Report of Predictive Maintenance

Dataset :

S. Matzka, "Explainable Artificial Intelligence for Predictive Maintenance Applications," 2020 Third International Conference on Artificial Intelligence for Industries (AI4I), 2020, pp. 69-74, doi: 10.1109/AI4I49448.2020.00023.

Meaning of variables :

1. UID: unique identifier ranging from 1 to 10000
2. product ID: consisting of a letter L, M, or H for low (50% of all products), medium (30%) and high (20%) as product quality variants and a variant-specific serial number
3. type: just the product type L, M or H from column 2
4. air temperature [K]: generated using a random walk process later normalized to a standard deviation of 2 K around 300 K
5. process temperature [K]: generated using a random walk process normalized to a standard deviation of 1 K, added to the air temperature plus 10 K.
6. rotational speed [rpm]: calculated from a power of 2860 W, overlaid with a normally distributed noise
7. torque [Nm]: torque values are normally distributed around 40 Nm with a SD = 10 Nm and no negative values.
8. tool wear [min]: The quality variants H/M/L add 5/3/2 minutes of tool wear to the used tool in the process.
9. a 'machine failure' label that indicates, whether the machine has failed in this particular datapoint for any of the following failure modes are true.

The machine failure consists of five independent failure modes

1. tool wear failure (TWF): the tool will be replaced of fail at a randomly selected tool wear time between 200 - 240 mins (120 times in our dataset). At this point in time, the tool is replaced 69 times, and fails 51 times (randomly assigned).
2. heat dissipation failure (HDF): heat dissipation causes a process failure, if the difference between air- and process temperature is below 8.6 K and the tools rotational speed is below 1380 rpm. This is the case for 115 data points.
3. power failure (PWF): the product of torque and rotational speed (in rad/s) equals the power required for the process. If this power is below 3500 W or above 9000 W, the process fails, which is the case 95 times in our dataset.
4. overstrain failure (OSF): if the product of tool wear and torque exceeds 11,000 minNm for the L product variant (12,000 M, 13,000 H), the process fails due to overstrain. This is true for 98 datapoints.
5. random failures (RNF): each process has a chance of 0,1 % to fail regardless of its process parameters. This is the case for only 5 datapoints, less than could be expected for 10,000 datapoints in our dataset. If at least one of the above failure modes is true, the process fails and the 'machine failure' label is set to 1. It is therefore not transparent to the machine learning method, which of the failure modes has caused the process to fail.

1. Data Overview

1.1 Shape and structure

- **Identification of the target :** Machine failure
- **Numbers of rows and columns :** 10000, 14
- **Variables types :**
    - Discret :
        - Not usefull (Id variables): “UID”, “Product ID”
        - Type,
        - Machine failure,
        - **TWF,*
        - **HDF,*
        - **PWF,*
        - **OSF,*
        - **RNF*

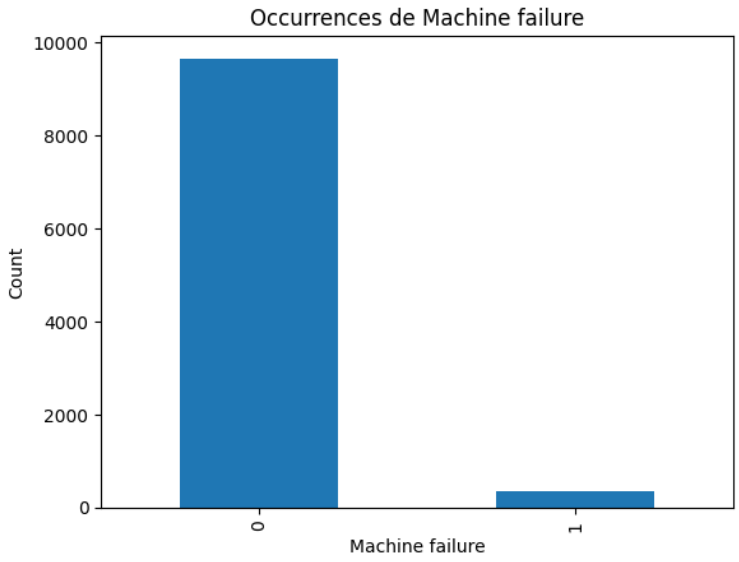
    **These variables are logically correlate with Machine failure, so because I need only one target and to don’t have data leak I will drop all these columns.*

    - Continuous :
        - Air temperature [k],
        - Process temperature [K],
        - Rotational speed [rpm],
        - Torque [Nm],
        - Tool wear [min],

1.2 Quality of Data

- **Identifications of the missing values :**
    - No missing values detected in this dataset
- **Identification of duplicate values**
    - No duplicate values

1.3 Target visualization :

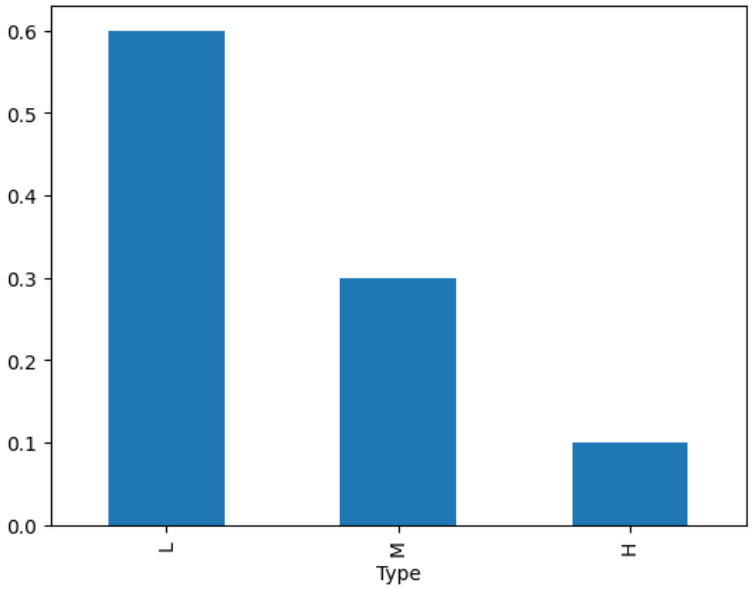


Machine failure

0 - 96.61%
1 - 3.39%

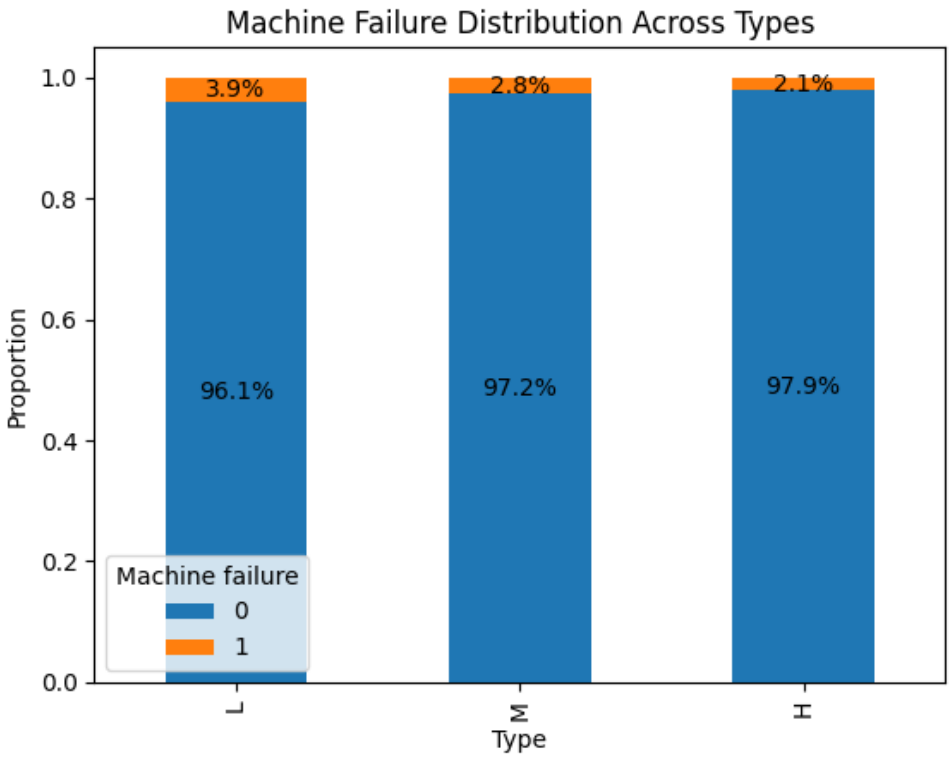
The dataset is unbalance, this is an important information to choose the criteria of evaluation of the Machine Learning model.

1.4 Distribution of the machines by type :



6000 (~60%) machines of TYPE L,
2997 (~30%) machines of type M,
1003 (~10%) machines of Type H.
The machine of type L are predominantly in this dataset.

1.5 Relation : Type / Machine Failure :

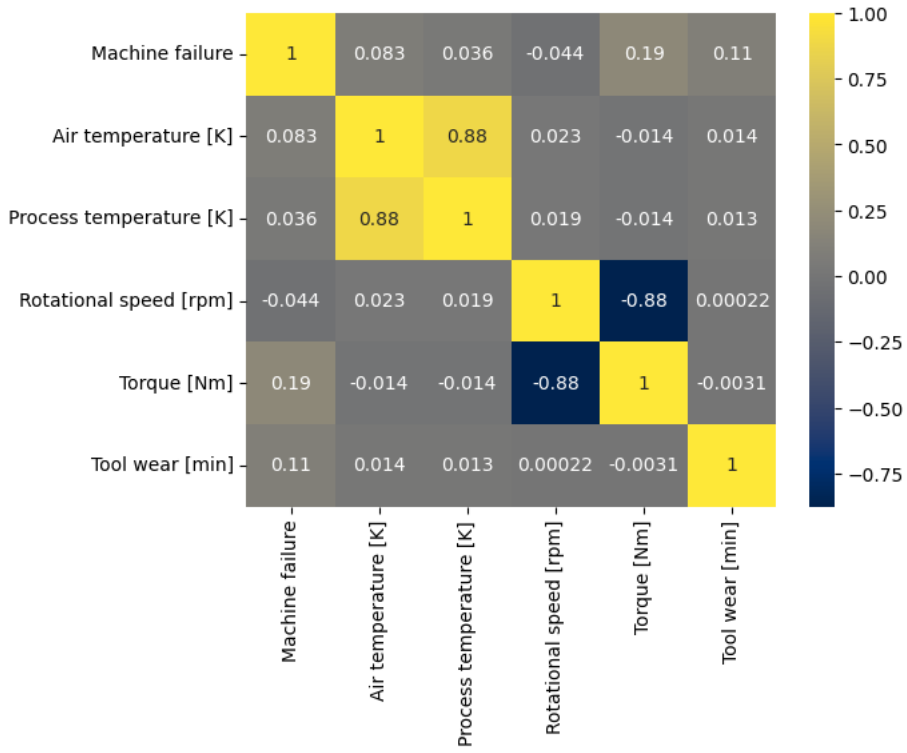


The failure rates are low for all machine types (below 4%). Type L shows a slightly higher failure rate (~3.9%) compared to types H (~2.1%) and M (~2.8%).

This difference remains moderate and would require a statistical to assess whether it is statistically significant.

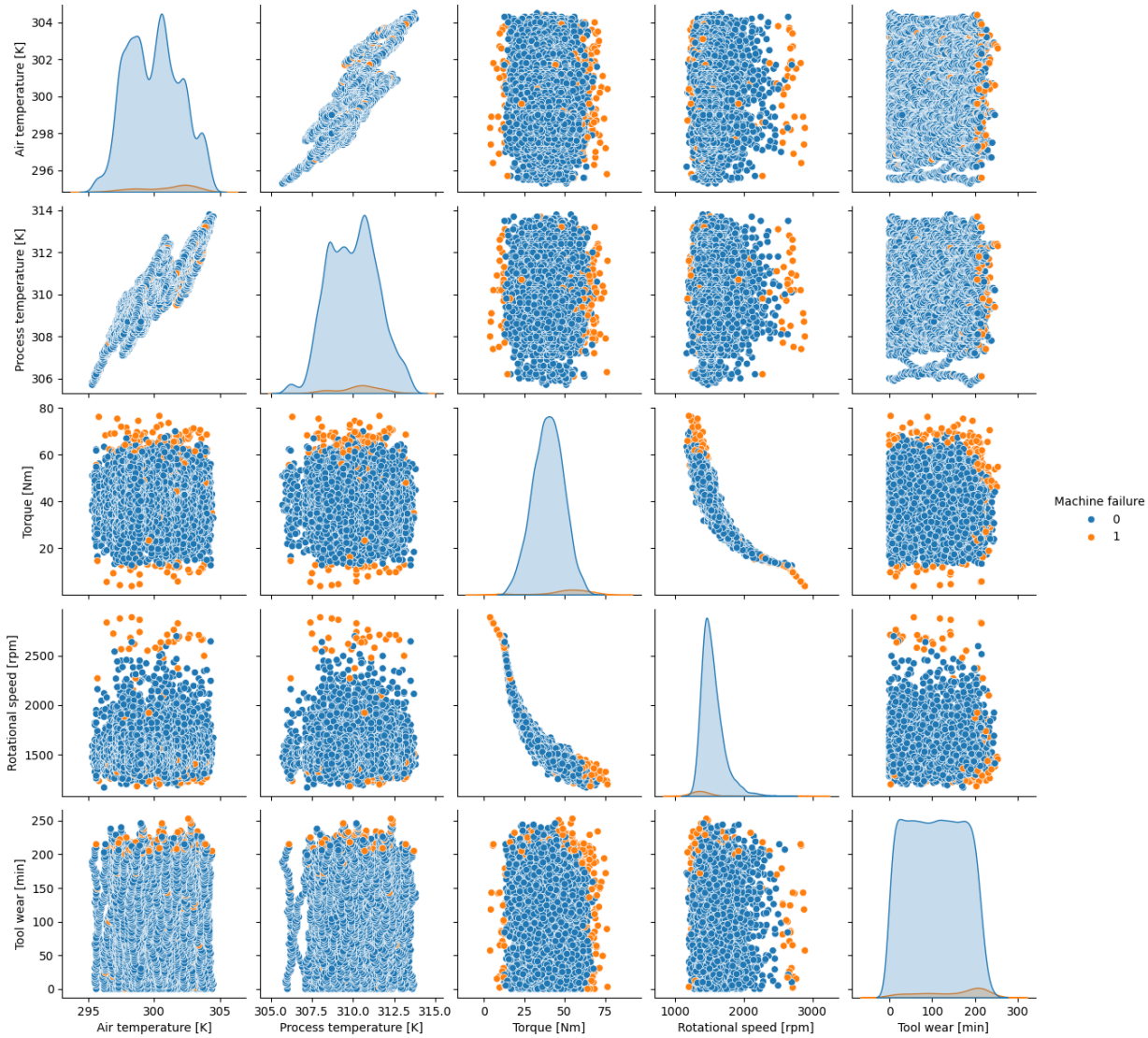
2. Global Exploration of interactions

2.1 Correlation matrix (heatmap)



- The Air temperature and Process temperature have a strong correlation (0.88)
- Torque and rotational speed are strongly correlate (-0.88)

2.2 Pairplot for Feature Relationships and correlation



This graph confirm a linear correlation between :

- Air temperature [k] and Process temperature [K]
- Torque [Nm] and Rotational speed [rpm]

In the following analysis, I will combine these variables in feature engineering.

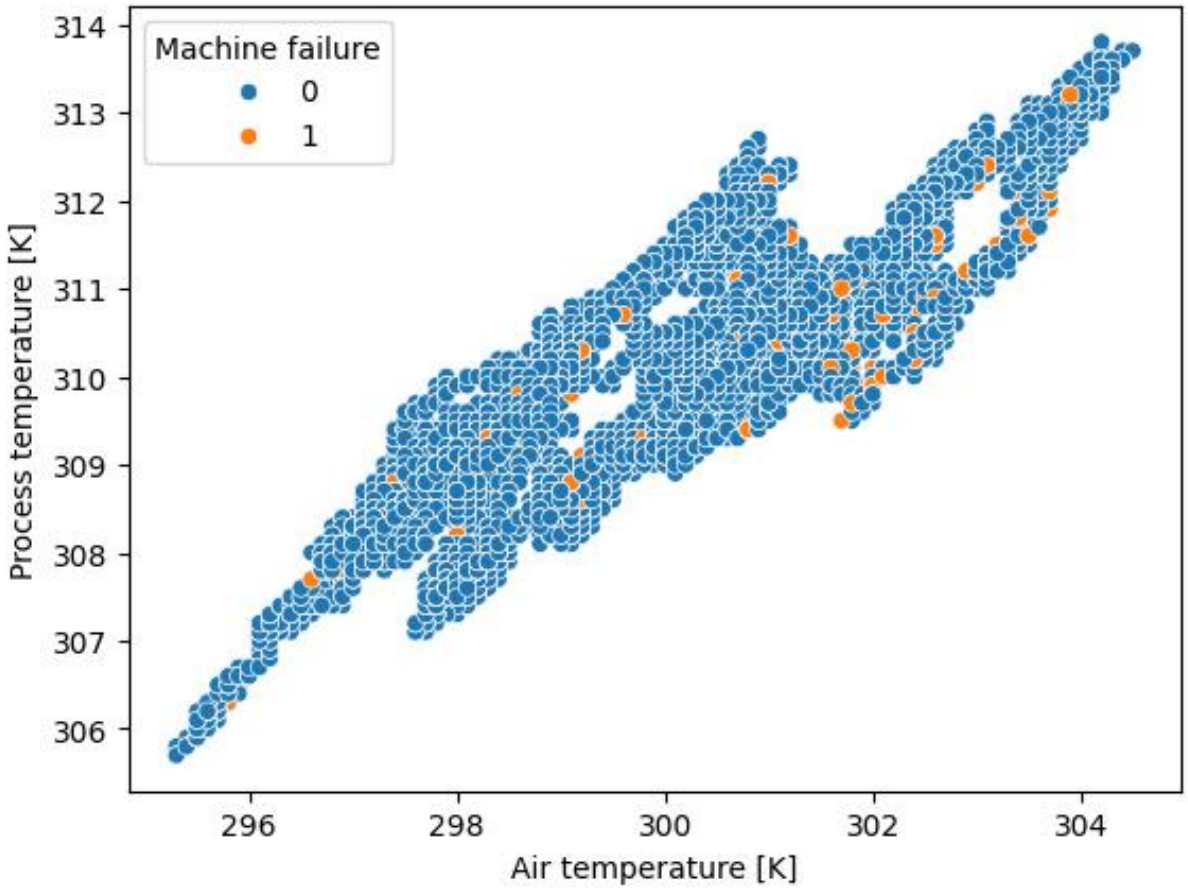
Temp_diff = Process temperature [K] - Air temperature [k]

Machine_power = Torque [Nm] * [pi() * Rotational speed [rpm] / 30]

3. In-depth Thematic Analysis

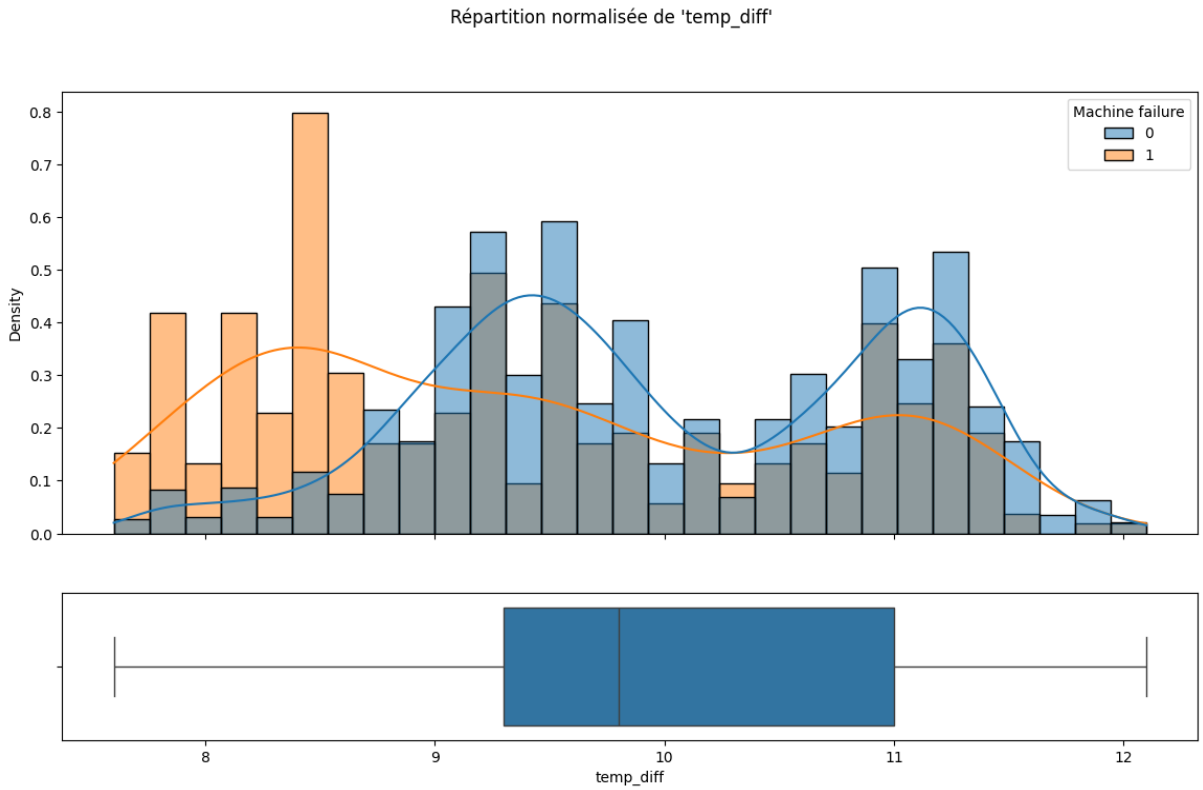
3.1 Thermal Analysis

A preliminary analysis of the raw variables Air temperature and Process temperature was conducted (details in **Appendix A**). While these variables follow stable distributions, they exhibit a significant overlap between normal operations and failures, limiting their individual predictive power.



As suggested by the correlation matrix, the scatter plot above confirms a strong linear relationship (0.88) between the two temperatures. This redundancy implies that monitoring both variables independently adds little value

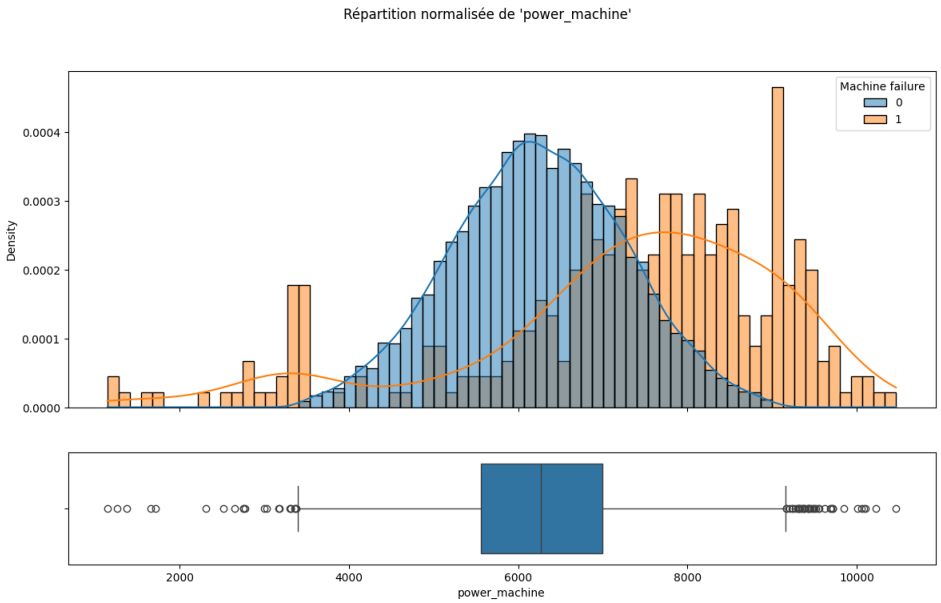
Based on the physical properties of the system (Heat Dissipation Failure - HDF), the critical risk factor is not the absolute temperature, but the machine's **inability to dissipate heat**. Therefore, we engineered a new feature: temp_diff (Process - Air).



The 8.6 K threshold is critical. As the graph demonstrates, a temperature difference below this value drastically increases the risk of failure.

The distribution above visually confirms this theory. To rigorously validate it, we will now verify if the failures strictly fall below the 8.6 K threshold mentioned in the domain documentation.

3.2 Mecanique and power analysis



This graphic is normalize to make easier the analysis. Clearly it appear that the outlier are outside of the normal range of the machines and correspond to failure points

3.3 Maintenance and usure analysis

4. Synthesis and features selection

4.1 Deleted variables

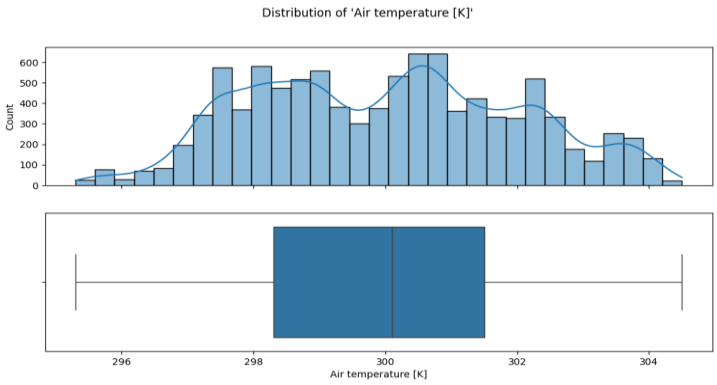
4.2 Useful variables

4.3 Conclusion

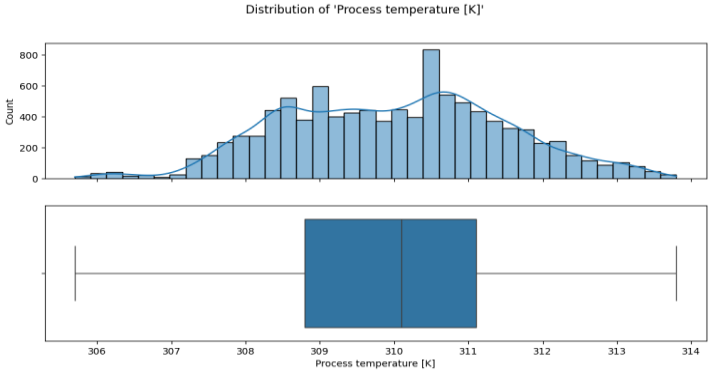
Annexe A :

Detailed Analysis of Raw Temperature Variables

1. Global Distribution & Statistics (Univariate)

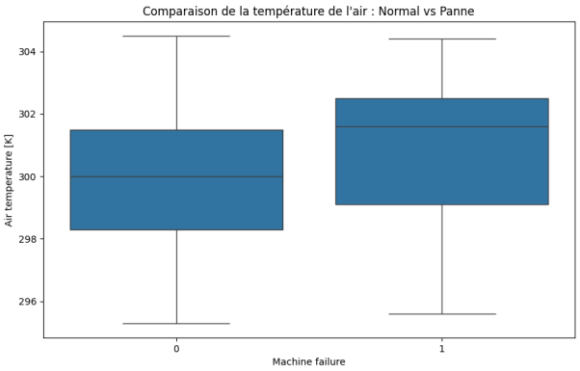


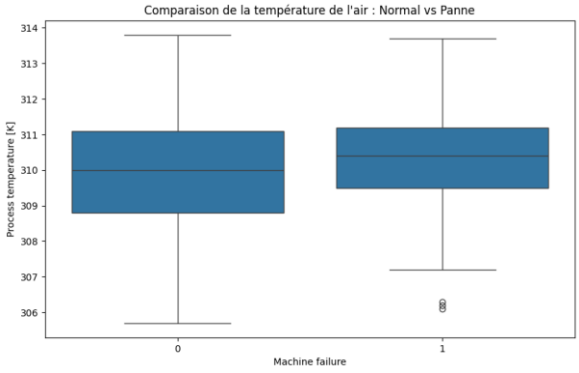
count	10000.000000
mean	300.004930
std	2.000259
min	295.300000
25%	298.300000
50%	300.100000
75%	301.500000
max	304.500000



count	10000.000000
mean	310.005560
std	1.483734
min	305.700000
25%	308.800000
50%	310.100000
75%	311.100000
max	313.800000

2. Discrimination Capability vs. Machine Failure (Feature-Target)

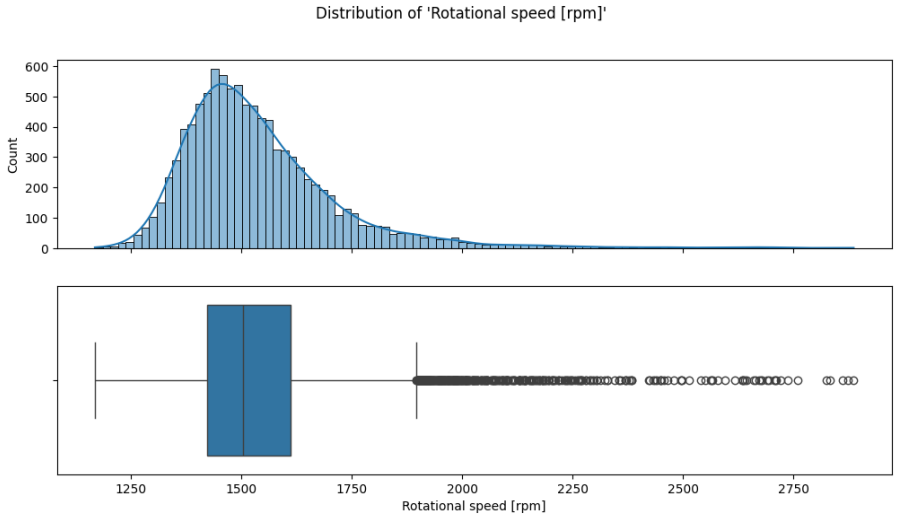




Observation: The univariate analysis (top) confirms that both Air and Process temperatures are stable and follow a multimodal distribution. However, the Feature-Target analysis (bottom) reveals a **significant overlap** between the interquartile ranges of normal operations and machine failures.

Conclusion: This confirms that raw temperature values lack sufficient discriminative power to predict failures on their own. This observation **justifies the Feature Engineering strategy** (creation of the temp_diff variable) presented in the main body of the report.

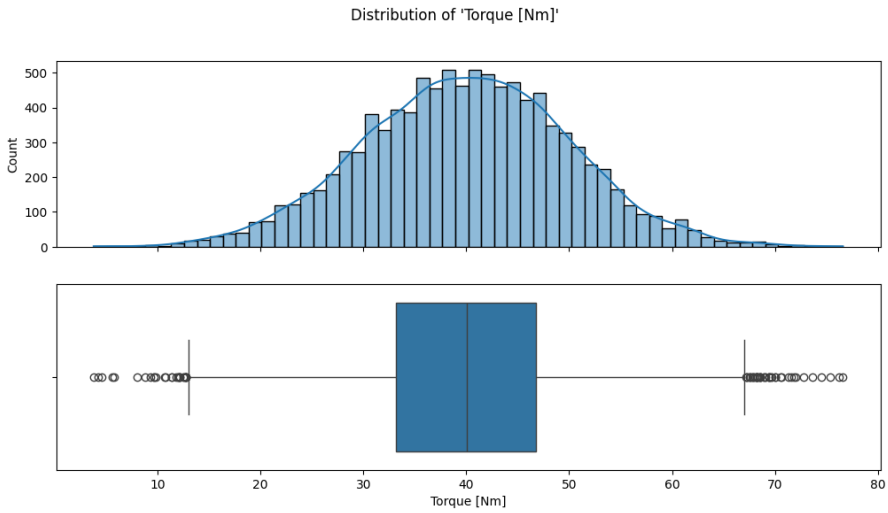
‘Rotational speed [rpm]’ :



count	10000.000000
mean	1538.776100
std	179.284096
min	1168.000000
25%	1423.000000
50%	1503.000000
75%	1612.000000
max	2886.000000

- **Distribution:** The distribution of Rotational speed is strongly **right-skewed (positive skew)**. A large concentration of the data is clustered at lower speeds, with a peak around 1500-1600 rpm. This is followed by a long tail of progressively less frequent, higher-speed values.
- **Outliers and Hypothesis:** The boxplot highlights a significant number of data points as outliers, all located at high rotational speeds. While a sensor malfunction could be an initial hypothesis, the outliers form a continuous and structured tail rather than random noise. This suggests they represent a **valid but less common mode of operation**. This leads to the hypothesis that these high-speed events correspond to specific physical conditions, such as periods of low machine load. This will be investigated further in the bivariate analysis. For this reason, these outliers are considered meaningful and will be retained.

4.3.1 Distribution of the variable ‘Torque [Nm]’ :

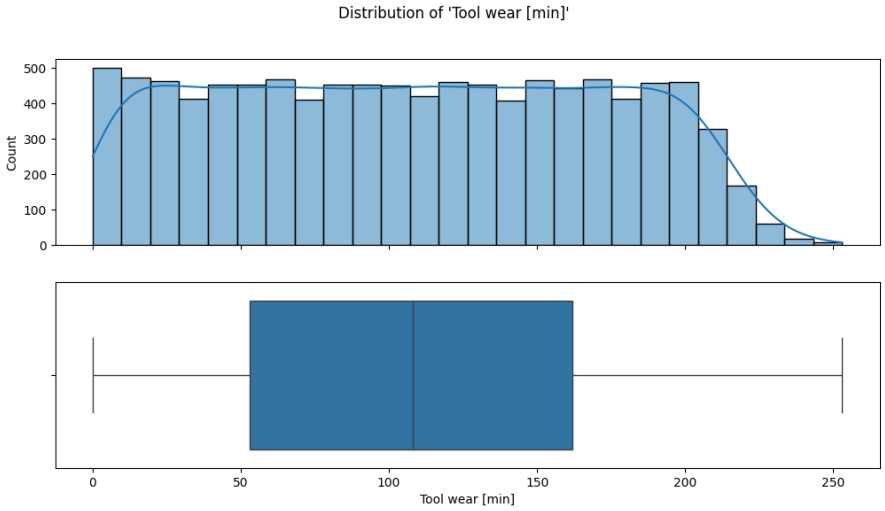


count	10000.000000
mean	39.986910
std	9.968934
min	3.800000
25%	33.200000
50%	40.100000
75%	46.800000
max	76.600000

- **Distribution:** The variable follows a **perfectly symmetrical Normal (Gaussian) distribution**, centered around a mean of **40 Nm** with a standard deviation of **10 Nm**. The mean and median are almost identical, indicating no skewness.

- **Outliers & Physical Interpretation:** The boxplot reveals outliers on both extremes of the distribution.
    - **Low Torque Outliers (< ~14 Nm):** These represent operating modes under very light loads, physically corresponding to the high-speed outliers observed in the Rotational speed analysis.
    - **High Torque Outliers (> ~66 Nm):** These represent operating modes under heavy strain. These points are critical as they likely correspond to high-stress situations where the risk of failure (specifically Overstrain or Power Failure) is significantly increased. They are valid and essential for the model.

4.3.2 Distribution of the variable 'Tool wear [min]' :



count	10000.000000
mean	107.951000
std	63.654147
min	0.000000
25%	53.000000
50%	108.000000
75%	162.000000
max	253.000000

- **Distribution:** Tool Wear follows a **Uniform Distribution**. The frequency is roughly constant from 0 to 200 minutes. The Mean (107.9 min) and Median (108.0 min) are identical, confirming the symmetry of the uniform distribution. The maximum value is of 253 miutes.

4.4 In-depth Analysis multivariate

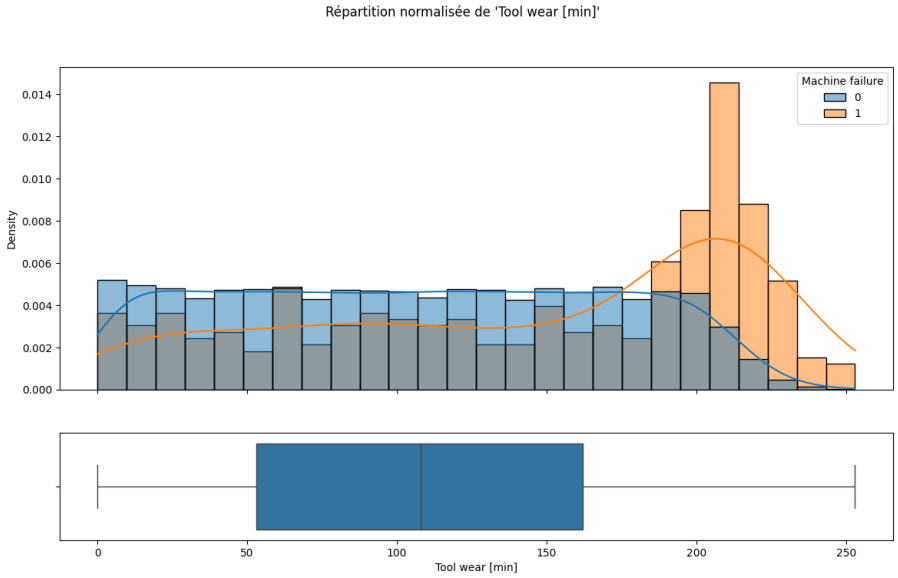
4.4.1.1 Pairplot for Feature Relationships and correlation

Identification of outliers :

The outliers are situate in low RPM and low/hight torque, however this is the area of failure so I will keep them for the training model

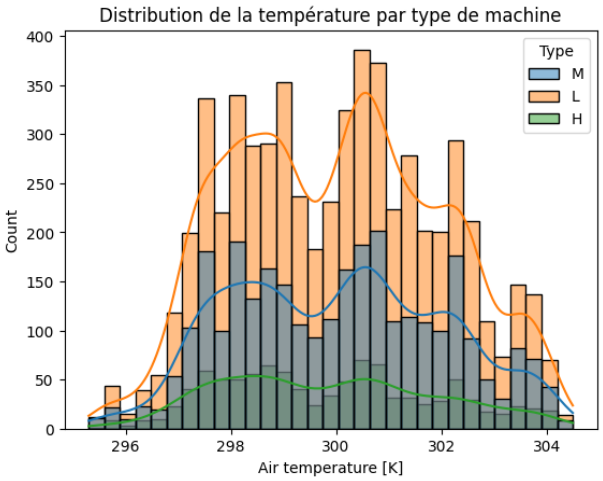
4.4.2 Visualization of the features – target relationship :

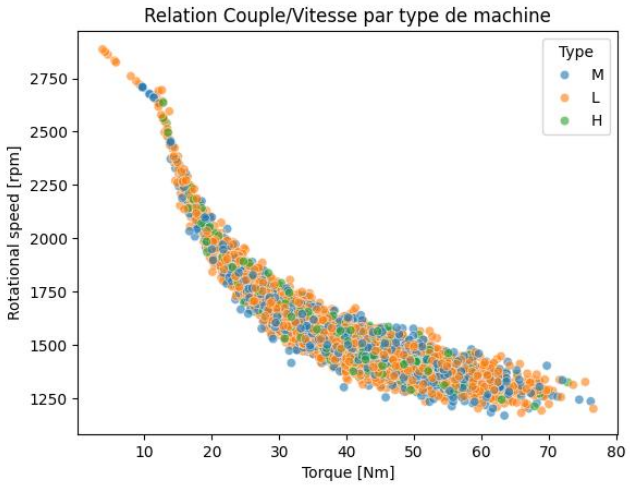
4.4.2.1 Tool wear [min] / Machine Failure



Visualization of the features – features:

4.4.2.2 Is the 3 types of machines in the same range of temp, torque and rpm ?





It appear that the 3 types of Machine run in the same range of temperature, torque and speed. We can see 2 types of working arround 298K and 301K. And we can see the 3 quality of machines, L have a lower thermique quality than M and H better than the 2 others.