**Report: Analysis and Optimization of k-NN Based Anomaly Detection for Bus Data**

**1. Introduction**

This report presents the results and analysis of an anomaly detection experiment conducted using k-nearest neighbors (k-NN) on bus data. The goal was to detect anomalies in the time series data of various buses in a smart grid using a CUSUM-like algorithm. Different values for parameters such as the number of neighbors k, tail probability threshold alpha, and the CUSUM threshold h were explored to determine their impact on the detection performance.

**2. Experimental Setup**

* **Dataset**: The dataset consists of time series data from several buses (115, 116, 117, 118, 119, 121, 135, 139), with the last 855 data points used for the analysis.
* **Parameters**:
  + k: Number of neighbors used in the k-NN algorithm, tested with values [5, 10, 15].
  + alpha: Tail probability threshold for detecting outliers, tested with values [0.001, 0.01, 0.05, 0.1].
  + h: CUSUM threshold, tested with values [5, 10, 20, 30, 50].

**3. Results**

The following observations were made based on the results obtained:

* **Precision and Recall**:
  + Most buses showed very high recall (nearly 1.0), indicating that the model successfully detected most of the actual anomalies.
  + However, precision was generally low, particularly for buses 119, 121, 135, and 139. This suggests a high number of false positives, where normal data points were incorrectly flagged as anomalies.
* **F1 Score**:
  + Buses 115, 116, 117, and 118 exhibited relatively higher F1 scores, indicating a better balance between precision and recall. In contrast, buses 135 and 139 showed lower F1 scores, pointing to poorer anomaly detection performance.
* **Effect of Parameters**:
  + Lower values of k (e.g., k = 5) and higher alpha values (e.g., alpha = 0.05) led to excessive false positives, resulting in low precision.
  + Increasing k to 15 and reducing alpha to 0.01 helped in reducing false positives, but precision remained lower than desired in many cases.
  + Varying the CUSUM threshold h showed that higher values (e.g., h = 30 or h = 50) led to fewer false positives, but did not significantly improve the overall detection performance across all buses.

**4. Analysis and Recommendations**

Based on the results, several key insights and recommendations are provided:

1. **Precision-Recall Trade-off**:
   * The high recall and low precision indicate that while the model is sensitive to detecting anomalies, it lacks specificity. This trade-off suggests that further tuning of parameters is necessary to reduce false positives while maintaining adequate detection of true anomalies.
2. **Parameter Optimization**:
   * Increasing k to values such as 15 or higher may help in smoothing the decision boundary, thereby reducing false positives.
   * Lowering alpha to very strict thresholds (e.g., 0.001) can help in better distinguishing between normal and anomalous behavior, which could lead to higher precision.
3. **CUSUM Threshold Tuning**:
   * The CUSUM threshold h plays a crucial role in controlling the sensitivity of the anomaly detection. Higher values of h should be further explored to reduce the occurrence of false positives, especially for buses where detection performance is suboptimal.
4. **Bus-Specific Optimization**:
   * The varied performance across different buses suggests that a one-size-fits-all approach may not be ideal. It is recommended to optimize the parameters k, alpha, and h individually for each bus to achieve better overall detection performance.

**5. Conclusion**

The experiments conducted demonstrate the capability of k-NN-based anomaly detection in identifying changes in bus data within a smart grid. However, the results also highlight the need for careful parameter tuning to balance precision and recall, particularly in reducing false positives. Future work should focus on bus-specific parameter optimization and further experimentation with higher values of k and stricter alpha thresholds to enhance detection accuracy.

This report serves as a foundation for ongoing work in refining anomaly detection methods for smart grid data, aiming to achieve robust and reliable performance across all monitored buses.