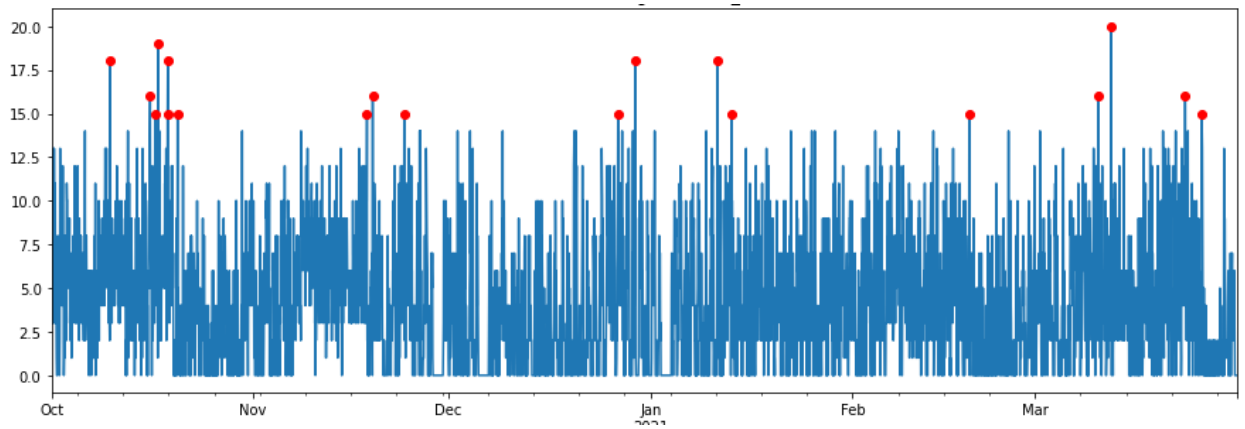
MANAGEMENT AND ANALYSIS OF PHYSICS DATASET

2020/21

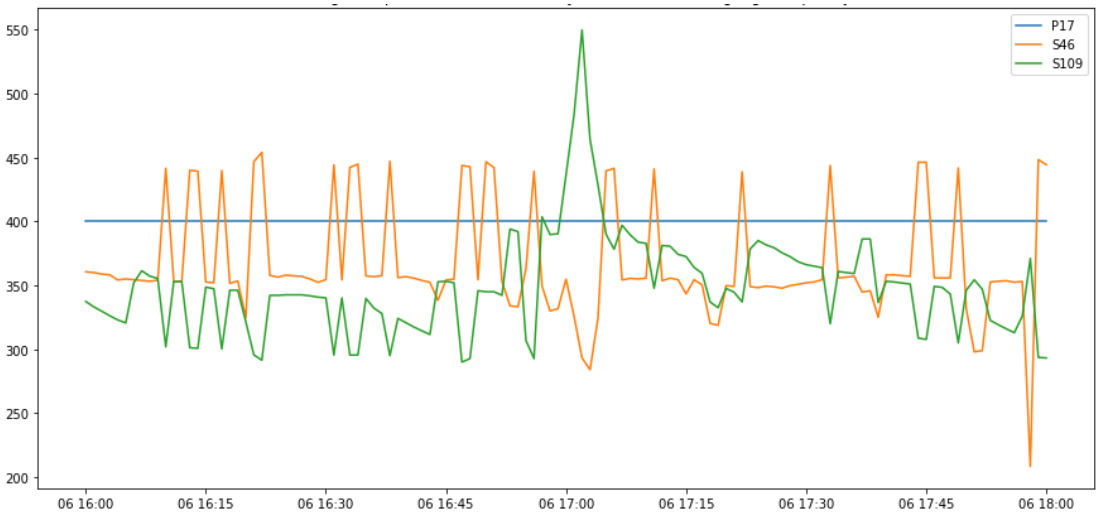
Cristiano Colpo – Mattia Sandri

Anomaly Detection

In this part, an investigation about the high frequency of ON/OFF switches on the motors is conducted. From now on, this high frequency of ON/OFF on the motors will be referred to as “anomaly”. The anomalies are searched in the data from the sensors S117, S118, S169 and S170. We need to define a numerical value for the number of turning ON / shutting OFF of the engines to be considered as an anomaly. For this purpose the 3 Sigma Rule was used: if the number of change of state of the motor is higher than $\mu + 3 \cdot \sigma$ it is considered as an anomaly. In the next image the red dots represent the anomalies, while the blue lines are the number of changes of state of the motor.



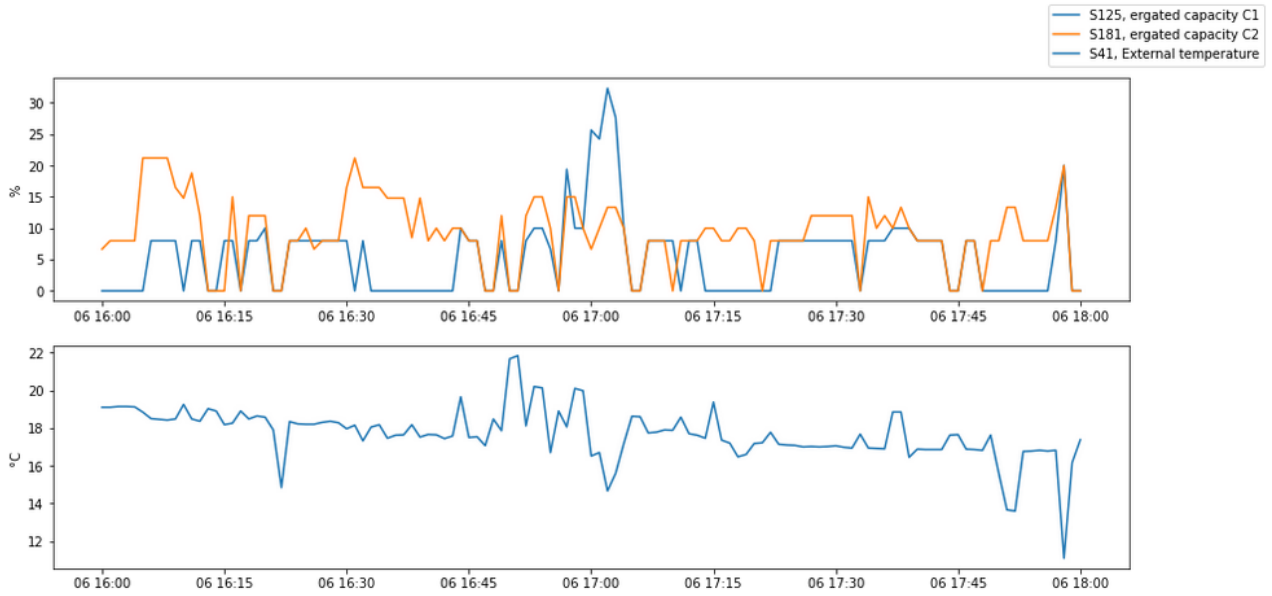
A comparison between the data regarding the anomalies and the data regarding the temperature on sensor S46 (Source water temperature) and S109 (Discharge water temperature) is performed. The hypothesis is that a bad-configured PID controller is causing the anomalies. The next image is a plot of this values during the occurring of an anomaly.



The image enforces our idea of a bad-tuned PID controller: the temperature (S46) “floats” with an oscillatory behavior under and then over the set point temperature (P17).

Anomaly Detection 2

Now the parameter to be considered is the percentage of load of the engines. The first thing is to get a visual comparison between the percentage of load of the engines and the external temperature.



The image gives a possible idea of what is the correlation between these two values: as the external temperature gets lower, the percentage of load on the engines goes up.

The Pearson Correlation Coefficient is calculated between the percentage of load on the engines and the external temperature, giving for both the engines a negative correlation. This indicates that while a value gets lower, the other one rises. This is additional evidence for the hypothesis of the engines used in a heating circuit.

The data regarding the percentage of load is then compared to the data regarding the anomalies used in the first part, but no correlation is found between them.

Predictive maintenance

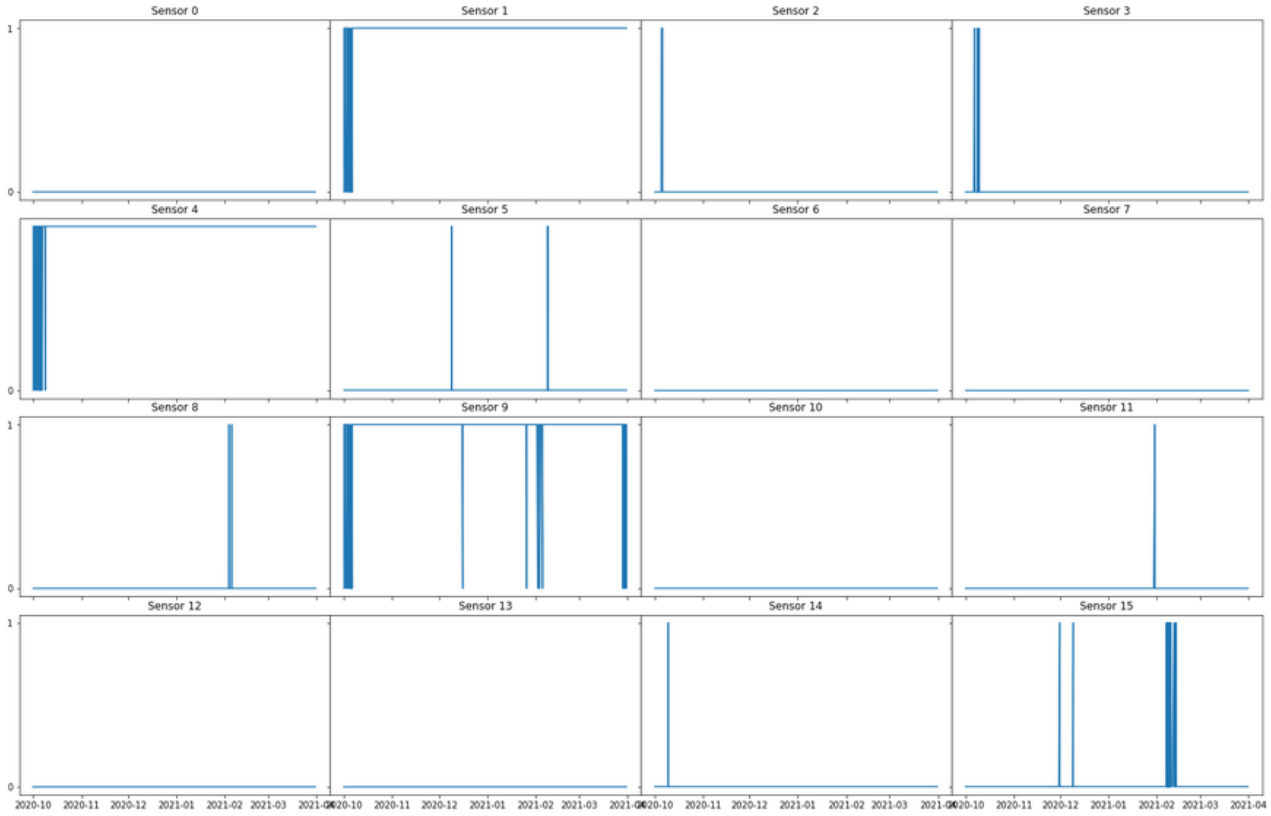
As the other two parts we use dask with the tasks that require a great amount of computation capacity and let panda manage the results, dataframes with a lower amount of rows.

So by following this philosophy we extract from the dataset only the variables that we are interested in, A5 and A9. We need to do this because in the first task we need to convert the integer values of A9 and A5 to binary strings. After selecting these sensors, as we have seen in the previous part, we perform a resampling with 1min frequency and max aggregation.

The max aggregation was chosen because if there are two samples, one with 0 and one not zero we want to define that timestamp as one that contains a problem, the value that is not zero.

The compute() method return a pandas dataframe with the data that we need to process, firstly we drop the first index (the returned pandas dataframe was a Multiindex object) and the we applied to each row a function to convert the value column from a integer to a binary string of length 16.

After converting the strings into lists of numbers we "exploded" these lists as each element created a new column named "Sensor 0" to "Sensor 15".



In order to solve task 2 we extracted the new data that we wanted to analyze, we chose the sensors based on the unit of measure because we know only for them what method to resample (mean).

We discarded the sensors with only a single unique value because a sensor that outputs only 1 value is not interesting.

Now from Sensor 6,7,8 we can see if there is overheating or not, by applying the function f to each row we can check if in a specific timestamp there is an anomaly.

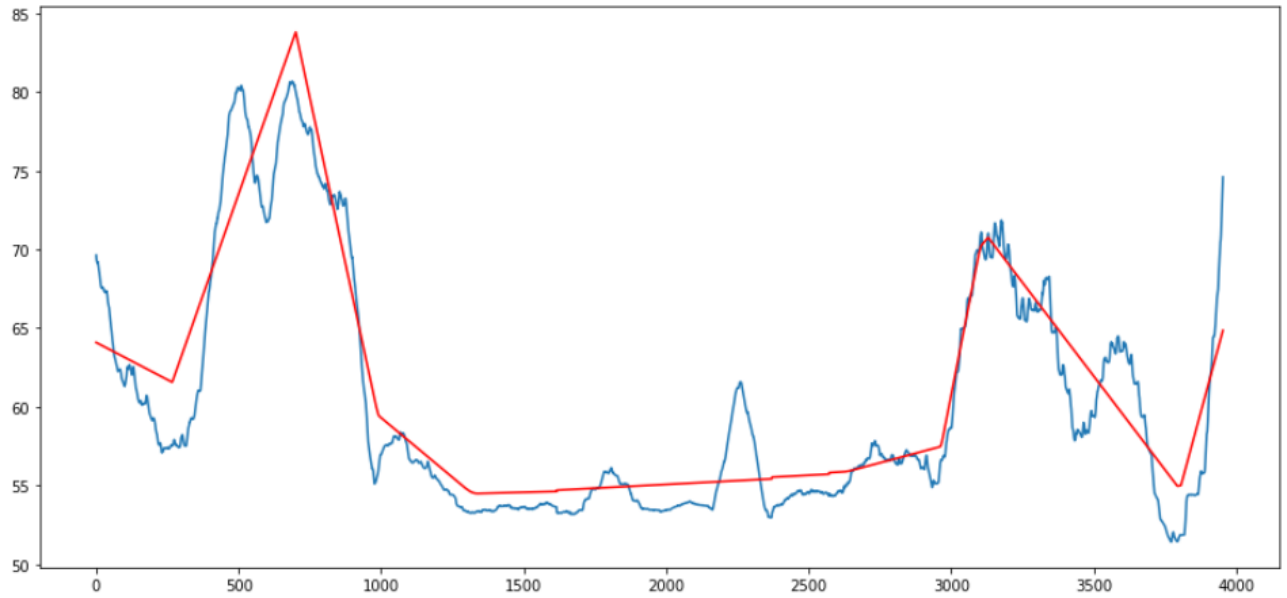
Task 2 requires checking the correlation between the overheating and the other sensors, so we compute correlation with each pair. We see that S128 has a high correlation so we use this sensor for the last task.

Task 3 requires us to predict an anomaly and in order to perform this task we need to compute two models, a classification model to classify the value returned by S128 as anomaly or not and a regressor model that predicts a future value of S128.

After associating each value of S128 to a fault {0,1} value we obtain a dataframe with 763 rows where only 11 are with value 1. This is a clear unbalanced dataset that we cannot use to train our models, so for the classification task we take a subset of the rows with fault 0, 100 rows and all the rows with value 1.

We train a classification model in order to see if there is overheating by knowing the output of S128. We need a model such a RNN that can predict the label by exploiting temporal information, but instead of using a delta-t of samples, for the sake of simplicity we use a simple MLPclassifier with single samples, the results are obviously worse.

Finally we train a MLPRegressor in order to predict a value of S128 in the future (1 april 2021), we convert the timestamp into julian date and then rescale.



We feed the classification model with the output of the regressor and we find that on April 1st 00:00 there is a chance that an anomaly will occur.