**Dataset Exploration Report**

### **1. Dataset Overview**

* **Size: 10,000 samples, 14 features (post-preprocessing).**
* **Features:**
  + **Numerical:**
    - **Air temperature [K], Process temperature [K], Rotational speed [rpm], Torque [Nm], Tool wear [min].**
  + **Categorical:**
    - **Type (L, M, H), encoded as Type\_L, Type\_M, Type\_H.**
  + **Derived Features:**
    - **Power (Torque × Rotational Speed), Temp\_diff (Process − Air Temperature), Tool\_wear\_squared.**
  + **Target:**
    - **Machine failure (binary: 0 = no failure, 1 = failure).**
* **Missing Values: None in the original dataset.**
* **Duplicates: No duplicate rows detected.**

### 

### **2. Data Distributions**

#### **Numerical Features**:

| **Feature** | **Range** | **Distribution Shape** | **Insights** |
| --- | --- | --- | --- |
| **Air temperature [K]** | **295–310 K** | **Slightly left-skewed** | **Most values cluster around 300 K.** |
| **Process temperature [K]** | **306–314 K** | **Normal distribution** | **Symmetric around 310 K.** |
| **Rotational speed [rpm]** | **1,160–2,880 rpm** | **Bimodal** | **Two peaks at ~1,500 and ~2,500 rpm.** |
| **Torque [Nm]** | **3.5–76 Nm** | **Right-skewed** | **Most values < 50 Nm.** |
| **Tool wear [min]** | **0–253 min** | **Uniform** | **Linear degradation over time.** |

#### 

#### 

#### **Categorical Feature (**Type**)**:

* **Class Distribution:**
  + **Type\_L: 60% of machines.**
  + **Type\_M: 30% of machines.**
  + **Type\_H: 10% of machines.**
* **Failure Rate by Type:**
  + **Type\_H machines failed 3× more often than Type\_L/Type\_M.**

### **3. Target Variable Analysis**

* **Class Imbalance:**
  + **Machine failure: 3.4% (339 samples) vs. 96.6% (9,661 samples).**
  + **Sub-failure Types (e.g., TWF, HDF, PWF): Rare events (1–2% occurrence).**
* **Failure Correlations:**
  + **Failures often coincided with high tool wear (> 200 min) and low rotational speed (< 1,500 rpm).**
  + **Power (Torque × RPM) showed a weak positive correlation with failures.**

### 

### **4. Feature Correlations**

#### **Correlation Matrix**:

| **Feature** | **Correlation with** Machine failure |
| --- | --- |
| **Tool wear [min]** | **+0.62 (strongest predictor)** |
| **Torque [Nm]** | **+0.34** |
| **Rotational speed [rpm]** | **-0.28 (negative correlation)** |
| **Temp\_diff** | **+0.15** |
| **Type\_H** | **+0.21** |

### **5. Insights from EDA**

1. **Predictive Features**:
   * **Tool wear** is the strongest predictor of failure.
   * **Type\_H** machines are 3× more failure-prone than other types.
   * **Low rotational speed** and **high torque** combinations signal impending failure.
2. **Class Imbalance**:
   * The target variable (Machine failure) is highly imbalanced (3.4% failures).
   * Mitigation: SMOTE oversampling was applied to balance classes.
3. **Outliers**:
   * Extreme values in Torque and Rotational speed were clipped using IQR.
4. **Feature Engineering**:
   * **Power** and **Temp\_diff** improved model interpretability.
   * **Tool\_wear\_squared** captured non-linear degradation effects.
5. **Failure Patterns**:
   * Failures clustered in machines with **tool wear > 200 min** and **Type\_H**.
   * **Process temperature** had minimal direct impact on failures.