Predictive maintainance for Industrial Equipment

The dataset consists of 10 000 data points stored as rows with 14 features in columns

- UID: unique identifier ranging from 1 to 10000
- productID: containing unique numbers to identify product
- Type: 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
- air temperature [K]: generated using a random walk process later normalized to a standard deviation of 2 K around 300 K
- 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.
- rotational speed [rpm]: calculated from powepower of 2860 W, overlaid with a normally distributed noise
- torque [Nm]: torque values are normally distributed around 40 Nm with an Ïf = 10 Nm and no negative values.
- tool wear [min]: The quality variants H/M/L add 5/3/2 minutes of tool wear to the used tool in the process.

Target column contains values 0 for no failure and 1 failures containing different types of failure.Random failures are included in target 0 as they cannot be actually detected as they occur for some unknown reasons.

All necessary modules like numpy, pandas,seaborn,matplotlib are imported and the csv file is read using pandas.

The column UID and Product Id are dropped as they have random unique values which are unnecessary for machine predictive maintenance.

	Type	Air temperature [K]	Process temperature [K]	Rotational speed [rpm]	Torque [Nm]	Tool wear [min]	Target	Failure Type
0	M	298.1	308.6	1551	42.8	0	0	No Failure
1	L	298.2	308.7	1408	46.3	3	0	No Failure
2	L	298.1	308.5	1498	49.4	5	0	No Failure

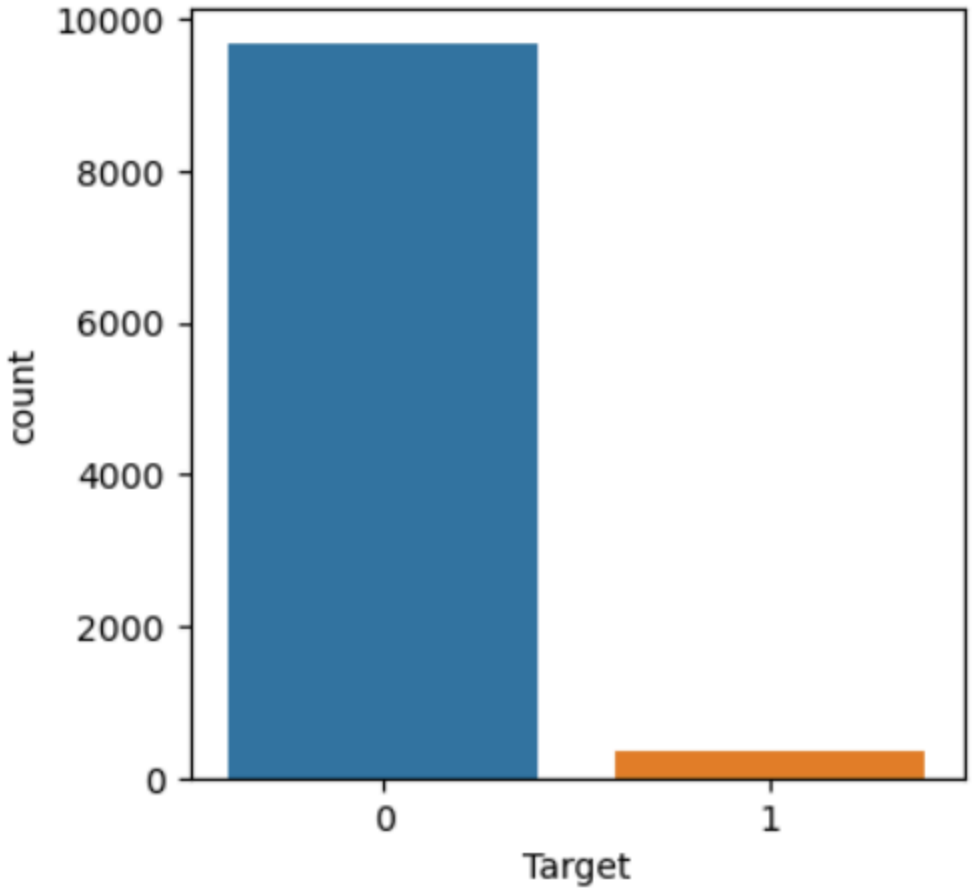
The dataset is checked for null and duplicate values ,it is found that there are no null and duplicate values.
The dataset is grouped in a way to find types of errors included in values of target column

		count
Target	**Failure Type**	
0	**No Failure**	9643
	Random Failures	18
	Heat Dissipation Failure	112
1	**No Failure**	9
	Overstrain Failure	78
	Power Failure	95
	Tool Wear Failure	45

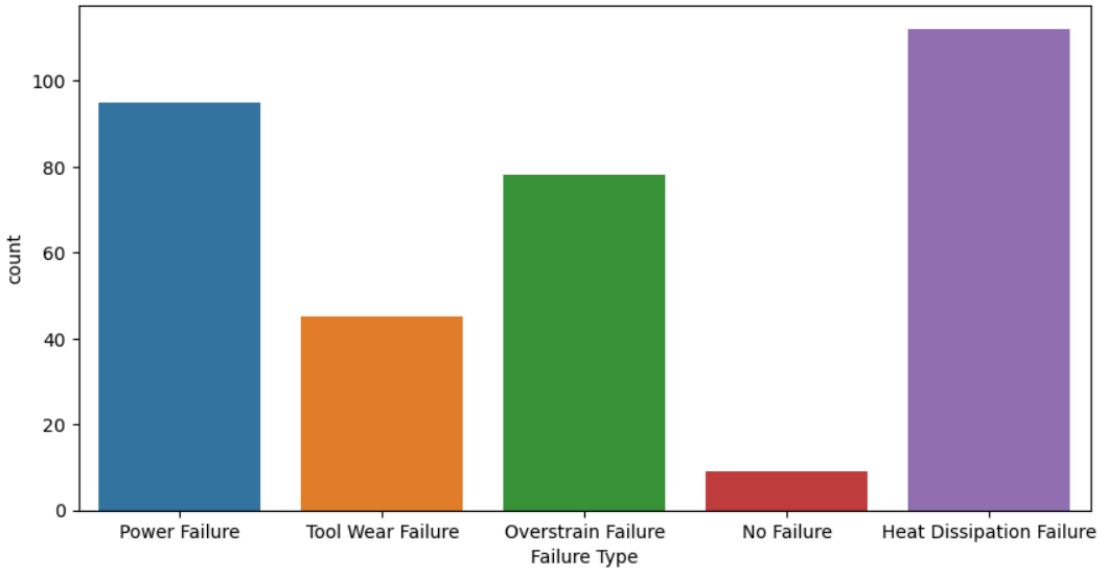
The data is converted into numbers for sake of processing using ordinal encoder from scikit learn module.
The data is grouped in order to visualize targets present in Type column containing labels H, Land M

		Air temperature [K]	Process temperature [K]	Rotational speed [rpm]	Torque [Nm]	Tool wear [min]	Failure Type
Type	**Target**						
0.0	**0**	299.7	309.9	1502.0	40.2	106.0	1.0
	1	302.0	310.2	1371.0	53.8	147.0	3.0
1.0	**0**	300.1	310.1	1508.0	39.7	107.0	1.0
	1	301.2	310.4	1362.0	53.9	182.0	2.0
2.0	**0**	300.1	310.0	1506.0	40.0	105.0	1.0
	1	302.0	310.6	1372.0	51.6	125.0	3.0

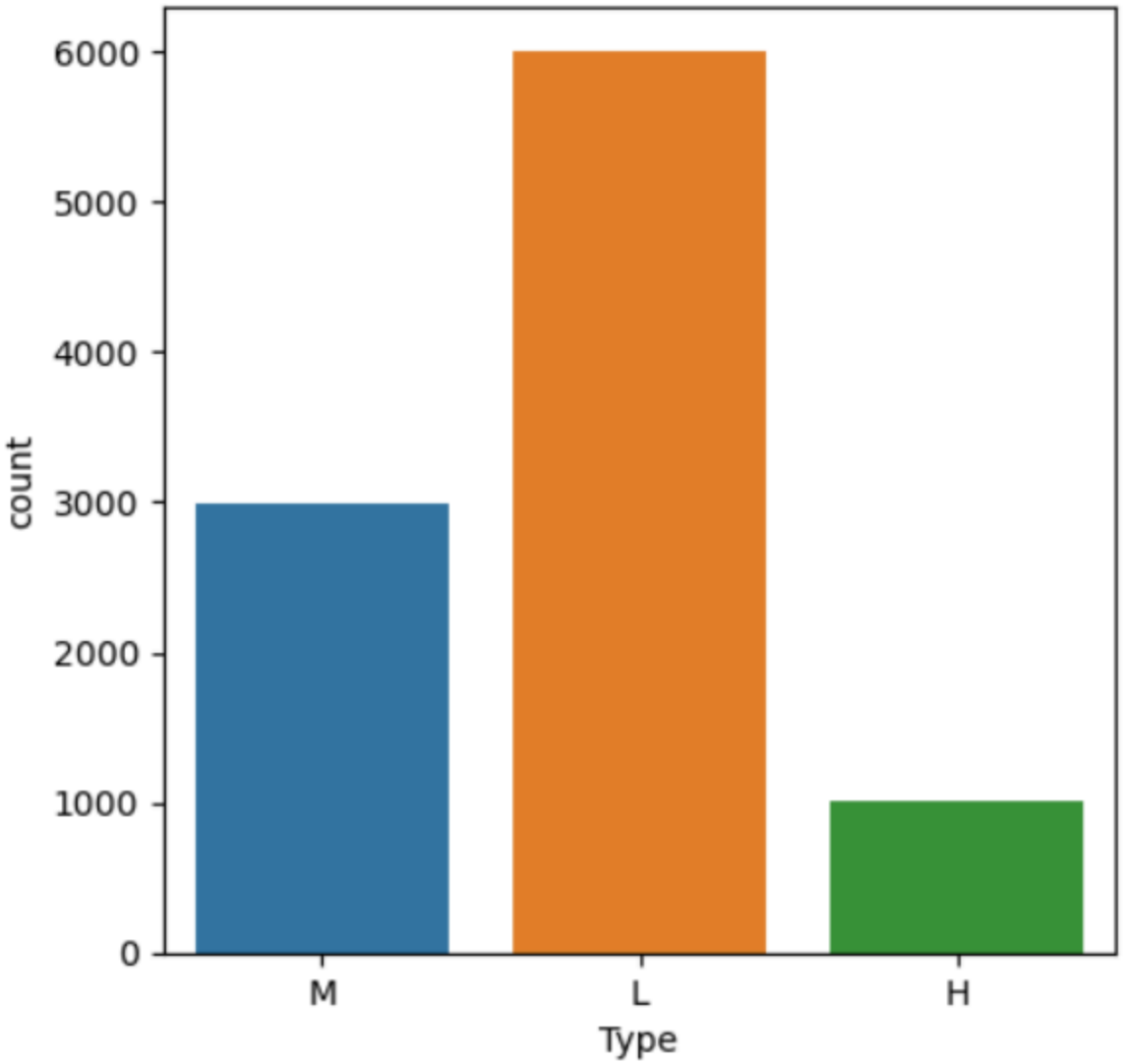
Here M is converted into numeric value 0.0 ,M into 0.1 and L into 0.2 as ordinal encoder module was used before.
Using matplotlib the counts for targets were plotted.It is found counts for target value 1 which represent failures is quite low.



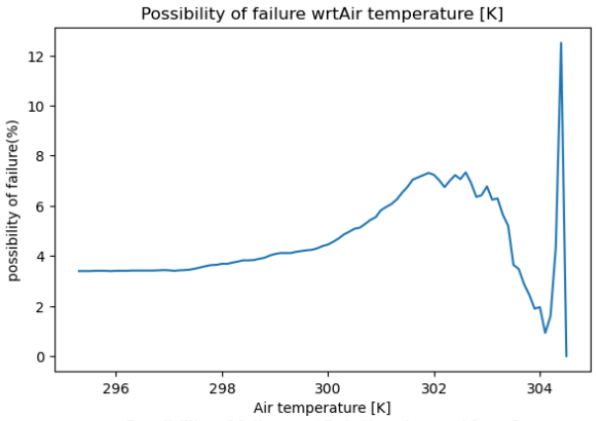
Counts for type of failures were also plotted.

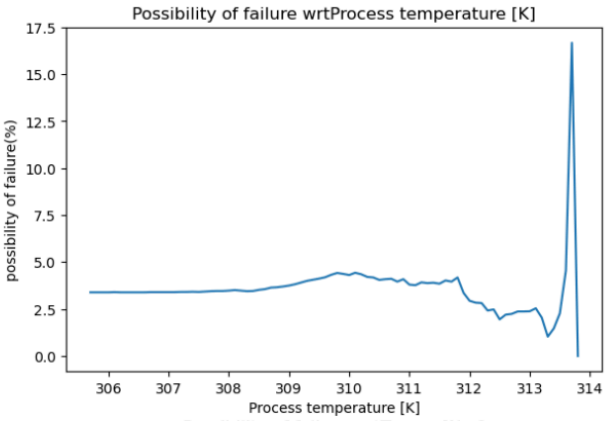


From this it can be concluded that number of heat dissipation failures and power failures and overstrain failures are high compared to others.
The count for L,M and H labels are also plotted. The count for label L is the highest.

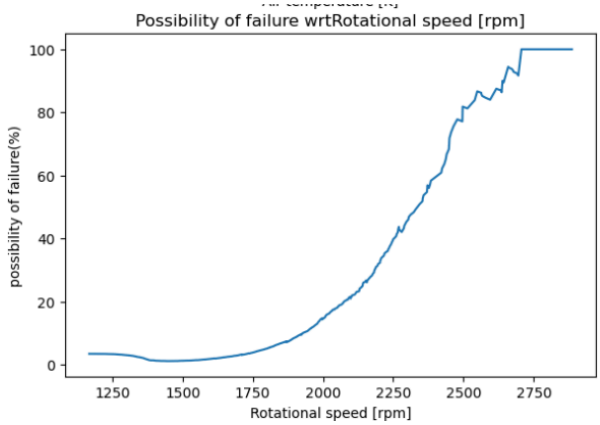


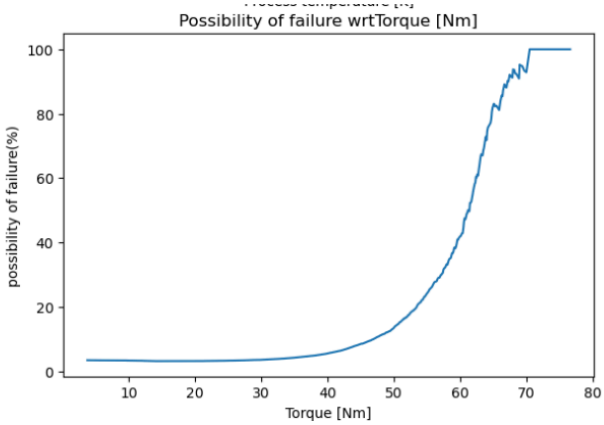
The graphs are plotted to find the effect of features like Air Temperature leading to machine failure.



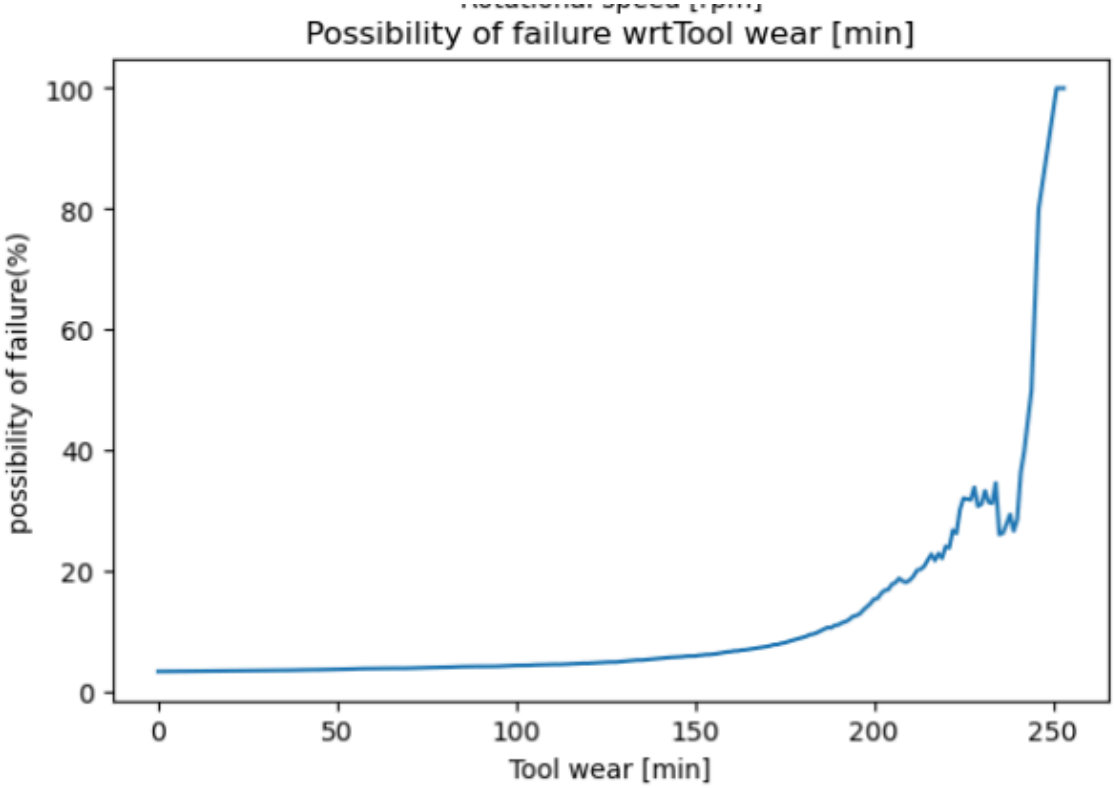


From this we can conclude that there is no increasing or decreasing pattern but probability of failure is quite high at certain temperature then it decreases suddenly for both Air and Process temperature.



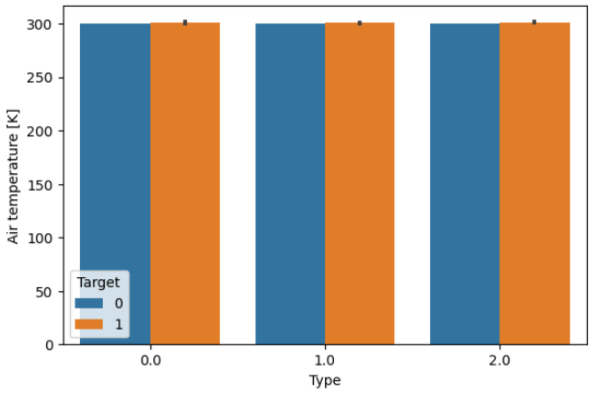


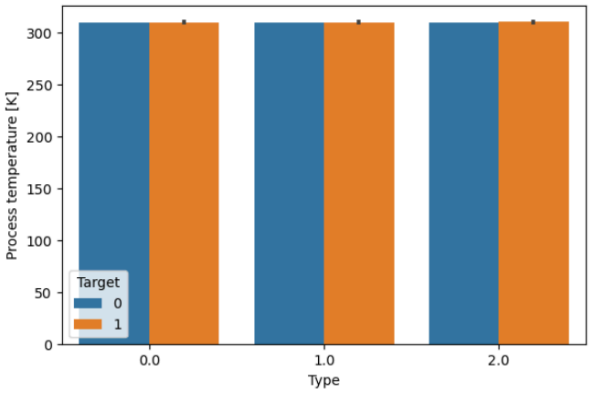
From this it can concluded that with increase in Rotional speed and torque probability of machine failure also increases and for rotational speed at about 2700 rpm and torque about 70 Nm and onwards probability of machine failure remains 100%.

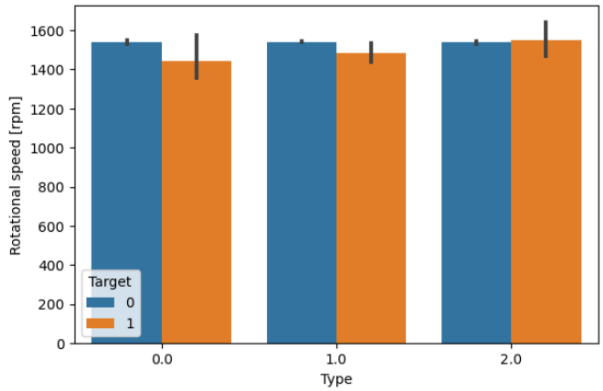


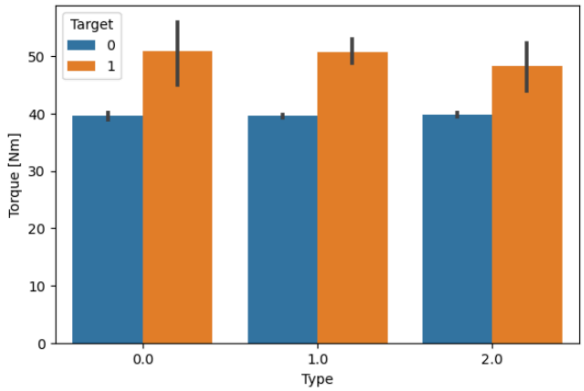
For tool wear also increase in minutes for tool waer also increases probability but the rate of increase in tool wear is quite slow and about 250 min for tool wear the probability of failure is 100%.
All this can conlcluded from graphs can be used for anomaly detection.

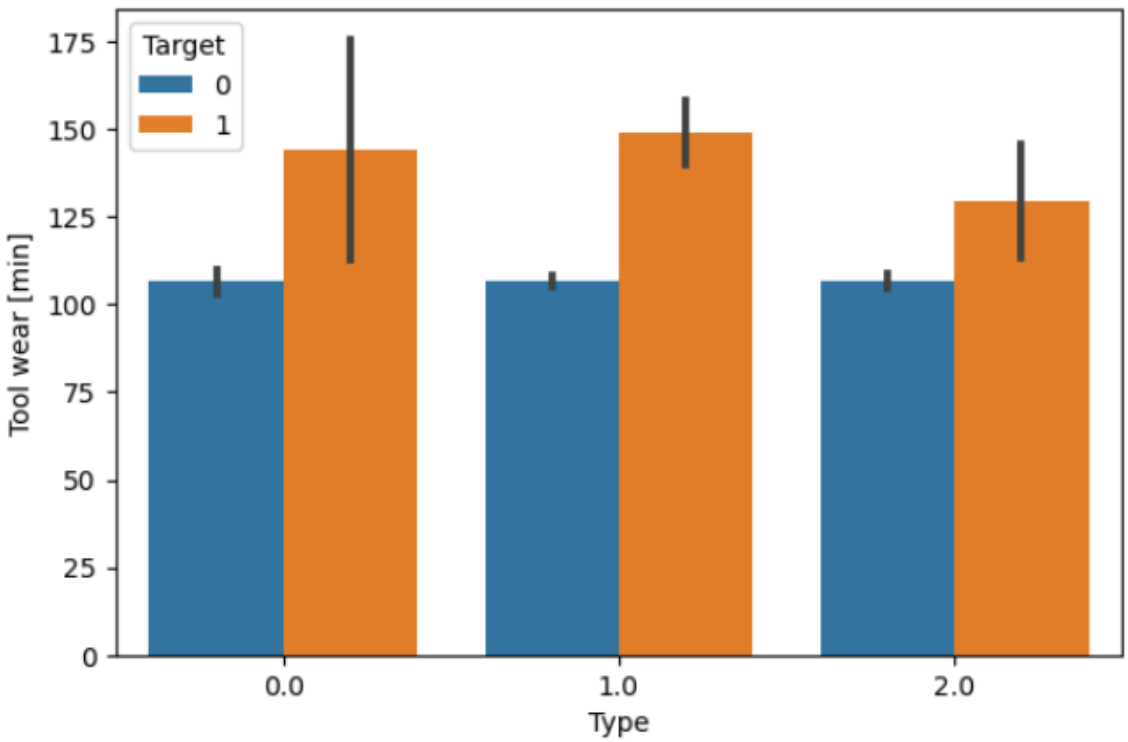
More 5 graphs are plotted to find the counts for features like air temperature having labels L,M and H having further counts of targets 0 and 1.



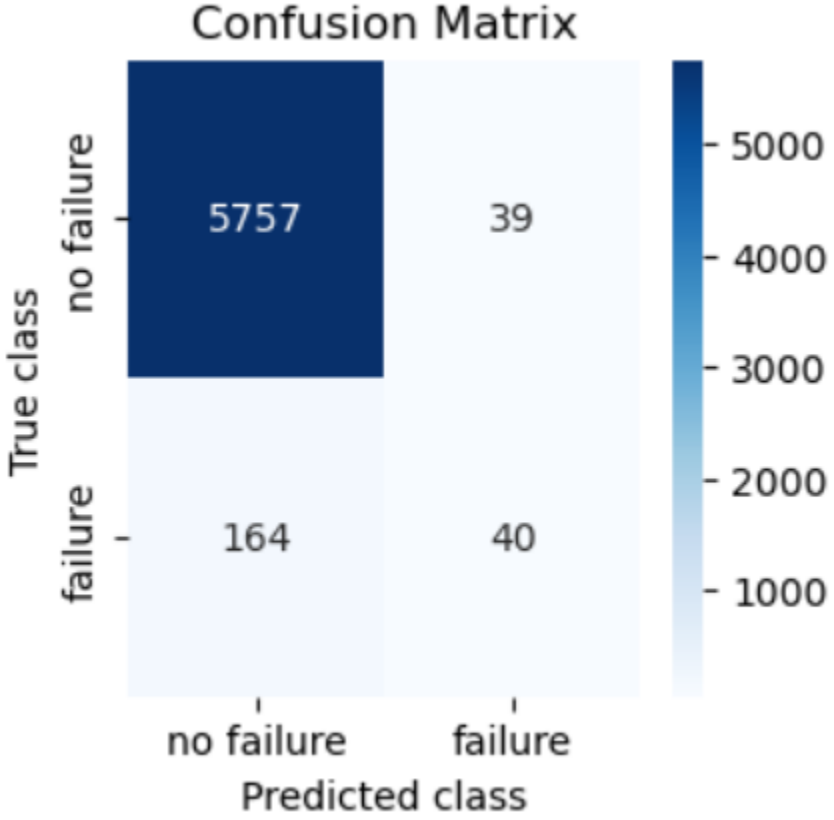




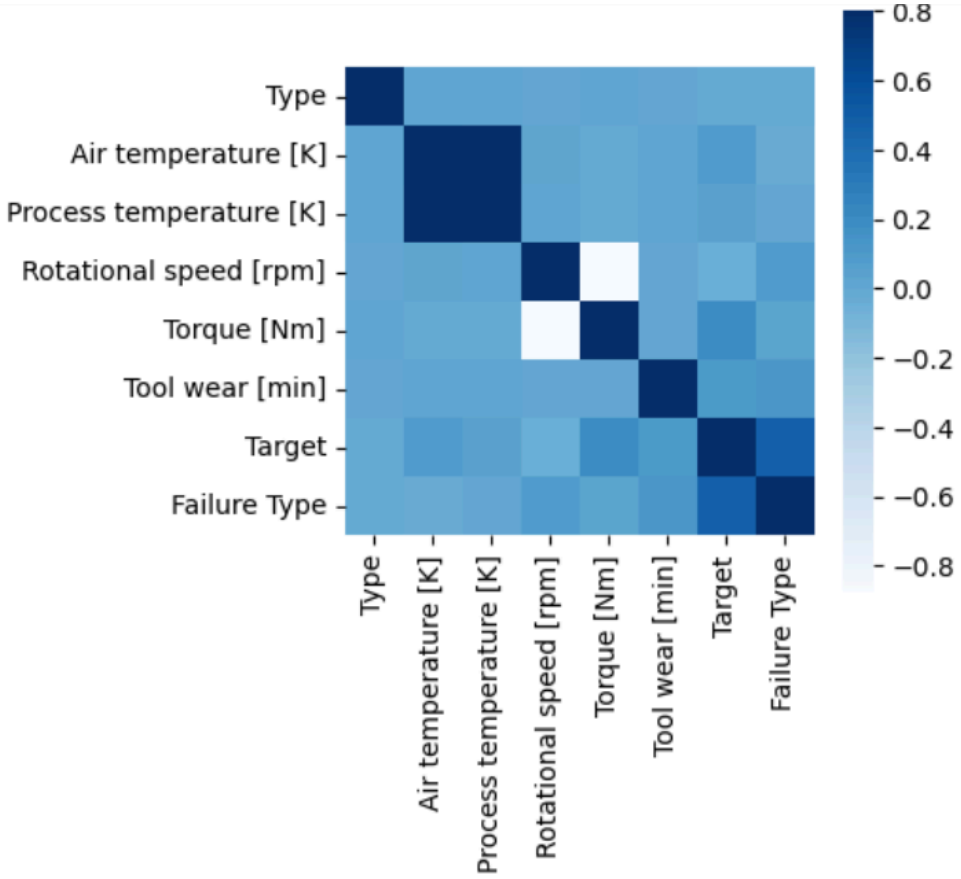




In this way the whole dataset is visualised.
The datset is is split into X an Y for having input and output datasets.
X contains all features except failure type and target and Y containing target column.
The X an Y dataset are further divided into training and testing datasets.
These training datasets are trained on models like kneighbors classifier, random forest classifier and logistic regression.and the model is used to predict results wrt testing dataset.
It is found that random forest classifier model gives the best results with accuracy about 98%.
The coefficient matrix is drawn for having better insight on the result.



From correlation matrix it can be concluded that process and air temperature are quite the same and either of colmns can dropped and types of failure is closely related to targets.


Type
Air temperature [K]
Process temperature [K]
Rotational speed [rpm]
Torque [Nm]
Tool wear [min]
Target
Failure Type
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8