**PrognosAI: Predictive Maintenance System**

**Milestone 1 Report: Data Ingestion and Preprocessing**

**By- Darshna Ujwal**

**1. Introduction and Goal Setting**

This phase established the foundational data pipeline for the PrognosAI system. Our primary goal was to take the raw, comma-separated NASA CMAPSS FD001 sensor data and transform it into a clean, scaled dataset that is directly usable by our deep learning model. Success in this milestone means the data is correctly structured and ready for sequence generation.

**2. Data Loading and Initial Cleanup**

We began by loading the raw data (train\_FD001.txt). Since the data is space-separated, we used Pandas to correctly label the columns, including unit IDs, cycle counts, three operational settings, and 21 sensor readings.

Two immediate cleaning steps were necessary:

* **Handling Empty Columns:** We stripped away any columns that were entirely empty (NaN), which often appear due to how the original text file is parsed.
* **Removing Static Sensors:** For time-series analysis, features that never change have no predictive value. We calculated the standard deviation for all sensors and removed the seven columns that showed zero variance. These static readings would have only added noise to the training process.

**3. Feature Engineering: Calculating RUL**

The core outcome of this phase was creating the **Remaining Useful Life (RUL)** target variable. The raw data only tracks the cycle count, not how many cycles are left until failure.

To generate the RUL for every row, we used the simple principle: **RUL = (Total life cycles for that engine) - (Current cycle number)**. This gave us a continuous degradation profile, where RUL counts down to zero at the point of failure.

**4. Preparing for Deep Learning: Min-Max Scaling**

Neural networks perform best when input data is normalized. We applied Min-Max Scaling to all dynamic sensor readings and operational settings, setting their values within a standardized range of [0, 1].

* This scaling stabilizes the training process for the LSTM.
* We deliberately excluded the engine ID, cycle count, and the RUL target from scaling to preserve their original meaning.

**5. Conclusion and Ready State**

Milestone 1 is complete. We have successfully executed the data preparation pipeline. The fully processed and scaled dataset, named **train\_FD001\_processed.csv**, has been saved to the data/processed/ directory.

The project is now ready to move directly into Milestone 2: Sequence Generation, where we will structure this data into the rolling windows required by the LSTM model.