PV System Anomaly Detection

1. Objective and Problem Statement

The task is to develop an unsupervised machine learning solution to detect anomalies in a PV system using time-series data of actual (P) and expected (P_exp) MPPT power at 1-minute intervals. Key requirements:

- Define faulty vs. normal data points based on power deviation patterns.
- Select and justify an unsupervised model.
- Train without explicit labels.
- Identify and visualize anomalies in the dataset.

2. Solution Approach and Methodology

Data Preprocessing

- **Data Cleaning**:
  Loaded and parsed datetime, handled missing values, and ensured 1-minute interval consistency.
- **Feature Engineering**:
  Created four key features to capture anomalies:
  a. `Delta_P = P_exp - P` (absolute deviation)
  b. `Ratio_P = P / P_exp` (relative deviation)
  c. `P_diff = P.diff()` (rate of change of actual power)
  d. `P_exp_diff = P_exp.diff()` (rate of change of expected power)
  These features quantify deviations in magnitude and trend consistency.

Model Selection: Isolation Forest

- **Justification**:
  - Unsupervised learning (no labels available).
  - Effective for detecting rare anomalies in high-dimensional data.
  - Computationally efficient with low memory requirements.
  - Robust to outliers by design.
- **Model Configuration**: `IsolationForest(contamination=0.01, random_state=42)`
  - `contamination=0.01` assumes ~1% of points are anomalies (adjustable based on domain knowledge).

Training Process

1. **Unsupervised Training**:
   - The model learns the "normal" feature distribution.
   - Anomalies are points easily isolated (fewer splits required).
2. **Anomaly Scoring**:
   - `model.fit_predict(features)` flags anomalies as `-1` (converted to `1` for clarity).
3. **No Explicit Labels Needed**:
   - Relies on the assumption that anomalies are statistically rare and distinct from normal data.

3. Key Findings

1. Data Distribution and Fault Definition

- **Data Distribution**:
    - Most points cluster where $P \approx P_{exp}$ (high Ratio_P, low Delta_P).
    - Occasional large deviations indicate potential faults (e.g., shading, dust, hardware issues).
- **Fault Definition**:
  A point is **faulty** if:
    - Delta_P or |P_diff - P_exp_diff| exceeds dynamic thresholds (learned by the model).
    - Ratio_P is abnormally low (e.g., < 0.8) or volatile.
- **Model Choice Justification**:
  Isolation Forest handles unlabeled data, ignores noise, and scales well for time-series.

2. Training Without Labels

- **Strategy**:
  The model infers anomalies solely from feature space density. Dense regions = normal; sparse regions = anomalies.
- **Assumptions**:
    - Anomalies are rare (< 1% of data).
    - Faults manifest as sudden drops in Ratio_P or inconsistent power-change rates.

3. Anomaly Detection Results

- **Visualization**:

  Actual vs. expected power with anomalies (red) highlighted.

- **Key Observations**:
  - Anomalies correlate with:
    - Sharp drops in P (e.g., from 900W to 300W while P_{exp} remains stable).
    - Sustained deviations (e.g., P consistently 20% below P_{exp}).
  - Example: On 10/1/2023 11:00, $P_{exp} = 1803.9W$ but $P = 1303.9W$ (Ratio_P ≈ 0.72).

4. Assumptions and Limitations

- **Assumptions**:
  - Clear-sky conditions dominate; anomalies imply system faults.
  - No environmental data (irradiance/temperature) used.
- **Limitations**:
  - May flag rapid weather changes as false positives.
  - Cannot classify fault types (shading vs. inverter failure).
- **Improvements**:
  - Add rolling-window features (e.g., 10-min mean/std).
  - Incorporate weather data to reduce false positives.
  - Use clustering (e.g., DBSCAN) for anomaly type identification.

5. Conclusion

The Isolation Forest model successfully identifies PV system anomalies by leveraging feature-engineered deviations in power data. This approach provides a scalable, unsupervised solution for real-time fault detection, enabling proactive maintenance.