Data analysis

1st entry of project diary.

Data analysis is done. I have a few interesting conclusions to draw from it to illustrate the modelling:

On the dataset in general

Dimension of the problem

The dataset is quite good for a regression task:

- 15 input features
- Output of dimension 24
- In training set, 220,000 samples

This is enough to go to more expensive methods like neural networks.

Dataset quality

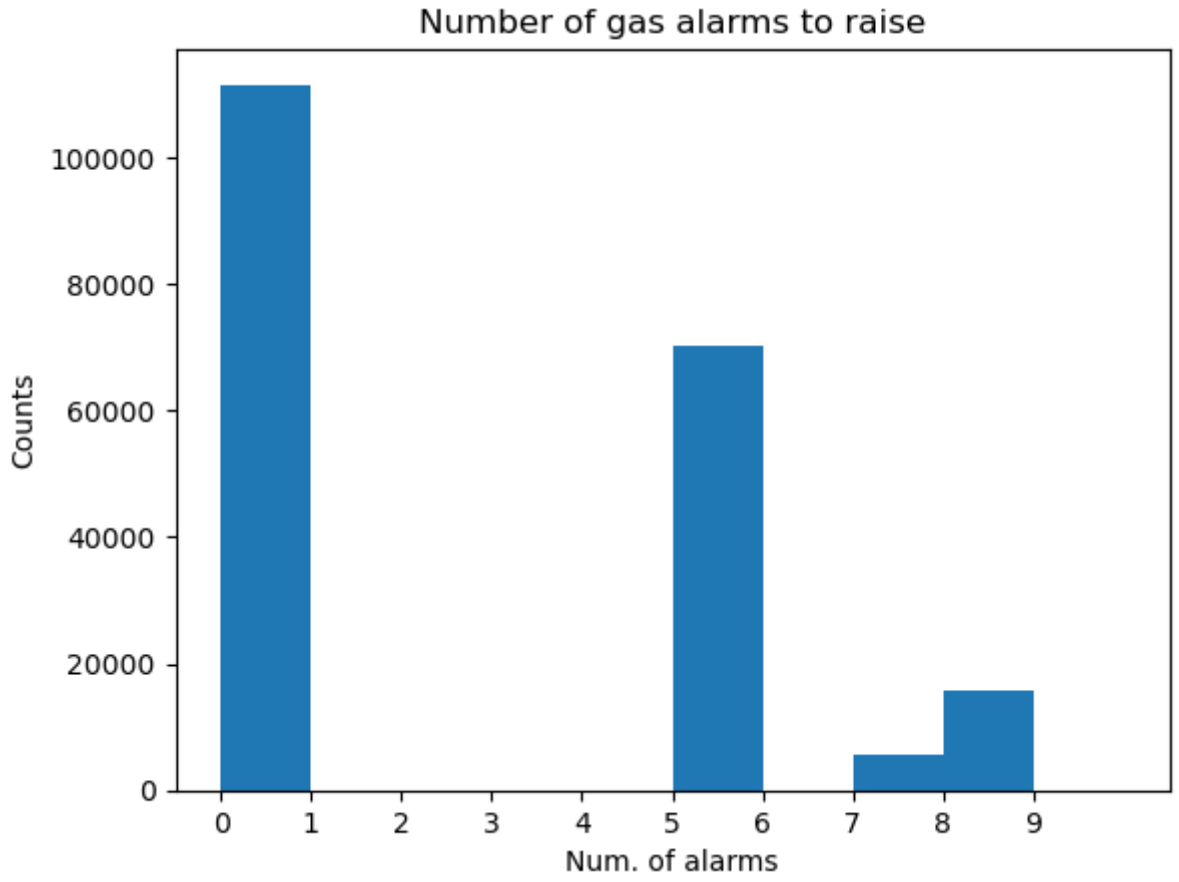
No missing values. Nothing to do in this regard.

On the characteristics of the labels.

The regression target is a vector of 24 components. Each component corresponds to a dangerous gas, and has a value between 0 and 1. Each number corresponds to an alarm level. In general, we can say that if the alarm level is above 0.5, alarm should be raised (this can be interpreted as a sort of probability of presence). The cost will punish more failures to predict activations larger than 0.5 in order to protect against false negatives.

How many observations include an alarming level?

We analyze the dataset by counting how many cells on each row of the target have alert values greater than 0.5:



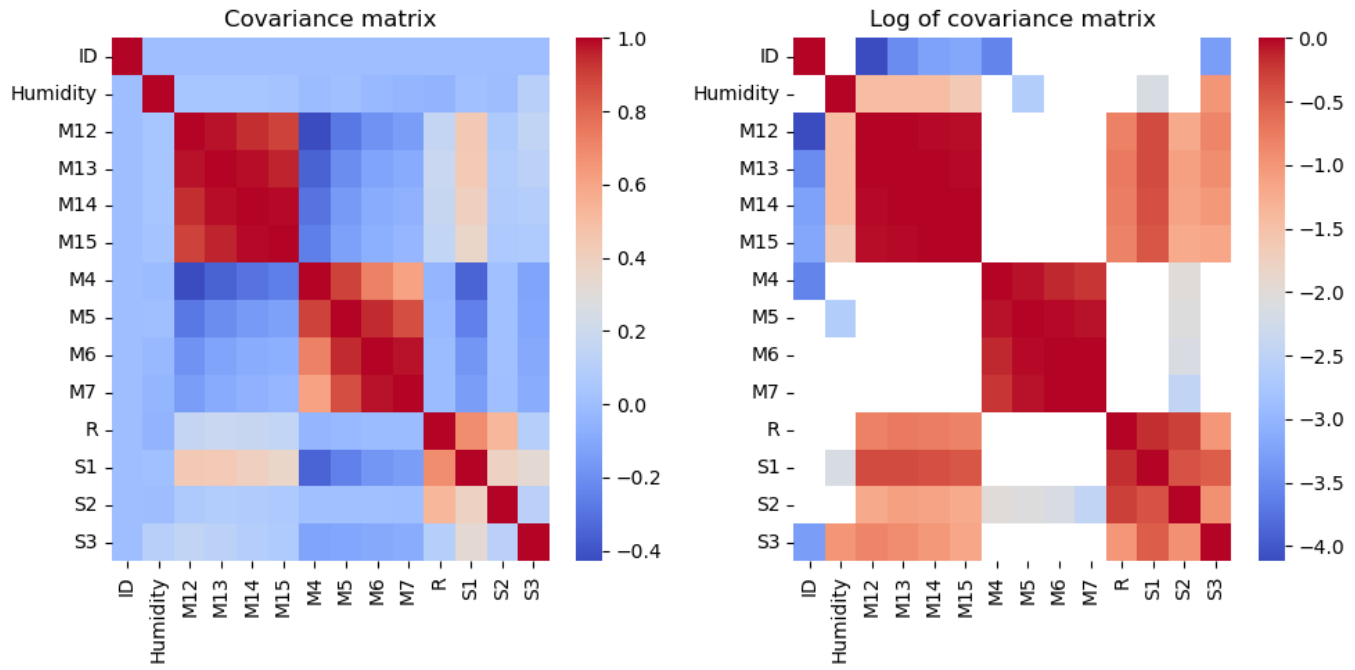
We see that most cases do not involve the raising of an alarm. But for the rest, multiple alarming levels of alert are found simultaneously. This is not the target of this challenge, but it seems unlikely that a cocktail of 5 toxic gasses could be found at once. Probably, to make a decision, the highest alert level should be selected. This is out of scope

Correlation between alarms

On the characteristics of the features

Machine learning is, in a way, a compression problem. Structure in the features is useful since we can implement feature transformations that reduce the dimensionality of the input, reducing model parameters and increasing information density.

Our features are a collection of 15 scalar numbers for each observation of the data. Each scalar is a numerical variable that encode the "activation level" of a given "element" of the detector.The correlation matrix of the data is:



There are a few observations to do.

- The ID column should be dropped, it is not related to anything, just a numeral.
- The humidity doesn't seem correlated to other features. As we are detecting gasses, more specifically chemical features of gasses, it suggests that all gasses are comparably soluble.
- The rest of the features are correlated in three blocks. This means that those 3 blocks probably represent some hidden variable binding together different categories.

The last fact is not surprising at all. Not all molecules are chemically possible, meaning that there is only a subset of combinations of detector readings that are physically realizable.