**Project:** Predictive Maintenance of Industrial Equipment

**Objective:** Use machine learning to predict failures of industrial equipment such as pumps, compressors, or valves based on historical maintenance data, sensor readings, and operational data.

**Project Steps:**

**1. Data Collection:**

* **Database:** Use the Machine Predictive Maintenance Classification dataset available on Kaggle, which allows for predicting machine failures (binary) as well as the type of failure (multiclass).
* **Data Types:** The data includes measurements of temperature, pressure, vibrations, operating hours, and failure history.

**2. Data Preparation:**

* **Data Cleaning:** Address missing values and remove outliers that could skew the analysis. For instance, abnormal temperature or pressure readings might indicate faulty sensors or data entry errors.
* **Feature Engineering:** Create additional variables from raw data, such as the rate of change in temperature or vibrations, which could be indicators of imminent failures.
* **Normalization:** Normalize the data to facilitate model training. For example, pressure and temperature should be scaled to prevent certain features from dominating others.

**3. Data Exploration:**

* **Historical Trend Analysis:** Visualize the data to understand conditions preceding failures. For example, identify if a sudden increase in temperature precedes a failure.
* **Clustering:** Use clustering techniques to identify recurring patterns in the data, such as abnormal vibration patterns associated with certain failures.

**4. Modeling:**

* **Regression Model:** Use a regression model to predict continuous values such as the remaining lifespan of equipment.
* **Classification Model:** Use a binary classification model to predict if equipment will fail within a certain timeframe (e.g., within the next 30 days). Experiment with different models, such as logistic regression, random forests, or neural networks (TensorFlow) to find the best fit for your data.

**5. Model Evaluation:**

* **Model Validation:** Split the data into training and testing sets. Use metrics such as accuracy, recall, and the ROC curve to assess model performance.
* **Cross-Validation:** Perform cross-validation to prevent overfitting and confirm the model's robustness across different data subsets.

**6. Optimization:**

* **Hyperparameter Tuning:** Adjust model hyperparameters to improve results, such as the number of trees in a random forest or the learning rate in a neural network.
* **Regularization:** Incorporate regularization methods to avoid overfitting, especially when models become too complex.

**7. Visualization and Reporting:**

* **Dashboards:** Create dashboards to visualize failure predictions and equipment performance indicators using libraries like Matplotlib, Seaborn, or Plotly for interactive charts.
* **Results Presentation:** Graphs should illustrate failure predictions, historical data trends, and model performance.

**8. Presentation and Documentation:**

* **Documentation:** Prepare a detailed PowerPoint presentation or report explaining the process, results, and practical implications of the project. Highlight how this project can be applied directly in OMV's industrial environment to reduce maintenance costs and optimize equipment reliability.
* **Industrial Implications:** Constantly relate your findings to real-world cases in the oil and gas industry to maximize the relevance of your project. For example, show how improved failure prediction could prevent unplanned shutdowns, saving time and money for the company.

**Let's Get Started!**

**Step 1: Data Collection**

**Step 2: Data Exploration and Preparation**

**Data Overview:**

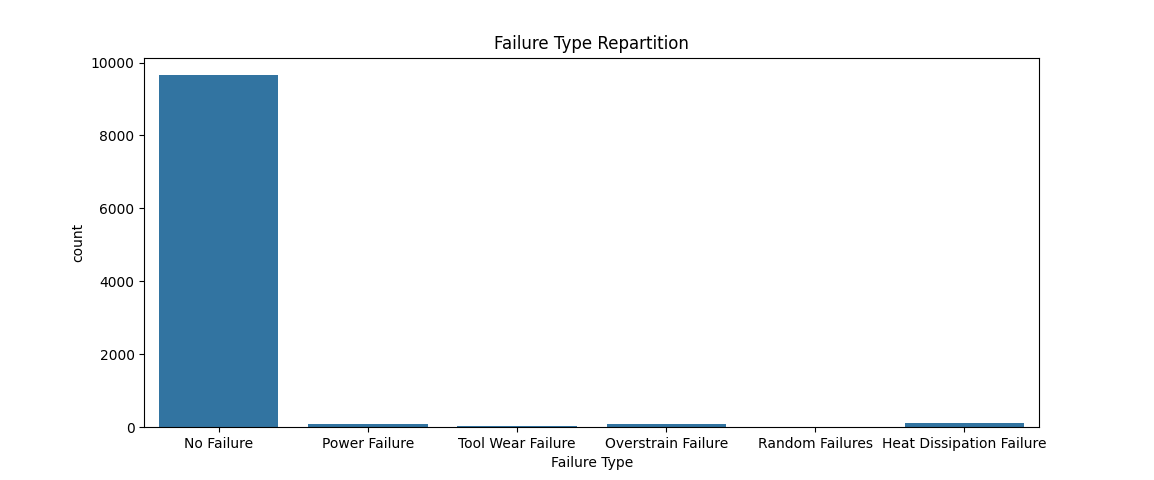
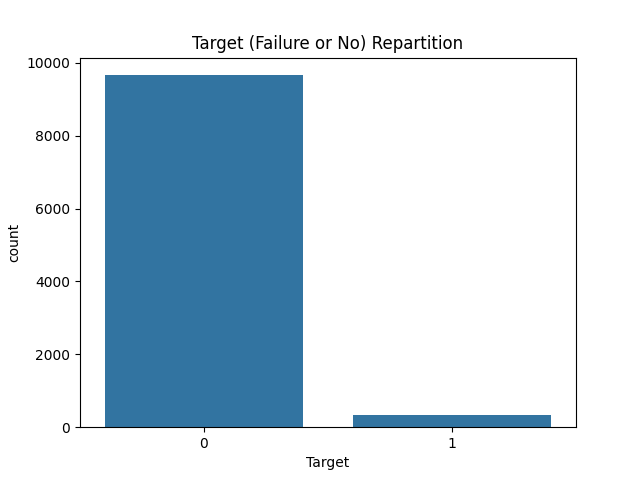
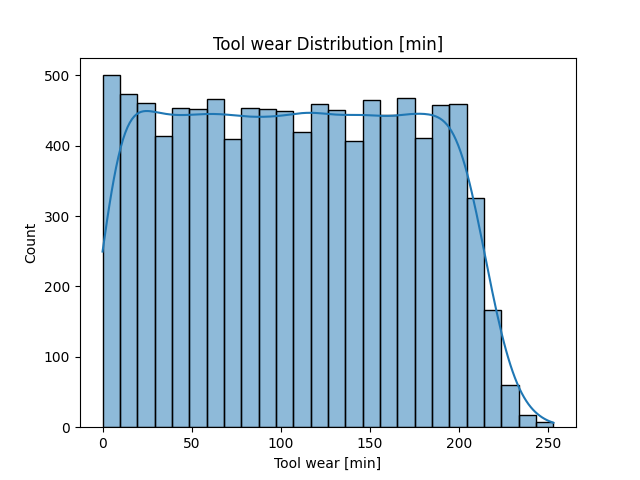
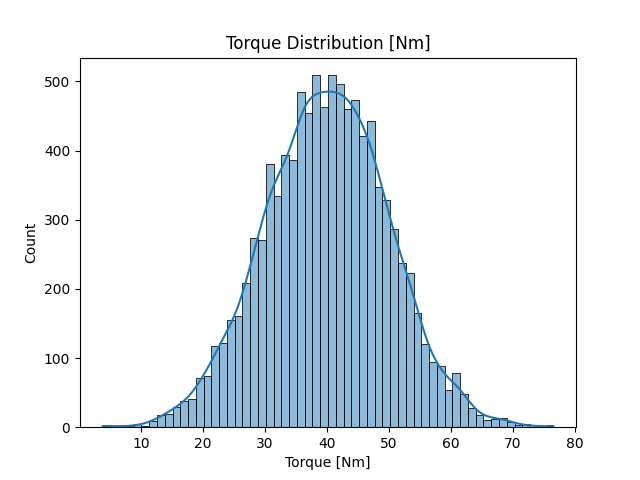
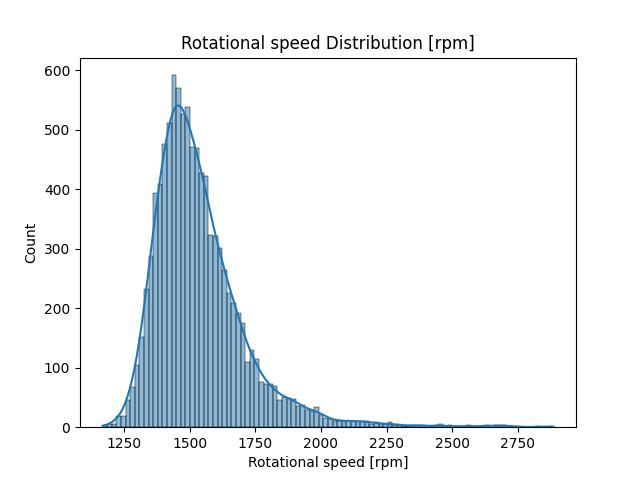
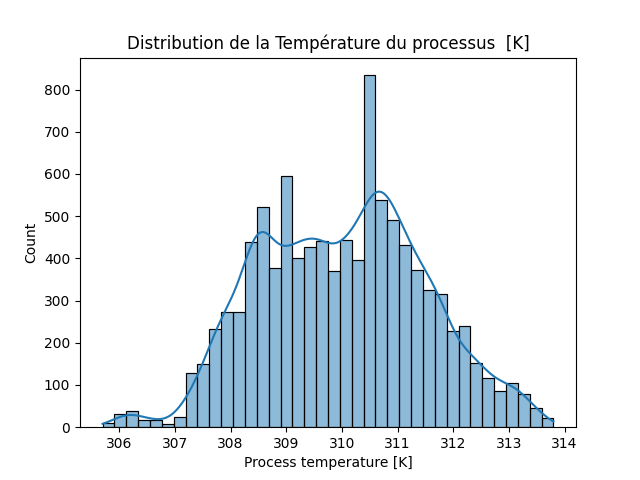
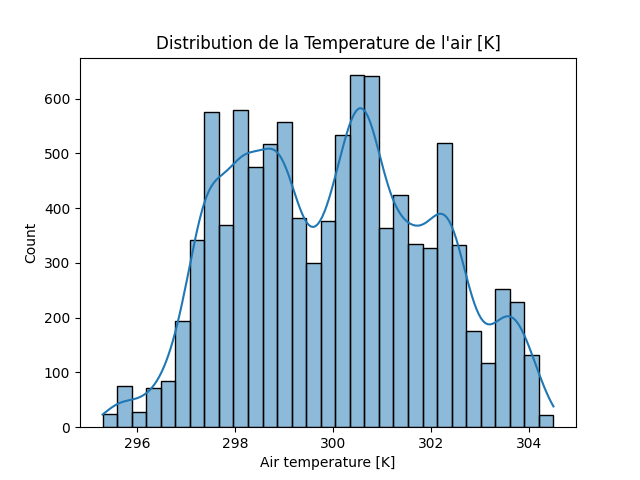
* **Total Number of Rows:** 10,000
* **Total Number of Columns:** 10

**Dataset Columns:**

1. **UDI:** Unique identifier for each record.
2. **Product ID:** Product identifier, including a letter indicating quality (L, M, H) and a specific serial number for the variant.
3. **Type:** Product quality (L for Low, M for Medium, H for High).
4. **Air temperature [K]:** Air temperature in Kelvin.
5. **Process temperature [K]:** Process temperature in Kelvin.
6. **Rotational speed [rpm]:** Rotational speed in revolutions per minute.
7. **Torque [Nm]:** Torque in newton-meters.
8. **Tool wear [min]:** Tool wear in minutes.
9. **Target:** Indicates if the machine failed (1) or not (0).
10. **Failure Type:** Type of failure, or "No Failure" if no failure is present.

**Descriptive Statistics:**

* **Air temperature [K]:** Mean of 300.00 K with a standard deviation of 2.00 K.
* **Process temperature [K]:** Mean of 310.01 K with a standard deviation of 1.48 K.
* **Rotational speed [rpm]:** Mean of 1538.78 RPM with a standard deviation of 179.28 RPM.
* **Torque [Nm]:** Mean of 39.99 Nm with a standard deviation of 9.97 Nm.
* **Tool wear [min]:** Mean of 107.95 minutes with a standard deviation of 63.65 minutes.
* **Target:** Only 3.39% of the data indicates machine failure.

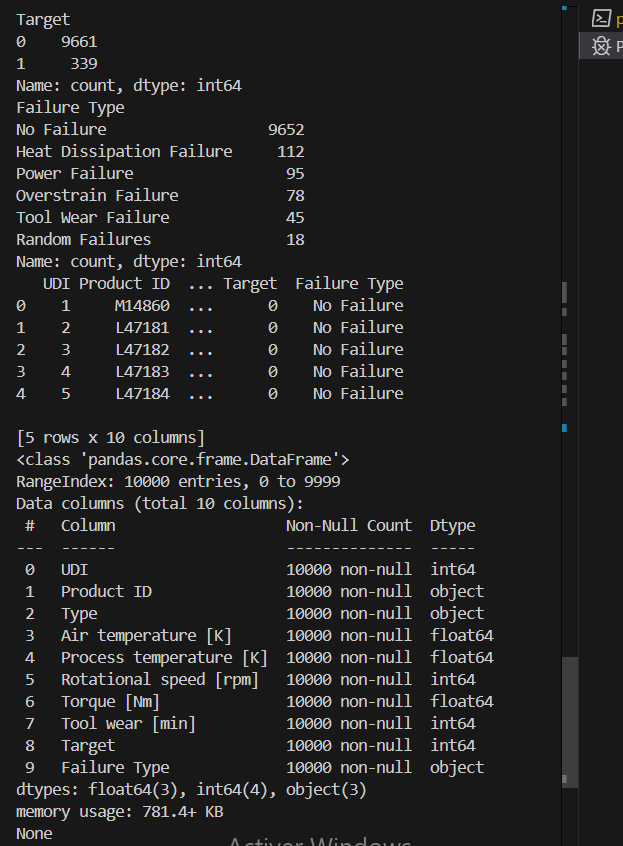


**Observations:**

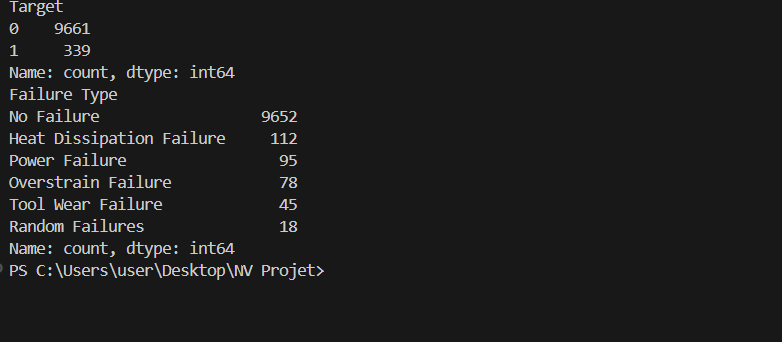
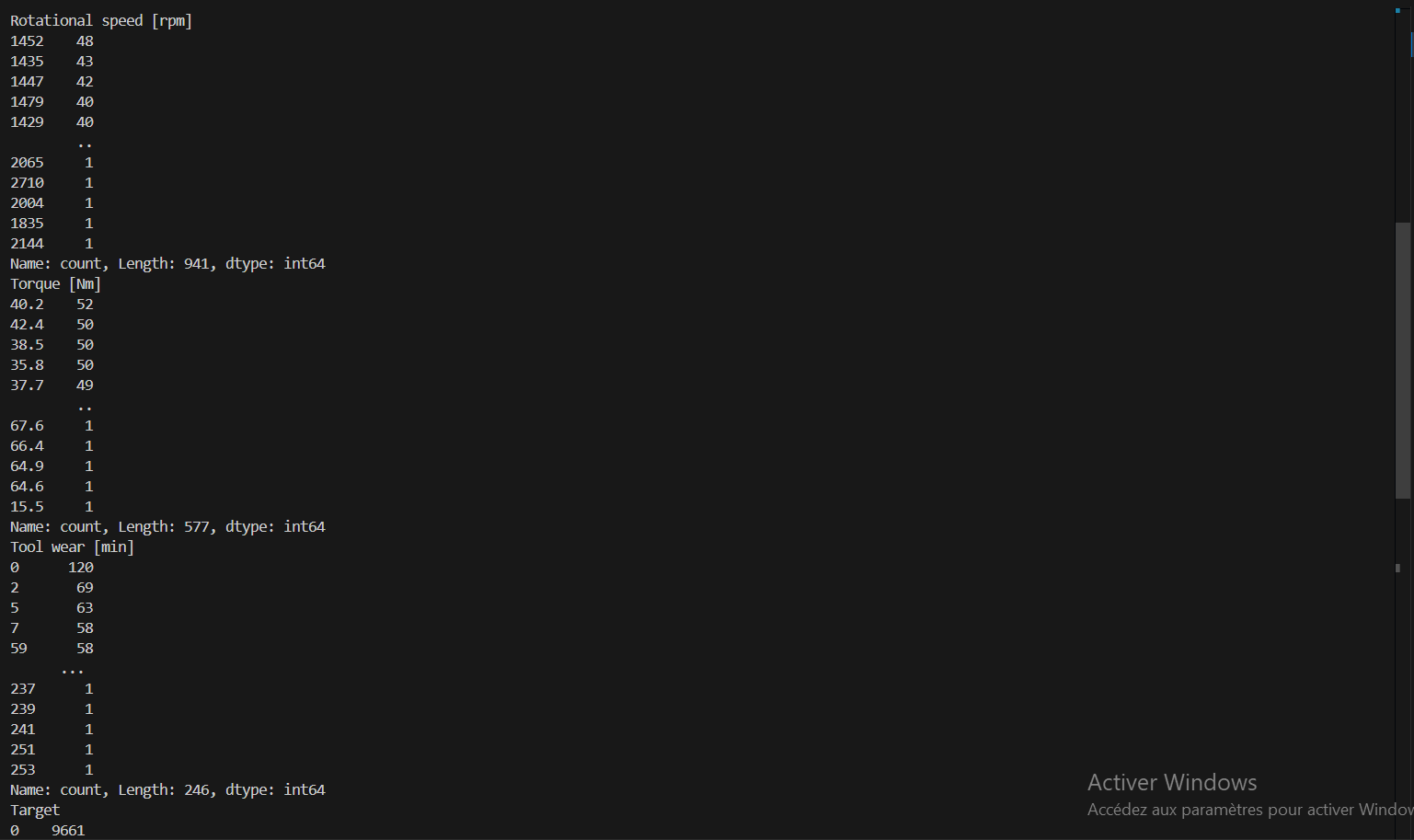
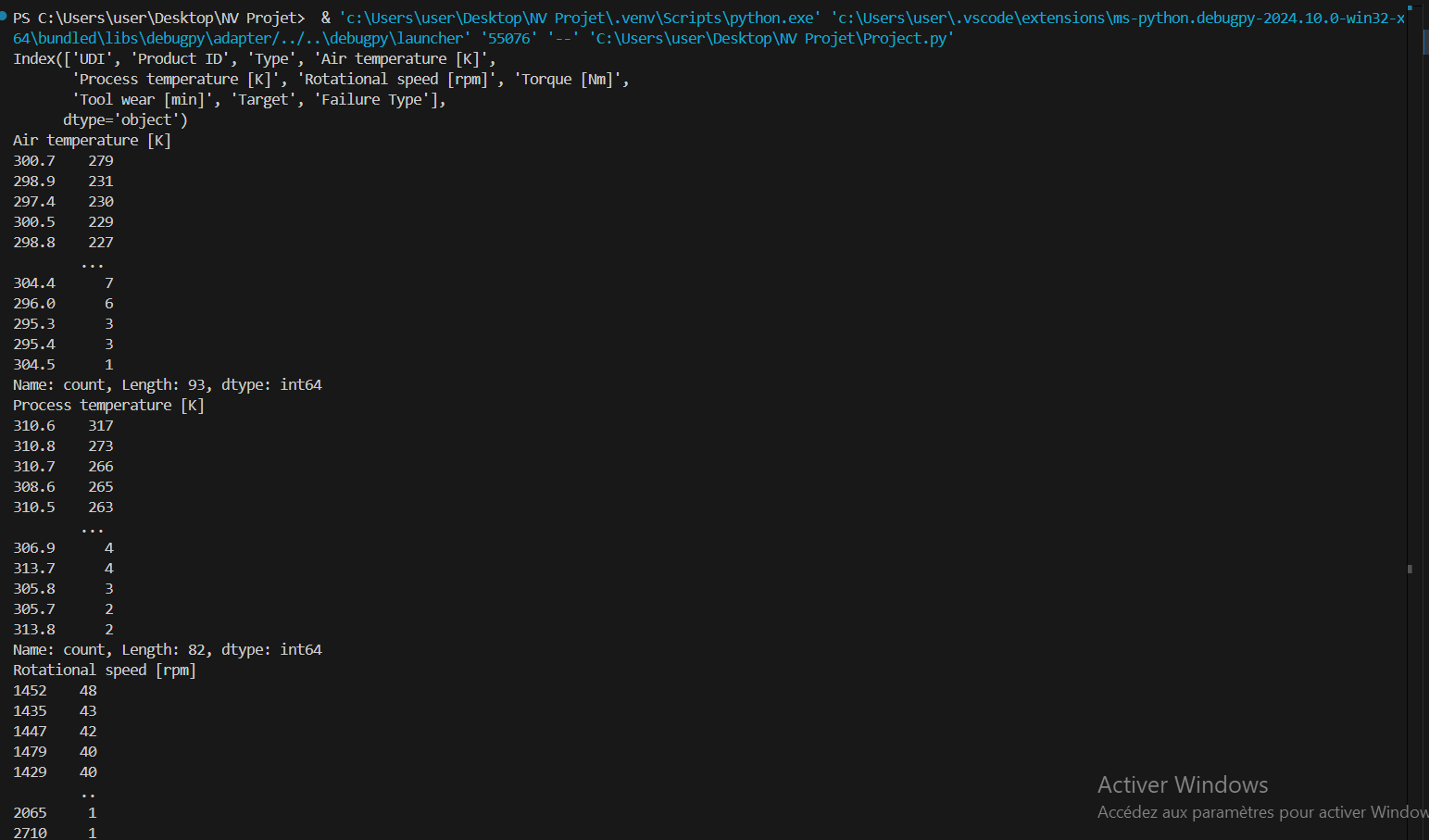
1. **Air Temperature [K]:**
   * **Observation:**
     + The distribution of air temperature is centered around 300 K with slight variance.
     + Most values fall within a narrow range, indicating thermal stability in the operating environment.
   * **Interpretation:**
     + A stable temperature is crucial for proper machine operation. If the temperature remains consistently in this range, it reduces the risk of failures due to thermal stress.
     + Extreme values (anomalies) may indicate periods when the environment was not well-controlled, potentially increasing the risk of failure.
2. **Process Temperature [K]:**
   * **Observation:**
     + The distribution is similar to the air temperature but slightly shifted upward, which is expected since process temperature is usually higher.
     + If the distribution shows a wide variance, it may indicate poor temperature control during the process.
   * **Interpretation:**
     + Significant deviations in process temperature could indicate an increased risk of failures, as excessively high or unstable temperatures can damage machines.
3. **Rotational Speed [rpm]:**
   * **Observation:**
     + The distribution may show a peak around a specific value (e.g., 1500-2000 rpm), with some machines operating at higher or lower speeds.
     + If the distribution is spread out or skewed, it could indicate diversity in machine operations.
   * **Interpretation:**
     + Higher rotational speeds can lead to increased wear on parts, potentially raising failure rates.
     + Low or excessively high rotational speeds, outside the norm, should be monitored as they may indicate misadjustment or improper operation.
4. **Torque [Nm]:**
   * **Observation:**
     + If the torque is normally distributed around 40 Nm, with some extreme values (either very high or very low), it reflects the load applied to the machines.
     + A symmetric distribution around 40 Nm with tails could indicate normal operating conditions.
   * **Interpretation:**
     + Excessive torque might indicate overload, a common cause of mechanical failure.
     + Low torque might suggest that the machine is operating below its capacity, which could be suboptimal for productivity.
5. **Tool Wear [min]:**
   * **Observation:**
     + The distribution of tool wear time should increase over time. If the distribution shows many low values, it could indicate frequent tool replacements.
     + Wide variation might indicate irregular wear due to material quality or operating conditions.
   * **Interpretation:**
     + Tools that wear out more quickly might signal a need for more frequent maintenance or a review of processes to improve their longevity.
     + Tools that last longer than expected could suggest optimal operating conditions.
6. **Machine Failure (Target):**
   * **Observation:**
     + If the target variable shows an imbalance with few failures compared to normal operations, this is typical in predictive maintenance data.
     + A balanced distribution between failures and normal operations would be less common.
   * **Interpretation:**
     + An imbalance in the failure class (e.g., more than 90% non-failures) can make prediction models harder to train as they may become biased toward the majority class.
     + Identifying common causes of failures among the variables could help prevent future incidents.
7. **Failure Type:**
   * **Observation:**
     + Different types of failures may show varied distributions based on other characteristics.
     + If certain failures are much more frequent, they might be directly related to specific conditions (e.g., high temperature associated with motor failures).
   * **Interpretation:**
     + Understanding the most common types of failures would help focus maintenance efforts on critical points, thereby improving overall machine reliability.

**Step 3: Data Preprocessing:**

1. **Explore the Data:**

****

1. **Calculate the Distribution of Categories:**

****

1. **Identify Potential Biases:**

* **Class Frequency:**
  + For the Failure Type column, most observations correspond to "No Failure," while other types of failures are much less frequent. This indicates a class imbalance that could introduce bias if used to train a model.
* **Value Distribution:**
  + For numerical columns such as Air temperature [K], Process temperature [K], and Rotational speed [rpm], values are concentrated around certain ranges. This may suggest that some operating conditions are much more common than others.
* **Class Balancing:**
  + Separate major and minor classes (e.g., different Failure Types).
  + Normalize or standardize numerical data to ensure that features are on a similar scale and reduce bias caused by differences in range.

1. After Cleaning :

