Week 1 – Data Generation and Exploratory Data Analysis

Introduction:

Data gathered from several sensors built into industrial machinery is crucial in today's production settings. These data streams record vital factors that can reveal the operational health and possible failure of equipment, including temperature, vibration, RPM, humidity, pressure, and voltage. Exploratory data analysis (EDA), a fundamental stage in the machine learning pipeline, is the focus of week one of the SmartFactory AI project. Its goals are to comprehend sensor behavior, identify trends, and validate presumptions required for downstream predictive modeling.

Objective:

Week 1's main objectives are to:

- Learn about the raw sensor dataset both intuitively and statistically.
- Examine the relationships between machine failures and sensor data.
- Determine anomalies or outliers by visualizing feature distributions.
- Show possible warning signs or early warning signs of equipment failure.
- In the coming weeks, lay the foundation for feature engineering.

Methodology:

The actions listed below are methodically carried out by the Week 1 script:

- **Preparation and Data Loading:**
  Bringing in factory_sensor_data.json, transforming it into a pandas DataFrame, and making sure the timestamp field is properly interpreted as a datetime object are the first steps in the script. In order to preserve temporal integrity, the data is arranged chronologically.
- **Descriptive Statistics:**
  The script uses .describe() to produce summary statistics for each numerical sensor characteristic. This helps to promptly identify anomalies (such as abnormally high vibration levels or negative voltages) by providing important data including mean, standard deviation, min, and max.

- **Correlation Heatmap:**
  All of the numerical features' pairwise Pearson correlation coefficients are presented as a heatmap. This finds probable causative linkages (e.g., high pressure before failure) and redundancy (e.g., strong correlation between RPM and temperature).
- **Feature Distribution Analysis:**
  For each of the main sensor characteristics, distribution plots—histograms superimposed with KDE curves—are produced. In order to choose the best transformation methods later on, these graphs aid in evaluating the skewness, modality, and variance of sensor values.
- **Boxplots by Failure Class:**
  Boxplots are used to depict each sensor feature in relation to the binary failure class. When failure happens, this makes it possible to visually see changes in distributions, such as a higher median temperature in failed samples.
- **Sensor Features Pairplot:**
  To prevent overplotting, a pairplot is made from a subset of 1000 samples. The hue='failure' parameter shows the feature space clustering of failed vs non-failed data points. Using raw sensor data, this provides preliminary insight into the separability of failure.
- **Failure Counts by Machine:**
  The top 20 machines with the greatest failure counts are displayed in a bar chart. This indicates whether production flaws, usage levels, or environmental variations may be the cause of failures that are concentrated in particular units.
- **Daily Average Temperature Trend:**
  The daily average temperature is plotted as a time series. This long-term trend can assist in detecting times of anomalous temperature conditions and can also indicate seasonal or process-related drifts.
- **Distributions of Temperature and Vibration by Failure Status:**
  Temperature and vibration values between failure and non-failure samples are compared in separate KDE plots. Finding characteristics whose distributions drastically change under failure situations requires the use of these comparison charts.
- **Temperature vs. Vibration Scatter Plot:**
  A temperature vs. vibration scatterplot for a 2000-row sample, colored by failure state, shows possible coupled behavior of the two characteristics under failure. Multivariate feature engineering may be preceded by clustering patterns.

- **RPM Boxplot by Failure:**
  A targeted boxplot illustrates the variations in RPM values under various failure scenarios. This aids in confirming if RPM can be utilized in predictive modeling as an engineered or direct characteristic.
- **Voltage vs. Humidity by Failure Status:**
  A scatterplot illustrates how voltage and humidity interact under various failure scenarios. For PCA or dimensionality reduction, these compound connections may provide useful information.
- **Temperature Trend for Top Failed Machines:**
  The top six machines (those with the highest failure rates) are plotted as time series. These charts show temperature anomalies unique to a machine, cyclical patterns, or abrupt shifts that point to pre-failure stages.

Summary:

Several important conclusions are drawn from the exploratory data analysis carried out in Week 1:

- RPM, temperature, and vibration all exhibit significant fluctuation with failure occurrences, which makes them attractive options for feature engineering.
- Under failure conditions, the temperature and vibration distributions change dramatically, with fatter tails and higher means.
- A small number of machines contribute disproportionately to failure incidents, which may be a sign of concentrated operational problems.
- Moderately positive correlations between several sensor characteristics are shown by correlation heatmaps, indicating that dimensionality reduction or sensor fusion may be beneficial.
- Time series analysis reveals intermittent spikes in some machines and slow drifts in the average temperature, indicating the importance of temporal aspects (lags, rolling statistics).
- In multivariate feature space, scatterplots and pairplots show both linear and non-linear separability between failed and non-failed classes.

Conclusion:

The analysis from Week 1 effectively creates a statistical and visual comprehension of the unprocessed sensor data. It draws attention to the possibilities of predictive maintenance through the use of sensor correlations, machine-specific patterns, and time-based characteristics. These revelations set the stage for Week 2, when more thorough

preprocessing and feature engineering will start to turn these EDA results into forms that are ready for models.

URL to Week 1 – Data Generation and Exploratory Data Analysis.ipynb