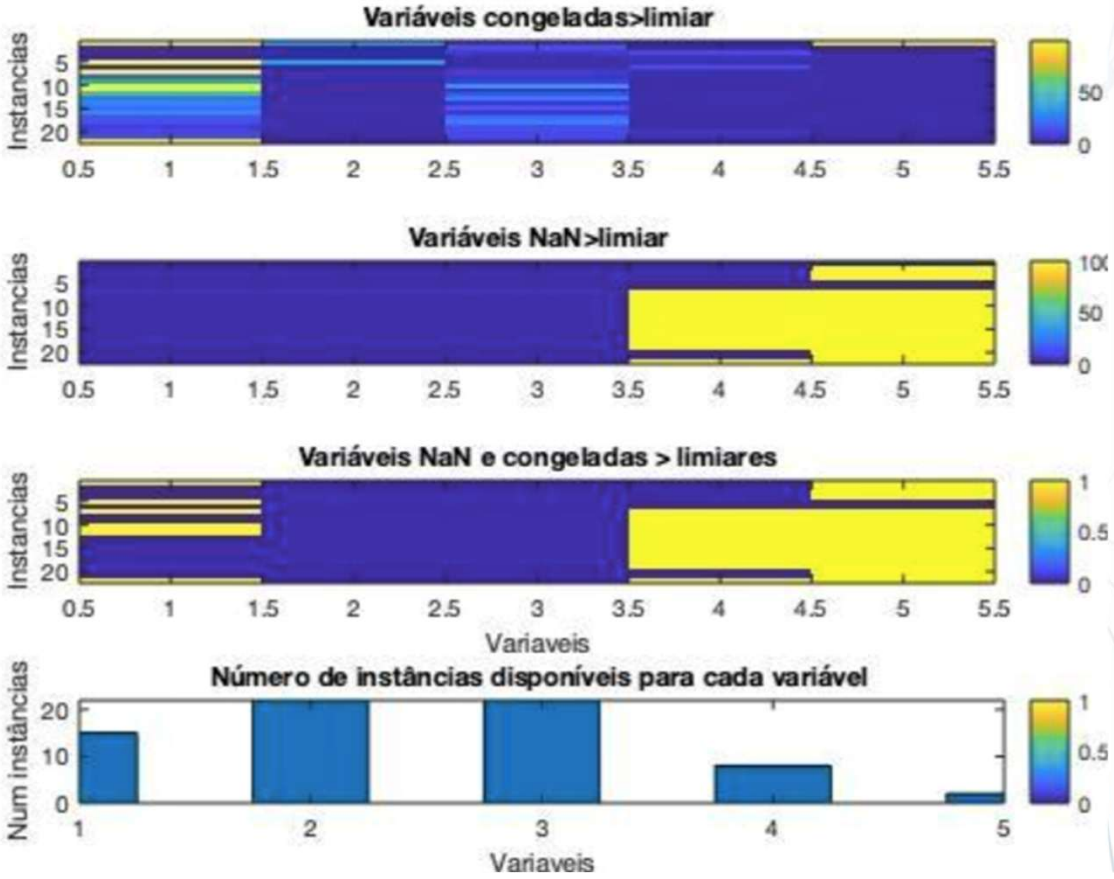
PETROBRAS

Instances per class:very unbalanced dataset.



Challenges

Real instances available for class 2, considering NaN and frozen variables

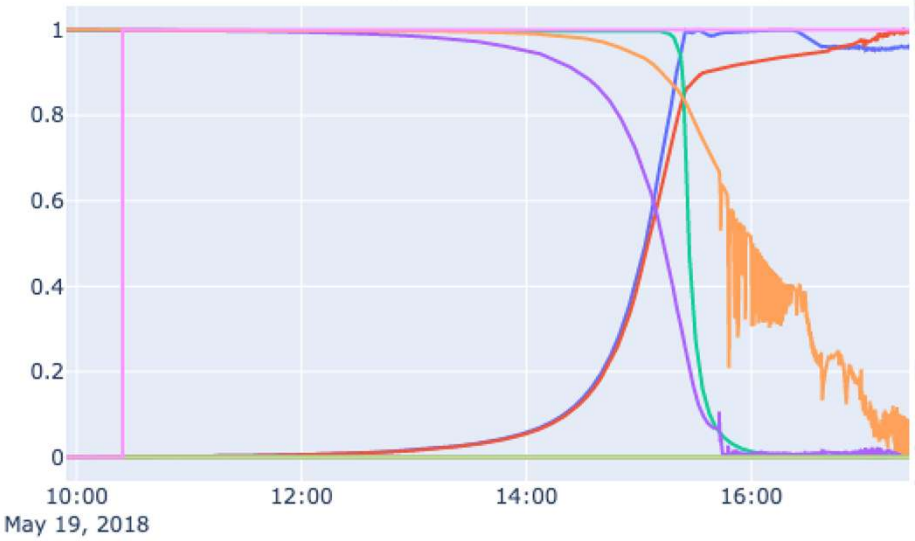


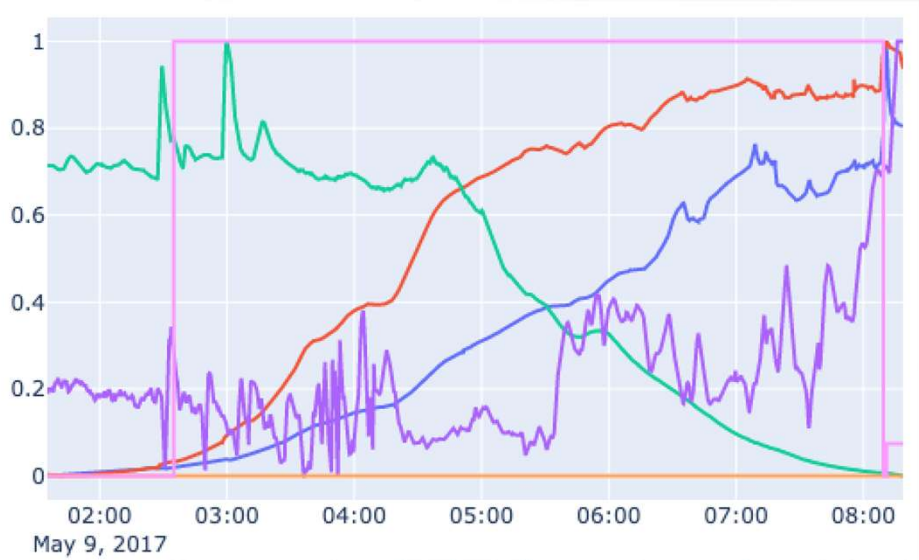
If more than 10 (from 22) real instances are required for training, only variables 1,2,3 can be used.

Challenges

PETROBRAS

Same class and different variable behaviors





Challenges

Instance of class 0 (normality) with diverse behaviors.

Selection of the instances for training, validation and test considering the available variables.

Decision about using real, simulated or hand drawn instances.

Definition of pre-processing procedures for NaN and frozen variables.

Definition of metrics suitable for the unbalanced classes.

Detection of unwanted events should work in real time with the available variables.

Thank you!