PRACTICAL WORK-2

AI4I Predictive Maintenance –
https://archive.ics.uci.edu/dataset/601/ai4i+2020+predictive+maintenance+dataset

---

Requirements
Submit - one .ipynb file only
Clear structure
Notebook must run **top → bottom without errors**

Scenario

You are hired as a **Data Analyst in a heavy mining company**.

Management wants to implement a **Predictive Maintenance System** to reduce:

- Unexpected machine breakdown
- Production downtime
- Safety risks
- Maintenance cost

Before Data Scientists build models, your task is:

> Audit, analyze, and prepare the industrial sensor dataset for ML deployment.

You are responsible for data quality, reliability, and industrial risk interpretation.

PHASE 1 — Data Audit & Risk Evaluation (Industrial Thinking)

Task 1: Data Integrity Check

1. Check missing values.
2. Detect impossible values (domain reasoning):
    - Negative torque?
    - Temperature outside physical range?
3. Check duplicated records.
4. Check inconsistent failure flags.

Question:

If this data feeds a real-time control system, what risks exist?

Deliverable:
Short industrial audit summary (5–7 bullet points).

PHASE 2 — Sensor Behavior Analysis

You must treat each variable as a real industrial sensor.

For each numeric feature:

1. Analyze distribution.
2. Compute dispersion (std, IQR).
3. Identify extreme operating zones.
4. Compare sensor behavior for:
    - Normal machines
    - Failed machines

Questions:

- Which sensors behave differently before failure?
- Which variables may be early warning signals?

Deliverable:
Table + interpretation paragraph.

PHASE 3 — Industrial Anomaly Detection

Instead of just removing outliers:

1. Detect anomalies using:
    - Z-score
    - IQR
2. Compare anomaly records with failure label.

Questions:

- Are anomalies noise or early breakdown signals?
- Should we remove them or mark them as risk indicators?

Deliverable:
Decision with justification.

PHASE 4 — Risk Segmentation (Business-Oriented)

Create industrial risk zones.

Examples:

- Tool wear:
  - 0–100 → Normal
  - 100–200 → Maintenance soon
  - 200+ → Critical
- Temperature difference (Process – Air):
  - Small → Stable
  - Medium → Warning
  - Large → Overheating risk

Then:

1. Create risk categories.
2. Calculate failure rate per risk zone.

Question:

Does risk segmentation meaningfully separate failure probability?

Deliverable:
Risk segmentation table + chart.

PHASE 5 — Data Preparation for Production

1. Apply scaling (justify choice).
2. Check class imbalance.
3. Apply 30% sampling.
4. Compare sample vs full dataset statistics.

Question:

If this dataset is used for model training, what bias risks exist?