Predicting Kubernetes Issues: Approach, Key Metrics, and Model Performance

- Team Wiz

1. Introduction

Kubernetes clusters face various failures, including pod crashes, resource limitations, and network problems. This project seeks to develop an AI/ML model that can forecast these failures in advance by analyzing both historical data and real-time metrics from the cluster.

2. Approach

2.1 Data Collection & Exploration

- **Dataset:** `balanced_shuffled_traffic.csv`
- **Columns:** Network traffic parameters (e.g., packet size, flow rate), resource usage (CPU, memory), and failure indicators.
- **Exploration:** Identified numerical features and the target variable (`Label`) indicating failures.
- **Key Findings:**
    - The dataset has **8,621 records** and **60 features**.
    - Failure labels are **balanced** (52% non-failure, 48% failure).
    - Strong correlations found with **packet size, network flow rates, and resource consumption**.

2.2 Data Preprocessing

- **Missing Values:** No missing values were found.
- **Outlier Handling:** Boxplots were used to detect outliers in key numerical features.
- **Feature Selection:**
    - Chose highly correlated features for training:
        - `Bwd Packet Length Std`
        - `Bwd Packet Length Max`
        - `Flow Bytes/s`
        - `Flow Packets/s`
        - `Total Length of Bwd Packet`

2.3 Model Training

- **Algorithm Used:** Random Forest Classifier
- **Training Steps:**
    - Splitting data into **training (80%)** and **testing (20%)** sets.
    - Standardizing numerical features.
    - Training the **Random Forest model**.

3. Key Metrics

The model's performance was evaluated using:

- **Accuracy**: Measures overall correctness.
- **Precision**: Assesses false positives in failure predictions.
- **Recall (Sensitivity)**: Ensures detection of actual failures.
- **F1-Score**: Balances precision and recall for reliability.

4. Issues Encountered & Resolutions

- **Error:** `NameError: name 'loaded_model' is not defined`
  - **Fix:** Assigned trained model (`rf_classifier`) to `loaded_model` before making predictions.
- **Correlation Calculation Error:** Strings in the dataset caused a failure in `df.corr()`.
  - **Fix:** Computed correlation only for **numerical** columns.

5. Future Enhancements

- Incorporate **real-time data streaming** from Kubernetes clusters.
- Expand scope to include **log-based anomaly detection**.