Problem Statement

The project aims to develop a robust machine learning model for anomaly detection in a multiphase flow facility. Anomaly detection involves identifying deviations from normal operating conditions within the facility, specifically focusing on differentiating between air leakage and air blockage anomalies. The primary challenge lies in accurately classifying these anomalies amidst various operating conditions, which are influenced by factors such as water and air flow rates, temperatures, pressures, and density. The goal is to create a system that can effectively identify and label anomaly points/clusters, ultimately enhancing the facility's operational efficiency and safety.

Methodology

The methodology involves leveraging machine learning techniques to address the anomaly detection problem. We will adopt a supervised learning approach, utilizing a combination of feature Selection and model training to classify normal and anomaly operational conditions. The process will consist of:

Data Preprocess the data to handle missing values, outliers, and ensure data quality.

Feature Selection: Extract relevant features from the collected data, categorizing them into control variables (directly influencing operating conditions) and external variables (indirectly influencing operating conditions) using PCA as the data are high dimensional or Heatmap and Grid Search

Model Development: Utilize machine learning algorithms such as Isolation Forest and Support Vector Machine (SVM) as the data are high dimensional with complex interactions to train the model on data from normal operating conditions. The model will learn to classify anomalies based on the extracted features and labeled data. Or GMM but the domain knowledge isn’t known.

Evaluation: Assess the performance of the trained model using metrics such as accuracy, precision, recall, and F1-score. Evaluate the model's ability to distinguish between air leakage and air blockage anomalies, particularly in unstable flow regimes.

Algorithm Selection and Justification

For this project, we have chosen the Isolation Forest algorithm and Support Vector Machine (SVM) classifier due to their suitability for anomaly detection tasks.
Isolation Forest: This algorithm is well-suited for anomaly detection in high-dimensional datasets with mixed variable types. It constructs isolation trees to isolate anomalies in the feature space, making it efficient for detecting outliers without requiring labeled data for training.
Support Vector Machine (SVM): SVMs are effective in separating data points into different classes based on their features. They are capable of handling high-dimensional data and nonlinear decision boundaries, making them suitable for classifying anomalies in complex datasets.
GMMs are a powerful probabilistic approach for modeling data distributions. They essentially represent the data as a mixture of multiple Gaussian distributions (bell curves).
Here's how GMMs work:
1. Training: Train the GMM on a dataset assumed to represent normal operating conditions. The model estimates the parameters (means and variances) of each Gaussian component within the mixture, capturing the different "clusters" within your data.
2. Anomaly Scoring: Once trained, calculate an anomaly score for each data point. This score reflects the probability of the data point belonging to the distribution learned by the GMM.
3. Anomaly Detection: Points with significantly low scores deviate from the expected patterns and are considered potential anomalies. These anomalies might warrant further investigation.

Timeline:

- Data preprocessing and feature selection: 1 week
- Model development , training and evaluation: 1 week
- Final reporting and documentation: 1 week