**1. Introduction**

In this task, I worked on fault prediction for wind farm data. The main aim was to compare **supervised learning (Random Forest)** with **unsupervised learning (Isolation Forest)** and decide which one works better.  
The dataset had time-based readings from different sensors, and a target column was created to indicate whether a failure occurred or not.

**2. Data Preparation**

* First, I loaded the CSV file containing wind turbine sensor data.
* I identified the **status\_type\_id** column as the label column, where non-zero values represent failures.
* A new **target column** was created to make the problem binary (0 = normal, 1 = failure).
* I cleaned the dataset by removing unwanted columns (like IDs and timestamps), filled missing values with the median, and dropped useless constant features.
* The data was then split into **70% training** and **30% testing** to evaluate the models.

**3. Models Used**

**3.1 Supervised Approach (Random Forest)**

* Random Forest was trained on the labeled data (train split).
* Since failures are rare, I used **class\_weight="balanced"** to handle imbalance.
* Evaluation was done on the test set using Precision, Recall, F1-score, and False Alarm Rate.

**3.2 Unsupervised Approach (Isolation Forest)**

* Isolation Forest was trained only on **normal (healthy) data**.
* It then predicted anomalies on the test data.
* Same metrics were calculated for comparison.

**4. Results**

| **Model** | **Precision** | **Recall** | **F1-score** | **False Alarm Rate** |
| --- | --- | --- | --- | --- |
| Supervised (Random Forest) | **High (close to 1)** | **Good** | **Balanced** | Low |
| Unsupervised (Isolation Forest) | High Precision but | Very Low Recall | Very Low F1 | Higher |

🔎 **Observation:**

* Random Forest detected a good number of failures (high Recall) while keeping false alarms under control.
* Isolation Forest had very poor recall (it missed most failures), even though its precision was high.

**5. Graphical Comparison**

I plotted a bar chart comparing **Precision, Recall, and F1-score** for both models.

* The graph clearly shows that **Random Forest performs better overall**.
* Isolation Forest almost fails in recall, meaning it cannot catch failures in advance.

**6. Conclusion**

From this task, I learned how both supervised and unsupervised methods can be applied to anomaly detection in wind farm data.

* **Supervised Random Forest** turned out to be the **better model** because it achieved higher recall and balanced F1-score.
* **Unsupervised Isolation Forest** was not reliable as it missed most failures, even though it gave fewer false alarms.

👉 **Final Answer:** For this dataset and task, **Supervised Learning (Random Forest)** is the best choice.