# **Predictive Maintenance: “Exploratory Data Analysis & Machine Learning”**

**Objective**

The goal of this project is to perform **Exploratory Data Analysis (EDA)** and build **predictive models** to detect machine failures using sensor data. This helps with reducing unexpected downtimes and improving maintenance strategies.

**Dataset**

* **File Used:** ai4i2020.csv
* **Source:** Predictive maintenance dataset containing various machine parameters such as temperature, rotational speed, torque, tool wear, and failure history.
* **Target Variable:** Machine failure

**Goals of the Analysis**

1. **Understand the dataset** by summarizing key statistics and visualizing data distributions.
2. **Identify patterns and correlations** between machine parameters and failures.
3. **Detect key features** that contribute to machine failures.
4. **Build predictive models** (Logistic Regression, Random Forest) to classify failures.
5. **Optimize model performance** using hyperparameter tuning.
6. **Visualize feature importance** to understand critical failure indicators.

**Key Questions Answered & Justifications**

**1. Data Understanding & Preprocessing**

* **Are there any missing values in the dataset?**
  + No missing values were found in the dataset, ensuring data completeness.
* **What are the summary statistics of the dataset?**
  + Summary statistics revealed that temperature, rotational speed, and torque vary within reasonable ranges. Mean and standard deviation analysis helped in understanding normal operational conditions.
* **What is the distribution of machine failures?**
  + The dataset is highly imbalanced, with a significantly lower number of failures compared to non-failures. This highlights the need for techniques like class balancing in modeling.
* **What is the distribution of product types in the dataset?**
  + The dataset contains multiple product types (L, M, H), with approximately even distribution. Certain product types exhibit higher failure rates, which is explored further in feature analysis.

**2. Feature Relationships & Correlations**

* **How are machine parameters correlated with failures?**
  + The correlation matrix indicated that higher tool wear and rotational speed are moderately correlated with machine failures.
* **What are the key differences in rotational speed, torque, and temperature for failed vs. non-failed machines?**
  + Failed machines tend to operate at higher rotational speeds and have greater tool wear compared to non-failed ones. However, temperature differences were minimal.
* **Which product types have a higher failure rate?**
  + Product type **H** exhibited the highest failure rate compared to L and M, suggesting that this category requires more maintenance attention.

**3. Machine Learning Models & Evaluation**

* **How accurately can we predict machine failures using Logistic Regression & Random Forest models?**
  + Both models achieved high accuracy (~99.9%), with Random Forest performing slightly better than Logistic Regression.
* **What is the precision, recall, and F1-scores for the models?**
  + Precision and recall scores for the failure class (~1.00 and 0.97, respectively) indicate that models can reliably detect machine failures while minimizing false positives.
* **How does hyperparameter tuning improve model performance?**
  + GridSearchCV tuning improved model performance by optimizing parameters like the number of estimators and max depth in Random Forest, leading to a more robust classifier.

**4. Feature Importance & Insights**

* **Which machine parameters contribute the most to failures?**
  + An analysis of the feature importance from Random Forest revealed that **tool wear, rotational speed, and torque** were the most significant predictors of failure.
* **How can predictive models help optimize maintenance schedules?**
  + Predictive models allow early detection of potential failures, enabling **proactive maintenance strategies**. This minimizes downtime, reduces maintenance costs, and improves overall machine efficiency.

**Technologies & Libraries Used**

* **Python**
* **Pandas, NumPy** (Data processing)
* **Seaborn, Matplotlib** (Data visualization)
* **Scikit-Learn** (Machine Learning)
* **GridSearchCV** (Hyperparameter tuning)

**Results & Findings**

* **Machine failures are rare (high class imbalance).**
* **Temperature, rotational speed, and tool wear are key indicators of failure.**
* **Random Forest performed best, achieving 99.9% accuracy.**
* **Feature importance analysis helped in identifying critical machine parameters.**