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. Exploratory Data Analysis

Shape Analysis

Identification of the target : Machine failure

Numbers of rows and columns : 10000, 14

Variables types :

Not usefull (Id variables):

- UID
- Product ID

Discrets :

- 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.*

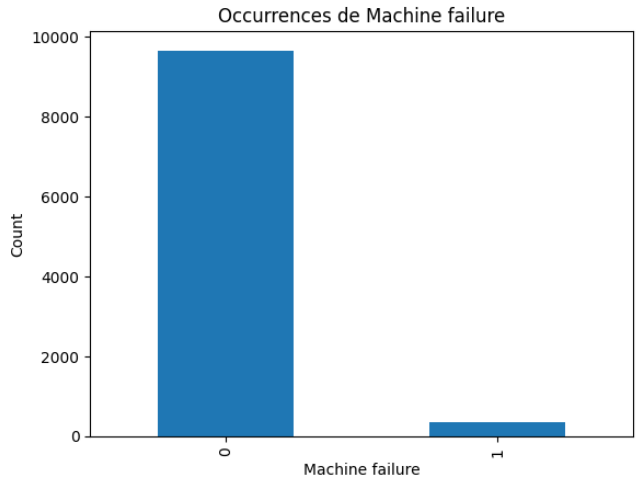
Continuous :

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

Identifications of the missing values : No missing values

In-depth Analysis

Target visualization :

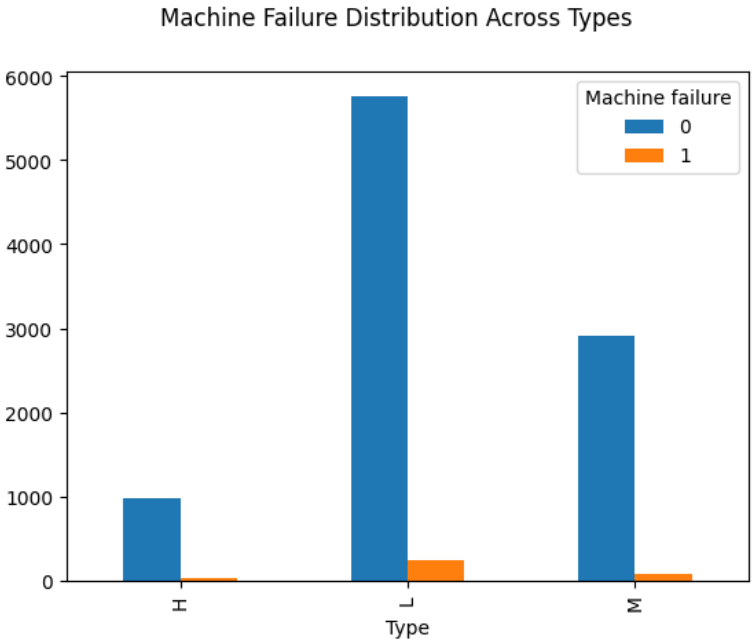


Machine failure

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.

Visualization of the features – target relationship :



In percentage

Machine failure	0	1
Type		
H	0.0982	0.0021
L	0.5765	0.0235
M	0.2914	0.0083

Counts

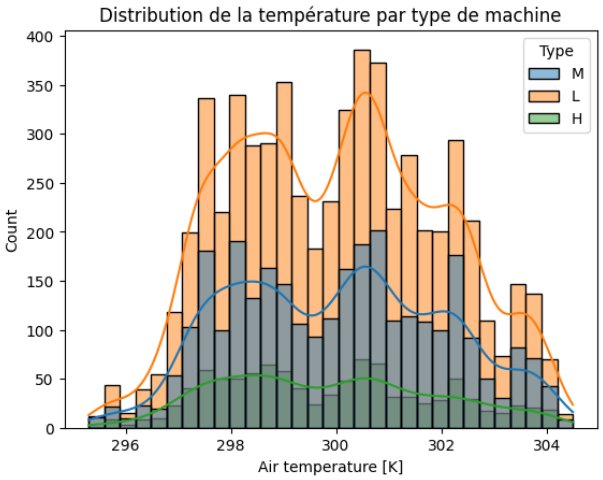
Machine failure	0	1
Type		
H	982	21
L	5765	235
M	2914	83

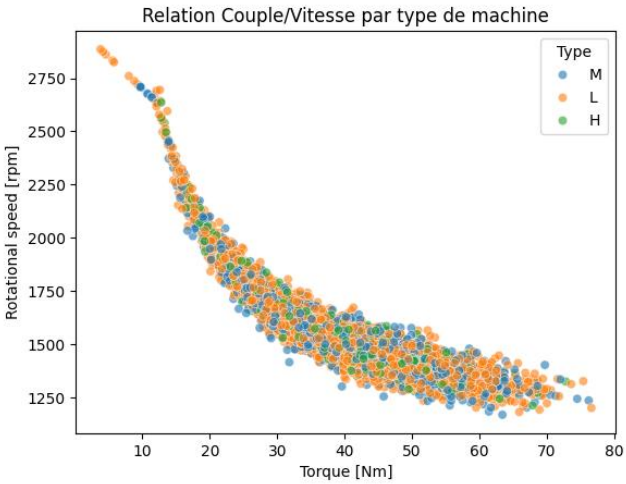
Type L : 2.35 % of failure

Type M: 0.83% of failure

Type H: 0.21% of failure

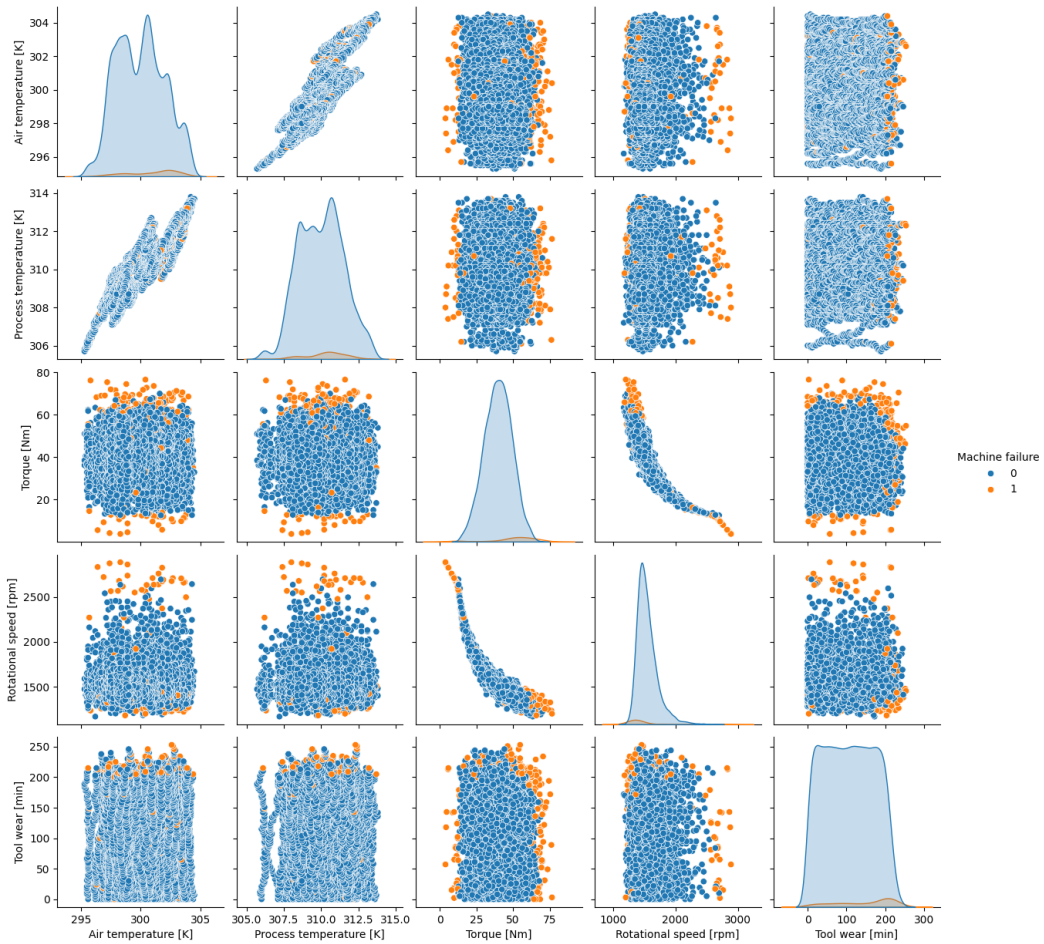
Is the 3 types of machines in the same range of temp, torque and rpm ?

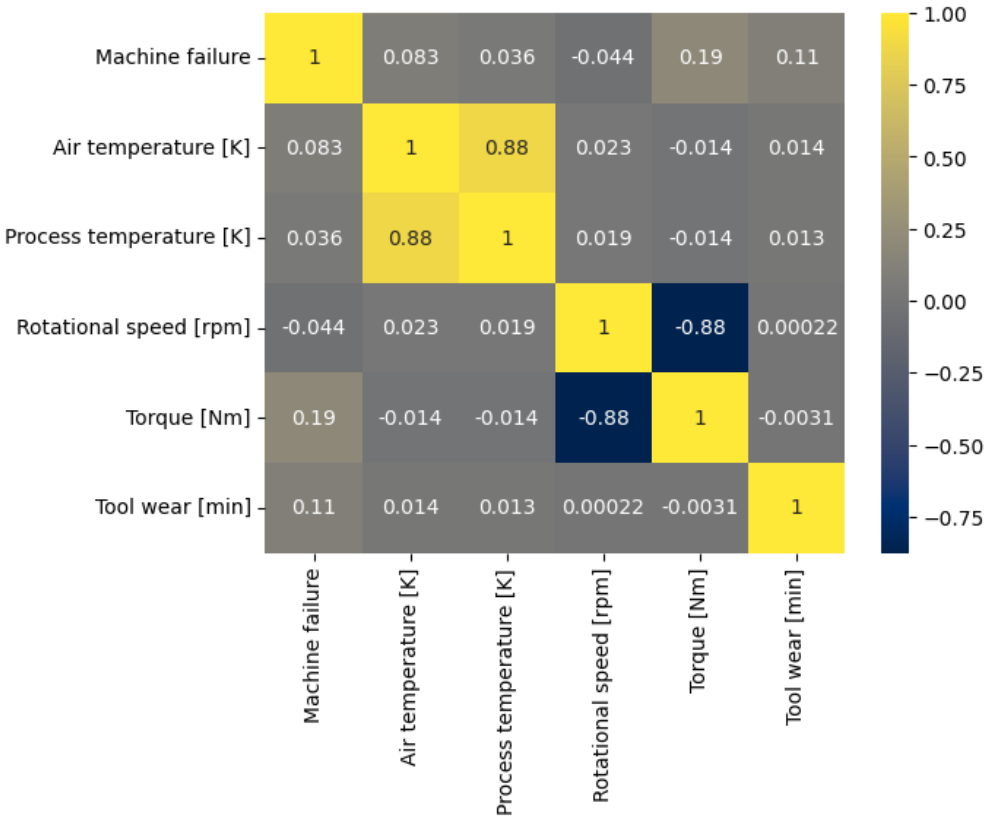




It appear that the 3 types of Machine run in the same range of temperature, torque and speed. We can see 2 types of working arround 298K and 301K. And we can see the 3 quality of machines, L have a lower thermique quality than M and H better than the 2 others.

Pairplot for Feature Relationships and correlation





The Air temperature and Process temperature have a linear correlation (0.88), so for later it can be interesting to combine this 2 variables.

Torque and rotational speed are negatively linear correlate (-0.88), combine this 2 variables can help the models with the problems of multilinearity.

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.