**DATA DICTIONARY -**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Type** | **Description** |
| UDI | Num | Unique identifier for data |
| Product ID | Char | Unique Product ID based on product types |
| Type | Char | Type of Product with 3 categories - L, M, or H |
| Air temperature [K] | Num | Air Temperature in Kelvin |
| Process temperature [K] | Num | Process Temperature in Kelvin |
| Rotational speed [rpm] | Num | Rotational Speed – number of rotations of the machine per unit of time. Measured in Revolutions Per Minute |
| Torque [Nm] | Num | Torque - A force that causes machinery to turn round. Measured in Newton-Meter |
| Tool wear [min] | Num | Gradual Failure of Cutting tools due to regular operations. Measured in minutes. |
| Target | Char | If the Machine failed or not- 0 - not fail, and 1 - fail |
| Failure Type | Char | Failure Type with 5 categories - No Failure, Heat Dissipation Failure, Power Failure, Overstrain Failure, Tool Wear Failure. |

**INSIGHTS –**

**Univariate Analysis:**

When the distribution of each variable is analyzed following observations were made:

* There are 6000 machines of "L" Type and 3000 machines of "M" Type and nearly 1000 machines are of "H type".
* The Air Temperature for the maximum number of machines is between 297 Kelvin to 303 Kelvin.
* The Process Temperature for the maximum number of machines is between 307 Kelvin to 313 Kelvin.
* The distribution of values of Rotational Speed is right-skewed. The maximum number of machines has Rotational Speed between 1250 to 1750 RPM.
* The value of Torque is approximately normally distributed. The maximum number of values is thus centered around the mean value.
* There are machines that might face failure due to tool wear up to 220 tool wear per minute.

**Bivariate Analysis:**

When the distribution of values of each variable is analyzed with respect to Target (fail or not), the following observations were made:

* Maximum number of failures occur for “L” Product Type.
* Maximum number of failures happens when Air Temperature is high.
* When either the Rotational Speed or Torque is too high or low, failure happens.

**Outlier Treatment:**

While checking for outliers in the data, the following observations were made:

* Outliers were present mainly for two variables, Rotational Speed and Torque.
* Maximum outliers present in these two variables corresponds to No Failure category of Failure Type.

**Correlation:**

While checking for correlation between the variables in the data, the following observations were made:

* There is a high positive correlation between Air Temperature and Process Temperature.
* There is a high negative correlation between Torque and Rotational Speed.

**Pair-Plot**:

While checking for relationship between variables with respect to Target (fail or not), the following observations were made:

* When the Torque is high and Rotational Speed is low (or vice versa), then the machine fails.

**Multicollinearity:**

While checking for multicollinearity, the following observations were made:

* Multicollinearity was checked using VIF. Since the VIF of the variables was not very high, the variables are independent and there is no multicollinearity in the data.

**Model Building Results –**

With the dependent variable as – Target:

|  |  |
| --- | --- |
| Model Name | Accuracy |
| Logistic Regression | 0.971 |
| Decision Tree Classifier | 0.977 |
| Random Forest Classifier | 0.985 |
| Support Vector Classifier | 0.974 |

With the dependent variable as – Failure Type:

|  |  |
| --- | --- |
| Model Name | Accuracy |
| Logistic Regression | 0.972 |
| Decision Tree Classifier | 0.971 |
| Random Forest Classifier | 0.982 |
| Support Vector Classifier | 0.972 |