Anomaly Detection

In this assignment we are going to explore different models for anomaly detection.

The chosen data is **realKnownCause/ambient\_temperature\_system\_failure.csv**: the ambient temperature in an office setting.

We opt to try three different models:

* A simple K-Means algorithm
* DBSCAN
* Isolation Forest

Each of them presents its advantages and disadvantages. While K-Means is straightforward to use, it required the definition of the number of clusters. The Elbow Method (h[ttps://en.wikipedia.org/wiki/Elbow\_method\_(clustering)](https://en.wikipedia.org/wiki/Elbow_method_(clustering))) can be used for that purpose. After that, to detect outliers, is necessary to define a distance function to compare the distance of each point to its centroid. Finally, it is necessary to define the percentage of expect outliers in the data set. These are going to be the points farthest to their centroids.

DBSCAN, on the other hand, doesn’t require one to chose the number of clusters. However, it is necessary to set a parameter representing the maximum distance between two samples for one to be considered as in the neighbourhood of the other. This will translate into the amount of points being considered outliers.

Finally, the Isolation Forest algorithm ‘isolates’ observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature. The number of times you had to go through this step is the isolation number. The lower the number, the more anomalous the observation is. Here, it is necessary to specify a parameter called contamination, which defines the expect percentage of anomalies in the data.

From that, we can note that it is necessary to have a good knowledge about the data set to properly define the parameter that will be used to detect outlier. A bad decision might lead to the majority of the data being detect as outlier or none of them being detected as such.