**Project Overview**

**Project Title:**

Enhancing Equipment Reliability with Machine Learning-Based Predictive Maintenance

**Project Summary and Background:**

Predictive maintenance aims to forecast equipment failures by analyzing sensor and operational data, helping reduce downtime and costs. This project focuses on the AI4I 2020 Predictive Maintenance dataset, which simulates industrial milling machine operations and records multiple failure modes (Paolanti et al., 2018). The dataset comprises sensor readings such as air temperature, process temperature, rotational speed, torque, and tool wear, along with a binary label indicating machine failure. This project applies advanced machine learning models XGBoost, LightGBM, CatBoost, and a Multilayer Perceptron (MLP) to predict machine failures, comparing their performance to identify the best predictive approach (Amer et al., 2023). The project will utilize SHAP (SHapley Additive exPlanations) as the primary Explainable AI (XAI) technique.

**Research Question:**

* How effectively can ensemble gradient boosting models (XGBoost, LightGBM, CatBoost) predict machine failures in the AI4I 2020 dataset compared to neural network approaches such as the Multilayer Perceptron (MLP)?
* Which features contribute most significantly to the prediction of machine failures in industrial milling machines?

**Project Objectives:**

* Conduct exploratory data analysis and preprocess the AI4I 2020 dataset.
* Implement and optimize XGBoost, LightGBM, CatBoost, and MLP models for failure prediction.
* Evaluate models using metrics including accuracy, precision, recall, and F1-score.
* Analyze feature importance to provide insights into key failure indicators.
* Ensure reproducibility through structured version control and documentation.

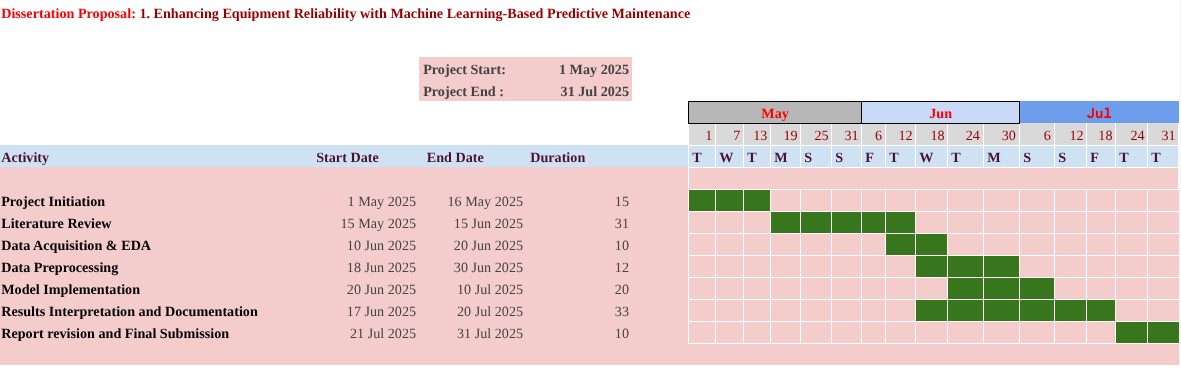
**References:**

Predictive Maintenance Dataset (AI4I 2020). (2022, November 6). Kaggle. https://www.kaggle.com/datasets/stephanmatzka/predictive-maintenance-dataset-ai4i-2020?resource=downloadhttps://www.kaggle.com/datasets/stephanmatzka/predictive-maintenance-dataset-ai4i-2020?resource=download

M. Paolanti, L. Romeo, A. Felicetti, A. Mancini, E. Frontoni and J. Loncarski, "Machine Learning approach for Predictive Maintenance in Industry 4.0," 2018 14th IEEE/ASME International Conference on Mechatronic and Embedded Systems and Applications (MESA), Oulu, Finland, 2018, pp. 1-6, doi: 10.1109/MESA.2018.8449150.

S. Amer, H. k. Mohamed and M. Badr Monir Mansour, "Predictive Maintenance by Machine Learning Methods," 2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS), Cairo, Egypt, 2023, pp. 58-66, doi: 10.1109/ICICIS58388.2023.10391130.

## **2. Project Plan: Task List and Timeline**



## **3. Data Management Plan**

### **Overview of the Dataset**

The AI4I 2020 Predictive Maintenance Dataset contains 10,000 records of milling machine sensor data collected to simulate operational conditions and failures. Features include:

* **UID:** Unique record identifier (1 to 10,000).
* **Product ID and Type:** Product quality variants (L=low 50%, M=medium 30%, H=high 20%).
* **Air Temperature [K]:** Random walk with SD 2 K around 300 K.
* **Process Temperature [K]:** Air temperature + 10 K plus noise (SD 1 K).
* **Rotational Speed [rpm]:** Derived from power with noise.
* **Torque [Nm]:** Normal distribution around 40 Nm with SD 10 Nm.
* **Tool Wear [min]:** Varies by quality variant (L=2, M=3, H=5 minutes).
* **Machine Failure Label:** Binary indicator (1 = failure; 0 = no failure), combining five failure modes (tool wear failure, heat dissipation failure, power failure, overstrain failure, random failures).

The Dataset was originally created by S. Matzka for the AI4I 2020 Third International Conference to promote explainable AI research in predictive maintenance.

### **Data Collection**

The dataset used in this project is the **AI4I 2020 Predictive Maintenance Dataset**, sourced from the AI4I 2020 conference and available on Kaggle. It contains structured data from a simulated milling process, including sensor readings and machine failure labels (Predictive Maintenance Dataset (AI4I 2020), 2022).

### **Metadata**

* Format: CSV files, approximately 10 MB in size.
* Number of records: 10,000 rows with 10 columns (features + label).
* Code files: Python scripts for data processing, modeling, and evaluation (~2 MB).

### **Document Control and Versioning**

* All code and documentation will be managed on GitHub: https://github.com/Manikumar1999/Predictive\_maintenance\_project
* Weekly commits will track incremental progress, including preprocessing, model development, and evaluation.
* **Version Control:** Version numbers will follow a semantic numbering system (e.g., v1.0, v1.1, etc.).
* **GitHub Settings:** The repository will be set to accessible, ensuring that markers can view the code throughout the evaluation process.

### **ReadMe File**

The final GitHub ReadMe file will include:

* Overview of the project and dataset.
* Instructions to run code scripts and reproduce results.
* Description of dataset features and preprocessing steps.
* Summary of implemented models and evaluation metrics.
* Contact information and citation instructions.

### **Security and Storage**

* **Backup Frequency:** Data will be uploaded once and project code files will be backed up weekly.
* **Data Sharing:** Files will be shared securely via GitHub, where access will be granted to relevant markers. Additionally, shared access to project files will be provided through Google Drive.
* **File Storage Locations: (**Primary: GitHub Repository), (Secondary: Google Drive)

### **Ethical Requirements**

1. **GDPR Compliance:**The dataset contains no personal or sensitive data; thus, GDPR does not apply.
2. **Conformance to UH Ethical Policies:**The project follows University of Hertfordshire ethical guidelines by using publicly available, anonymized industrial data for research.
3. **Permission to Use Data:**The AI4I dataset is publicly released for academic research and is cited accordingly.
4. **Ethical Data Collection Assurance:**The dataset creators (S. Matzka) have documented the synthetic and real components of data generation transparently in the referenced publication, ensuring ethical use.