Anomaly Detection in Time Series Data

1. Objective

The aim of this project is to detect anomalies in CPU utilization time series data using statistical methods and machine learning algorithms.
The dataset consists of timestamped CPU usage values, and the goal is to identify unusual usage patterns that may indicate issues.

---

2. Dataset

- **Source**: ec2_cpu_utilization.csv
- **Features**:
    - timestamp: Time of CPU usage measurement.
    - value: CPU utilization value.
- **Known anomalies**: Provided at timestamps:
    - 2014-02-26 22:05:00
    - 2014-02-27 17:15:00

---

3. Methods Implemented

3.1 Manual Labeling of Known Anomalies

- All timestamps except the known anomalies are labeled as **Inliers (1)**.
- Known anomaly timestamps are labeled as **Anomalies (-1)**.

3.2 Visualization

- Scatter plots of CPU usage showing anomalies in red and inliers in blue.
- Kernel Density Estimation (KDE) plot to understand the distribution of CPU usage.

3.4 Isolation Forest

- **Training Data**: First 3550 rows.
- **Test Data**: Remaining rows.
- **Contamination Rate**: 1len(train)\frac{1}{\text{len(train)}}len(train)1
- Detects anomalies by isolating points with fewer splits in a decision tree.

3.5 Local Outlier Factor (LOF)

- LOF detects anomalies based on local density deviations compared to neighbors.

---

4. Evaluation

- **Metric Used**: Confusion Matrix (True Positives, True Negatives, False Positives, False Negatives).
- Evaluations were performed for:
    - MAD method vs. manual labels.
    - Isolation Forest predictions vs. manual labels.

---

5. Observations

- The MAD method works well when data distribution is stable and anomalies are far from the median.
- Isolation Forest adapts better when anomalies are subtle and not extremely deviant.
- LOF is suitable for local density-based anomaly detection but may overfit in high noise data.

---

6. Conclusion

- Multiple anomaly detection methods were tested to handle both obvious and subtle anomalies.
- Combining statistical methods (MAD) with machine learning models (Isolation Forest, LOF) can provide robust anomaly detection in time series.
- This approach is applicable for real-time monitoring of server CPU utilization.