A MACHINE LEARNING APPROACH FOR PREDICTING MOLE FRACTION OF DISTTILATE USING MULTILINEAR, POLYNOMIAL REGRESSION AND PCA ANALYSIS

Pentapalli Venkata Satyanarayana

234107207

Submission Date: April 16, 2024

Final Project submission

Course Name: Applications of Al and ML in chemical engineering

Course Code: CL653

Executive Summary

The project focuses on creating a machine learning model to forecast the mole fraction of distillate in distillation processes, encompassing various stages from data collection to model deployment. Three distinct methodologies—linear regression, polynomial regression, and Principal Component Analysis (PCA)—are explored for modelling. The primary objective is to accurately predict the mole fraction of distillate, crucial for enhancing distillation operations' optimization and control. Data pertaining to temperatures, pressures, compositions, and flow rates are gathered from the distillation process to construct the dataset, which undergoes meticulous cleaning to handle missing values, outliers, and inconsistencies. Feature engineering techniques may be employed to extract pertinent information, followed by the segregation of features and target variables. Model selection involves opting for a simple linear model initially to establish a baseline prediction, followed by exploring non-linear relationships using polynomial regression to capture intricate data patterns. Additionally, Principal Component Analysis (PCA) is leveraged to reduce dataset dimensionality while retaining variance, potentially enhancing model performance. The dataset is then split into training and testing subsets for both linear and polynomial regression models, facilitating evaluation through metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), or R-squared (R^2). PCA-transformed data is also utilized for training and evaluation to gauge the impact of dimensionality reduction on predictive accuracy. Hyperparameters of chosen models undergo fine-tuning to optimize performance, which may involve adjusting regularization parameters for linear and polynomial regression or selecting the optimal number of principal components for PCA. Upon developing satisfactory models, they are deployed in a production environment for real-time predictions on new data, with integration into existing systems and continuous performance monitoring being imperative for successful deployment. Throughout the project, comprehensive documentation and transparent communication of findings are maintained to ensure reproducibility and knowledge transfer. Collaboration with domain experts is encouraged to interpret model results effectively and further refine predictive accuracy.

Introduction

Understanding and optimizing distillation processes are essential for improving efficiency, reducing costs, and ensuring product quality. One key aspect of this optimization is accurately modelling and fitting distillation data to different types of curves.

In this project, we explore a machine learning approach for fitting distillation data to various types of curves, including multilinear, polynomial, and principal component analysis (PCA) curves. By leveraging machine learning algorithms.

we aim to develop models that can effectively capture the complex relationships present in distillation data and provide accurate predictions for process optimization.

The primary objectives of this project are as follows:

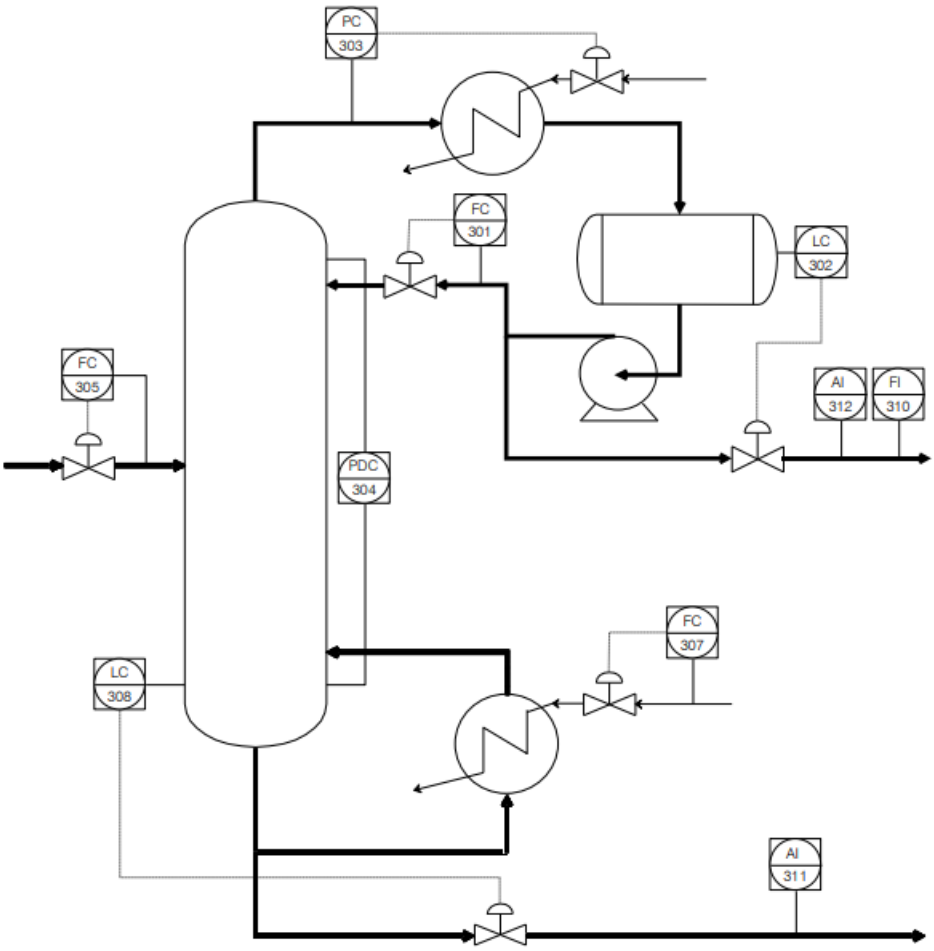
Data Collection and Preprocessing: We gather distillation data from relevant sources and preprocess it to ensure its quality and compatibility with machine learning algorithms. This involves cleaning the data, handling missing values, and possibly normalizing or standardizing the features.

Curve Fitting Algorithms: We explore and implement machine learning algorithms suitable for fitting distillation data to multilinear, polynomial, and PCA curves. This may involve techniques such as linear regression, polynomial regression, and PCA-based regression.

Model Training and Evaluation: We train our machine learning models using the pre-processed distillation data and evaluate their performance using appropriate metrics such as mean squared error (MSE), R-squared value, and cross-validation scores. This step helps us assess the accuracy and generalization capabilities of our models.

Comparison of Curve Fitting Techniques: We compare the performance of different curve fitting techniques, including multilinear regression, polynomial regression, and PCA-based regression. This comparison allows us to identify the strengths and weaknesses of each approach and determine the most suitable technique for different types of distillation data.

Overall, this project aims to showcase the effectiveness of a machine learning approach for fitting distillation data to various types of curves, thereby facilitating improved understanding and optimization of distillation processes in industrial settings.



This is the model of the process where the sensor data is collected and these are the data collection points in the distillation column. Distillation is method of separation of components from a liquid mixture which depends on the differences in boiling points of the individual components and the distributions of the components between a liquid and gas phase in the mixture.

Description of the Project

Distillation processes involve the separation of components in a mixture based on their differing volatilities, typically through vaporization and condensation. Understanding and optimizing these processes are critical. Traditional methods for modelling distillation data often rely on deterministic models or empirical correlations, which may struggle to capture the intricate relationships present in the data. Machine learning (ML) offers a promising alternative for distillation data analysis by leveraging algorithms that can learn from data patterns and relationships. This section provides a theoretical foundation for the proposed approach of using ML for distillation data fusion, focusing on multilinear regression, polynomial regression, and principal component analysis (PCA) curve fitting techniques.

Multilinear Regression: Multilinear regression is a statistical technique used to model the relationship between multiple independent variables and a dependent variable. In the context of distillation data fusion, multilinear regression can be applied to fit data where multiple factors influence the distillation process simultaneously. It assumes a linear relationship between the independent variables and the dependent variable, enabling the estimation of coefficients to represent this relationship.

Polynomial Regression: Polynomial regression extends the concept of linear regression by allowing the relationship between variables to be modelled as an nth-degree polynomial. This flexibility enables the capture of nonlinear relationships present in distillation data, which may not be adequately represented by linear models. By fitting polynomial curves to the data, polynomial regression can provide a more accurate representation of the underlying trends and patterns.

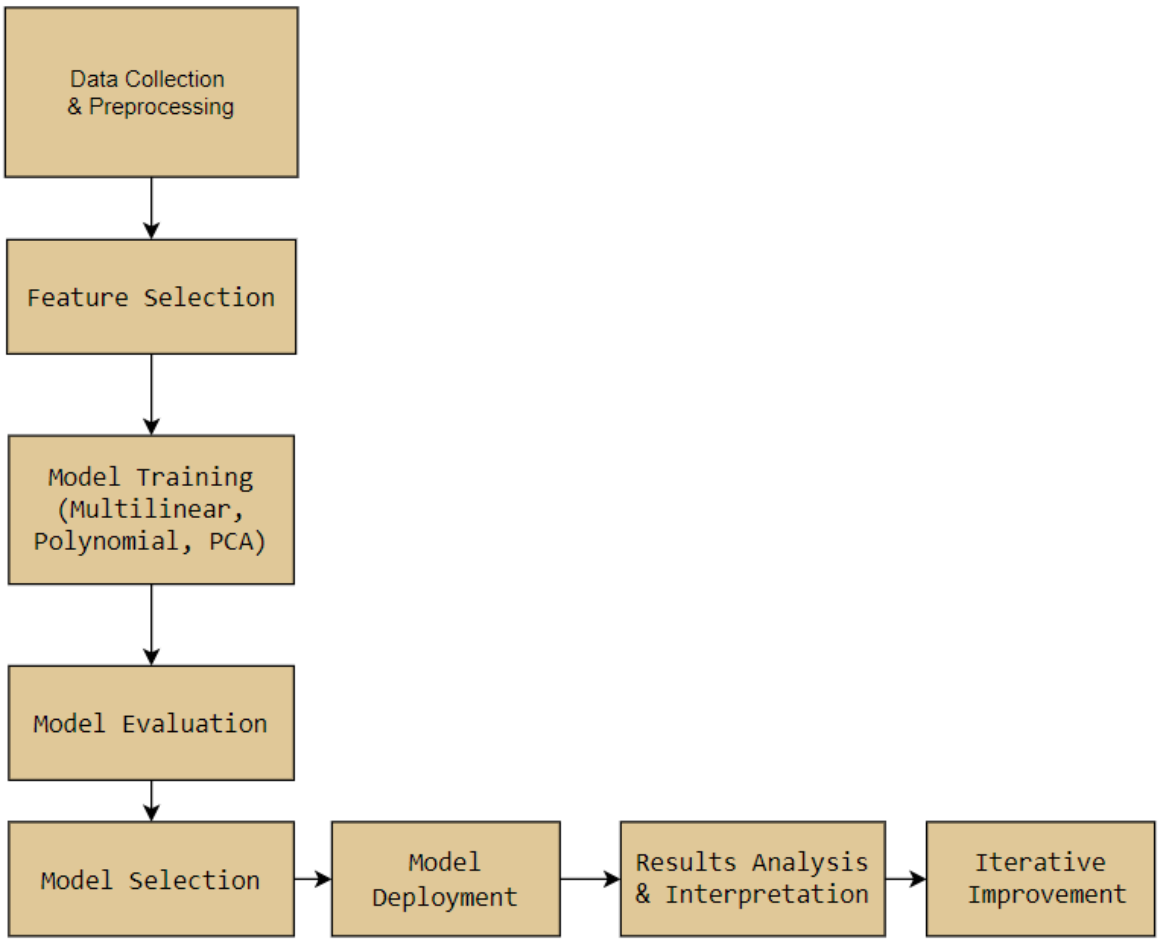
Principal Component Analysis (PCA): PCA is a dimensionality reduction technique commonly used to identify patterns in high-dimensional data by transforming it into a lower-dimensional space while preserving the most important information. In the context of distillation data fusion, PCA can be applied to extract the principal components representing the dominant patterns or variations in the data. By fitting curves to these principal components, PCA-based regression can capture the essential features of the distillation data while reducing complexity.

By combining these techniques within a machine learning framework, we aim to develop a comprehensive approach for distillation data fusion. This approach seeks to address the limitations of traditional methods by leveraging the flexibility and predictive power of machine learning algorithms. Through empirical evaluation and analysis, we endeavour to demonstrate the effectiveness of multilinear, polynomial, and PCA curve fitting techniques in accurately representing distillation data and facilitating improved process understanding and optimization.

Problem Statement

Accurately modelling and analysing distillation data remains a challenge. Traditional methods often struggle to capture the complex relationships inherent in such data. This project aims to address this challenge by proposing a machine learning approach for distillation data fusion. Specifically, investigate the effectiveness of multilinear, polynomial, and PCA curve fitting techniques in accurately representing distillation data.

Block Diagram/Flowchart of Process & Model Implementation



This flowchart shows how we take data from distillation processes, teach a computer to understand it using machine learning techniques, and then use that understanding to improve the distillation process. Each step helps us get closer to accurately representing the data and making useful predictions.

Data Collection & Preprocessing:

- First, we collect information about how distillation works.

- Then, we tidy up this info to make sure it's correct and well-organized.

Feature Selection & Engineering:

- Next, we pick out the most important parts of the info we gathered.

- Sometimes, we make new info from what we already have to understand things better.

Model Training:

- Now, we teach our computer about distillation by showing it examples.

- We use different methods to help the computer learn, like showing it lots of examples and using different math tricks.

Model Evaluation:

- After teaching the computer, we check if it learned correctly.

- We use special tests to see if what the computer predicts matches what really happens.

Model Selection:

- Based on how well the computer learned, we pick the best way for it to understand distillation.

Model Deployment:

- Once we know the best way, we put it into action.

- The computer starts looking at real distillation data to help with different jobs.

Result Analysis & Interpretation:

- We study what the computer tells us about distillation.

- This helps us understand how distillation works and find ways to make it better.

Iterative Improvement:

- Finally, we keep making our computer smarter and our methods better.

- This way, we keep getting better at understanding and improving distillation processes.

Data Source

Link :- https://www.kaggle.com/datasets/amarhaiqal/aspen-hysys-distillation-column-data

This dataset simulates readings from distillation column which used to distillate HX and TX component.

- Sensor1: Liquid Percentage in Condenser
- Sensor2: Condenser Pressure
- Sensor3: Liquid Percentage in Reboiler
- Sensor4 & Sensor5: Mass Flow Rate in Feed Flow and Top Outlet Stream
- Sensor6: Net Mass Flow in main tower
- Sensor7: Mole Fraction HX at reboiler
- Sensor8: HX Mole Fraction in Top Outler Stream
- Sensor9 & Sensor10: Feed Mole Fraction
- Sensor11: Feed Tray Temperature
- Sensor12: Main Tower Pressure
- Sensor13: Bottom Tower Pressure
- Sensor14: Top Tower Pressure
- Sensor15: Reflux Ratio
- Sensor16: Duties Summary
- Mole Fraction TX(Toluene)
- Mole Fraction HX(Hexane)

As we know the mole fraction the sum of mole fraction of hexane and toluene in overhead & bottom is 1. The aim of this project to develop a model equation from this data by using multilinear, polynomial regression and reduce the dimensionality to two feature and analyse the best fit for this data.

Methodology

Description of Data

The data for this project includes information collected from distillation processes. This data is almost same with respect to time but the varies with + or – 1% with previous value so this data can be considered as steady state.

Before conducting analysis or modelling, it's essential to preprocess the data to ensure its quality and compatibility with the chosen techniques.

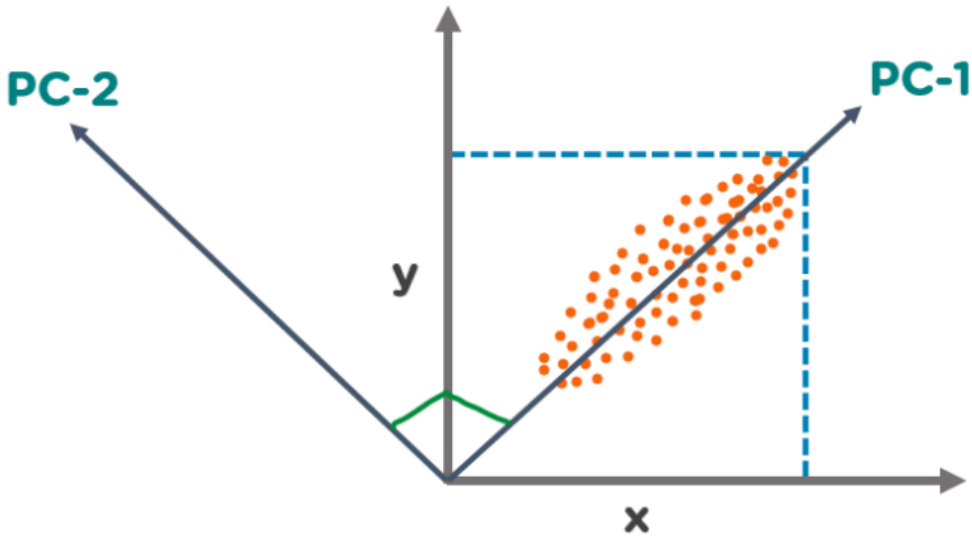
Dimensionality Reduction:

Apply techniques such as PCA (Principal Component Analysis) if dealing with high-dimensional data to reduce the number of variables while preserving important information.

Strategies for AI/ML Model Development

Developing AI/ML models for distillation data fusion involves some important steps. First, we need to really understand how distillation works and what things affect it. Then, we look at the data we have, using graphs to see what's happening. We might need to change the data a bit to make it easier for the computer to understand. Once we've got our data ready, we try out different computer programs to see which one works best. We also make these programs as accurate as possible by fine-tuning them. To make sure they're good, we test them with different sets of data. Sometimes, we even put a few different programs together to make them stronger. All along the way, we make sure these programs are easy to understand and explain. And once they're doing their job, we watch them to make sure they keep doing it well, and we make them better if needed. By doing all this, we can make really helpful tools that make distillation processes better.

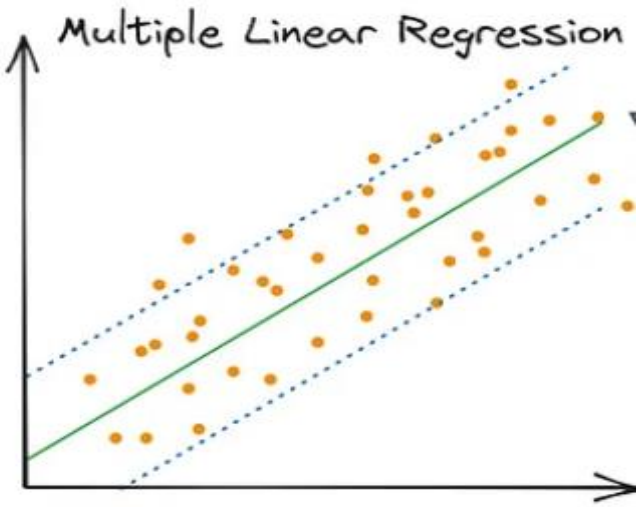
Principal component analysis



Principal component analysis, or PCA, is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize, and thus make analyzing data points much easier and faster for machine learning algorithms without extraneous variables to process.

Multiple linear regression

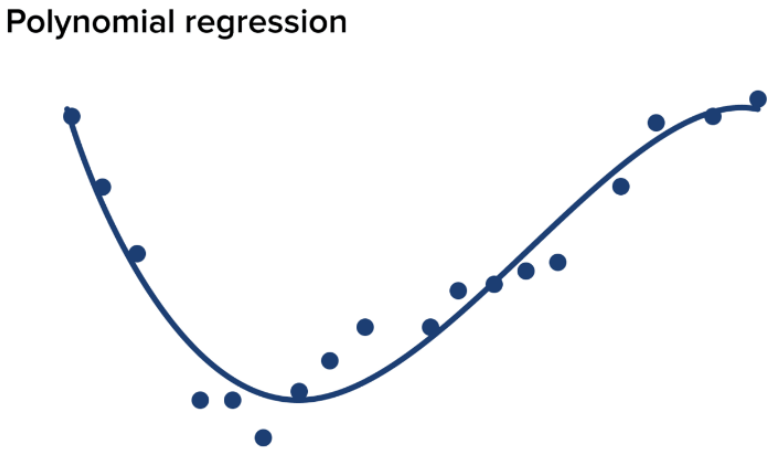


Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression is to model the linear relationship between the explanatory (independent) variables and response (dependent) variables. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable.

The mathematical representation of Multiple Linear Regression is given by the equation: -

$y = \beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta nxn + \varepsilon$, where y is the dependent variable, x1, x2, …, xn are the independent variables, β0, β1, β2, …, βn are the coefficients, and ε is the error term.

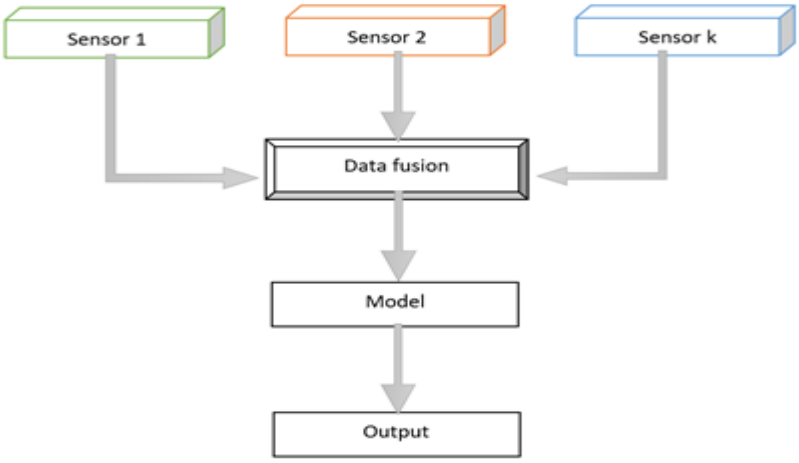
Polynomial regression



Polynomial regression is needed when there is no linear correlation fitting all the variables. So instead of looking like a line, it looks like a nonlinear function. Let's delve deeper into this type of regression. Polynomial regression is useful in many cases. Since a relationship between the independent and dependent variables isn't required to be linear, you get more freedom in the choice of datasets and situations you can be working with. So, this method can be applied when simple linear regression underfits the data. Polynomial regression is a simple yet powerful tool for predictive analytics. It allows you to consider non-linear relations between variables and reach conclusions that can be estimated with high accuracy.

Deployment Strategy

To deploy AI/ML models for distillation data fusion, like multilinear, polynomial, and PCA curve fitting, we need a good plan. First, we prepare our trained models and any steps needed to get data ready for them to use. We make sure these are in a format that's easy to move around, like Python files. Then, we fit these models into the systems already in place for distillation. This might mean putting them into systems that control the process or analyse the data. We also train the people who will be using the models, so they know how to get the most out of them. And we set up a way to get feedback, so we can keep making our models better over time. By doing all this, we make sure our AI/ML models are doing their job well and helping improve distillation processes.



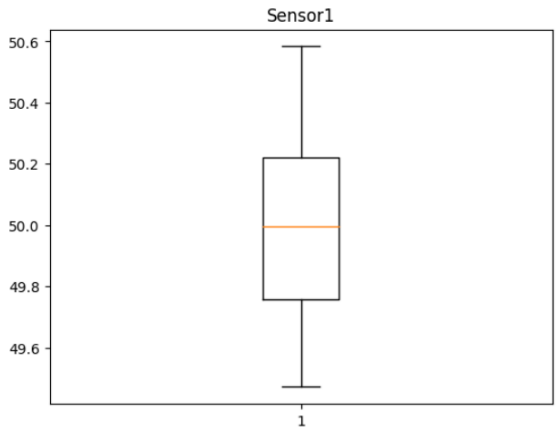
This method is referred to as input level fusion. Research proposes two possible approaches for early fusion technique. The first approach is combining data by removing the correlation between two sensors. The second approach is to fuse data at its lower dimensional common space. There are many statistical solutions which can be used to accomplish one or both methods, including principal component analysis (PCA),

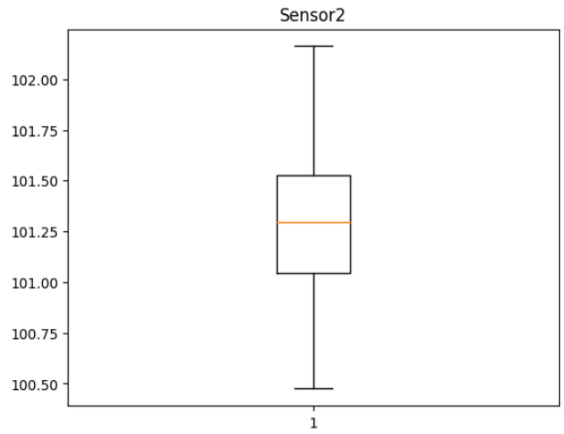
Results and Discussion

The pre-processing of data

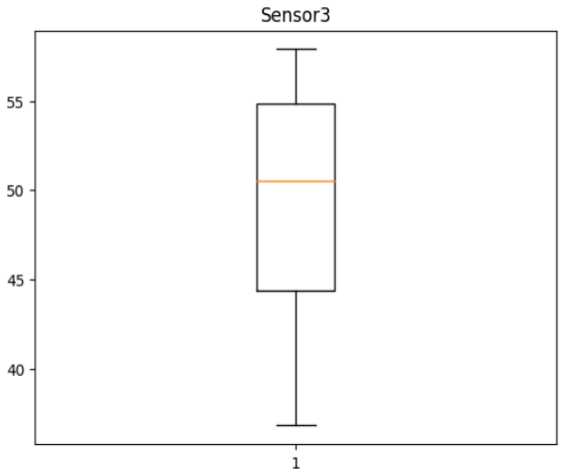
The sensor 16 data have been removed since this data contain mole fraction of the other component the combined mole fraction of the overhead product is one. If we predict the mole fraction of the one compound, we can calculate the mole fraction of the other component by just subtracting from one.

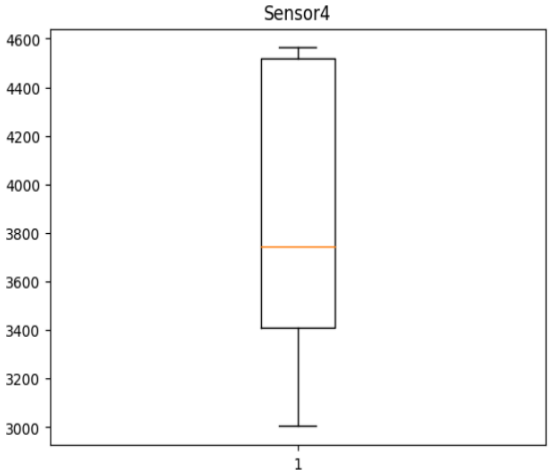
preprocess the data, remove the outlier and plot the individual outlier plot.



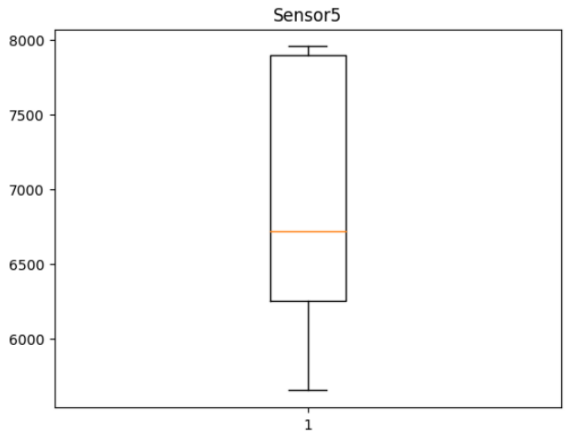


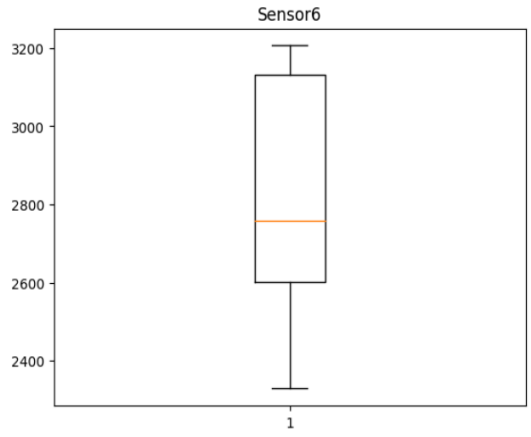
Sensor	Median	IQR Value	Lower	Higher
1	50.0	0.4	49.6	50.6
2	101.25	0.5	100.5	102.5



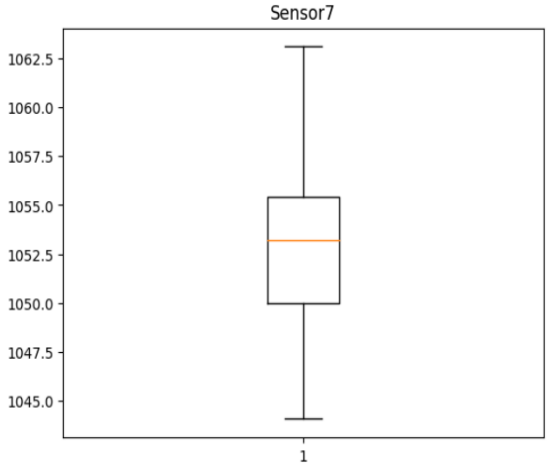


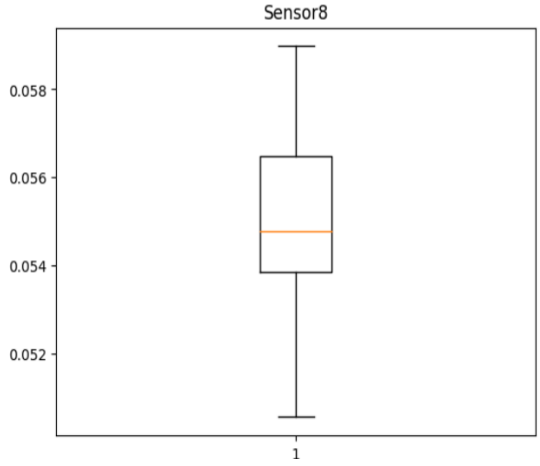
Sensor	Median	IQR Value	Lower	Higher
3	50.0	10	39	56
4	3750	900	3000	4600



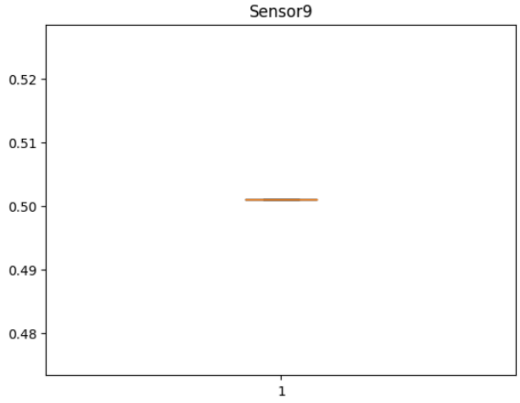


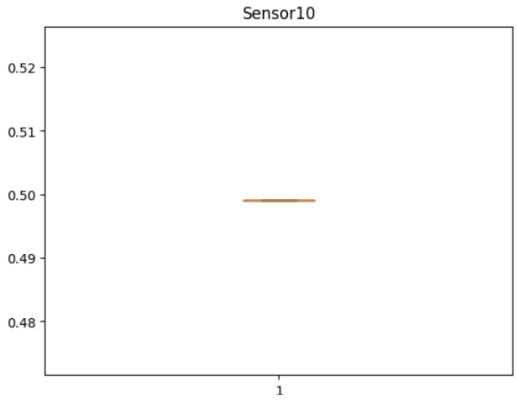
Sensor	Median	IQR Value	Lower	Higher
5	6750	1700	5850	8000
6	2750	500	3200	2400



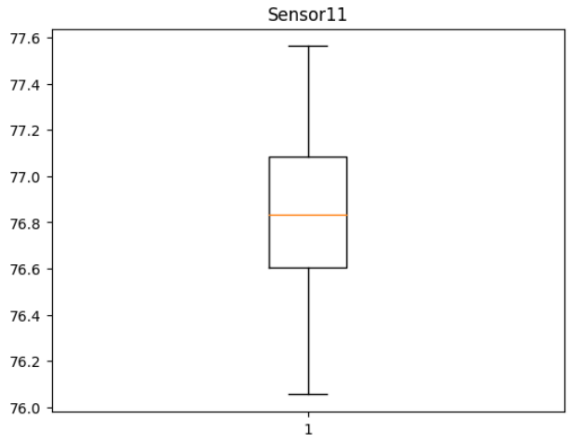


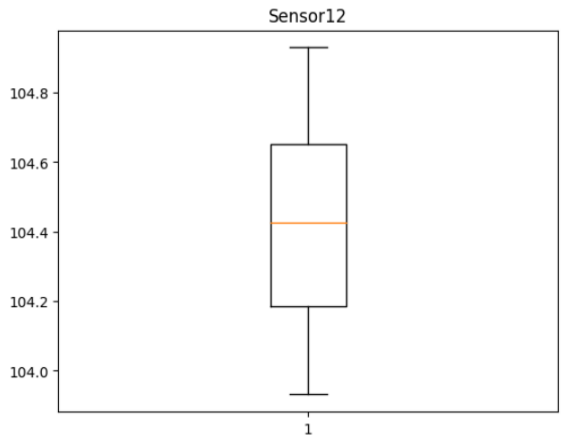
Sensor	Median	IQR Value	Lower	Higher
7	1053.5	5	1045.5	1062.5
8	0.055	0.002	0.052	0.059



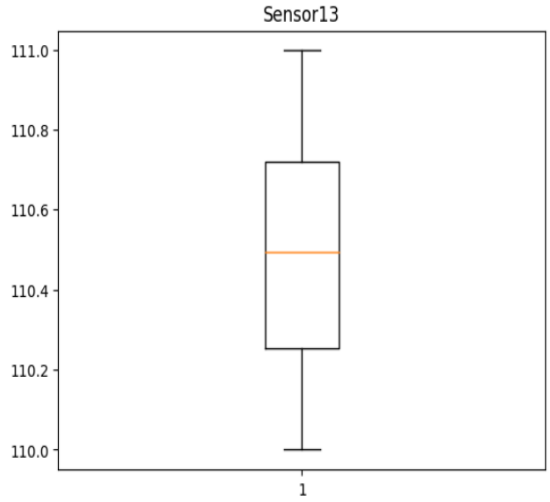


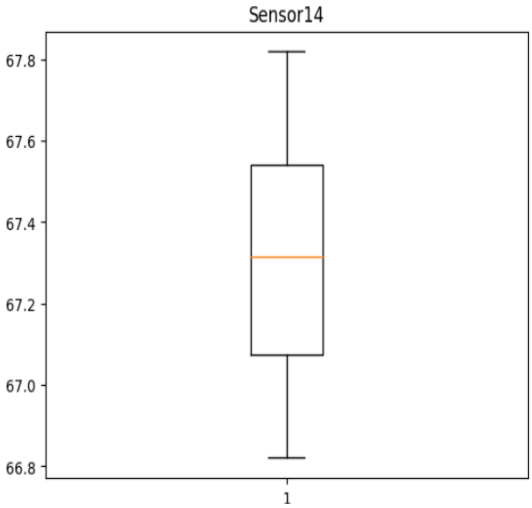
Sensor	Median	IQR Value	Lower	Higher
9	0.5	0.0001	0.5	0.50001
10	0.5	0.00001	0.5	0.50001





Sensor	Median	IQR Value	Lower	Higher
11	76.8	0.4	76	77.6
12	104.4	0.5	104	104.9

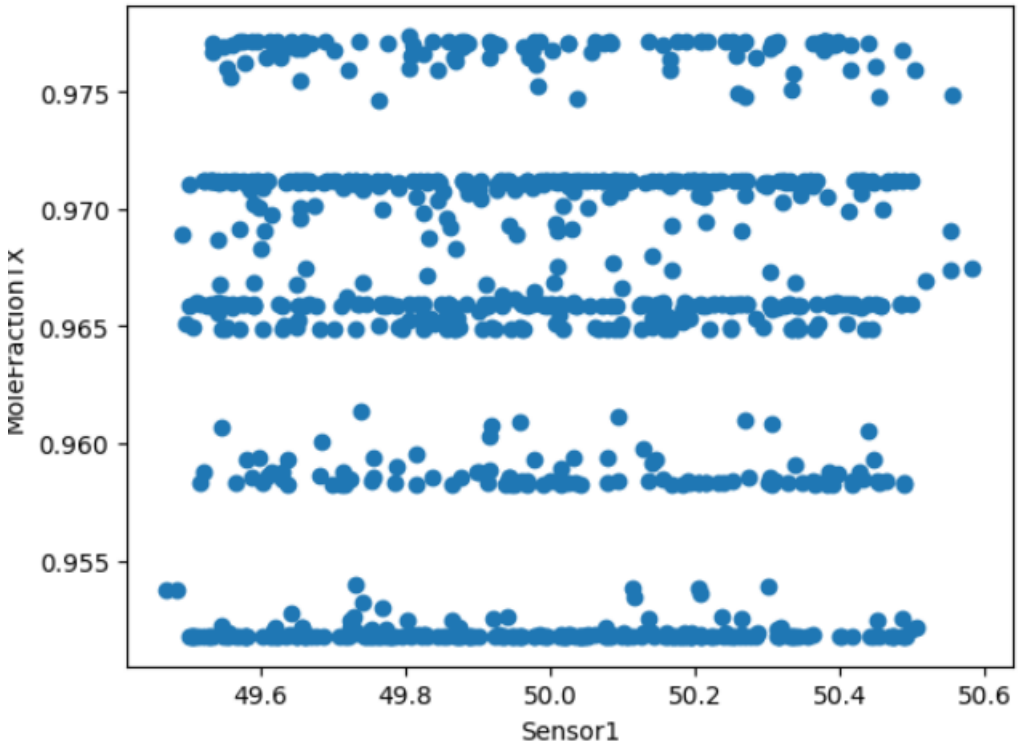




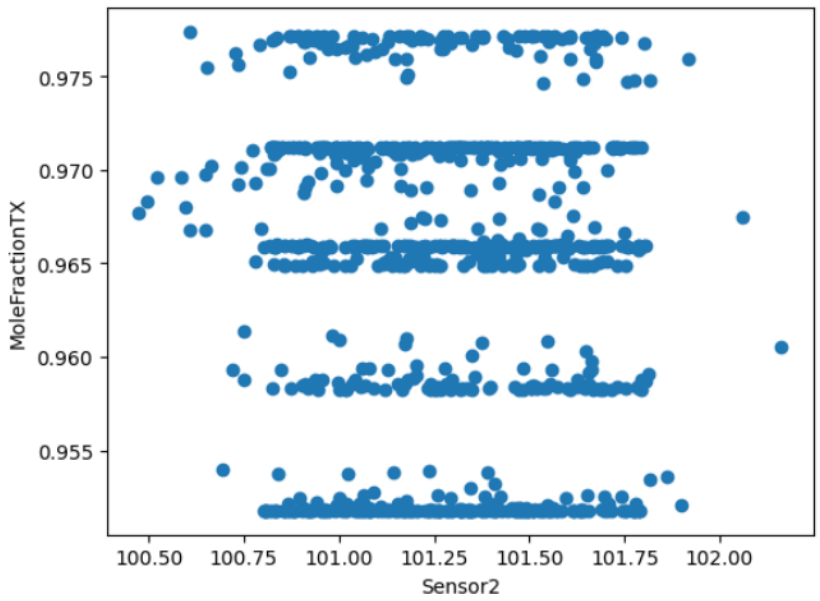
Sensor	Median	IQR Value	Lower	Higher
13	110.5	0.5	110.0	111
14	67.3	0.4	66.8	67.8

The values which in between lower and higher are considered as data points and remaining which don’t fall under lower and higher are taken as outliers.

Variation of Mole fraction of TX with sensor data

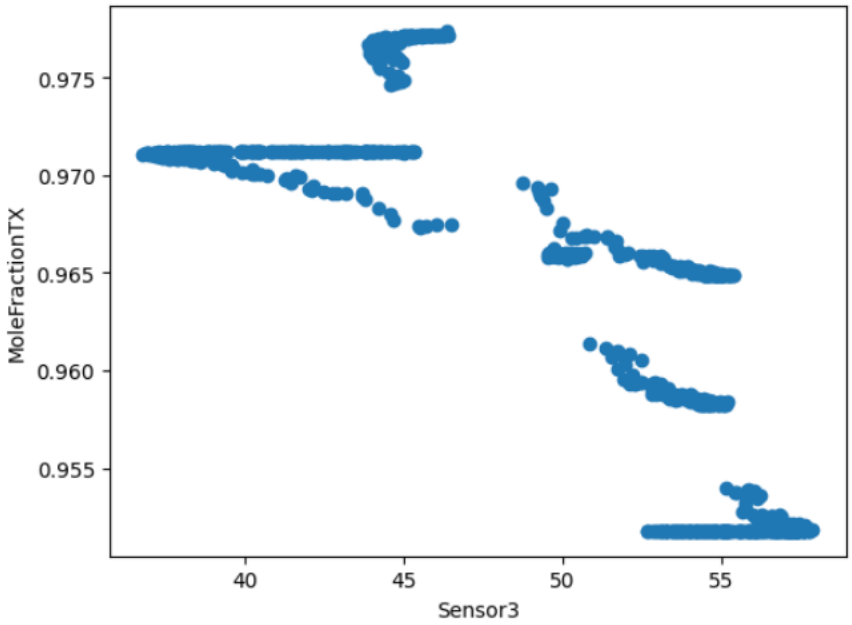


The plot shown in the above figure is mole fraction of TX (on y axis) versus sensor 1 data (on x axis) the conclusion point from this plot is the variation in mole fraction does not depend much on sensor data since the of is 0.975 when the data value 49.6 and same 0.975 when the data value is 50.6, we can conclude that sensor 1 data contribute in the value of mole fraction but doesn't show much variation when disturbed.

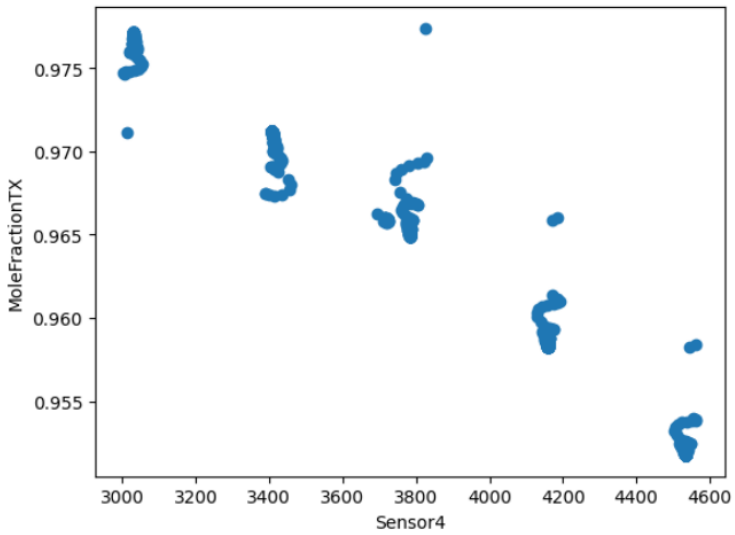


The plot shown in the above figure is mole fraction of TX (on y axis) versus sensor 2 data (on x axis) the conclusion point from this plot is the variation in mole fraction does not depend

much on sensor data since the of is 0.975 when the data value 100.5 and same 0.975 when the data value is 102.00, we can conclude that sensor 2 data contribute in the value of mole fraction but doesn't show much variation when disturbed.

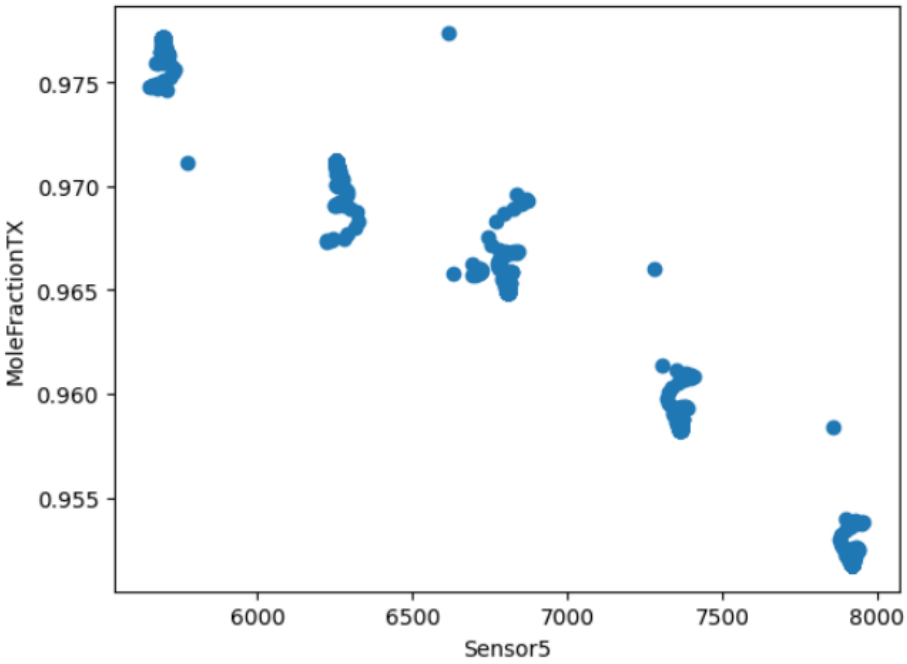


The plot shown in the above figure is mole fraction of TX (on y axis) versus sensor 3 data (on x axis) the conclusion point from this plot is the variation in mole fraction as the value of sensor data increased from 40 to 55 the value of mole fraction decreased from 0.975 to 0.955 so we have to tune the value of sensor 3 so that we can achieve the maximum value of mole fraction.

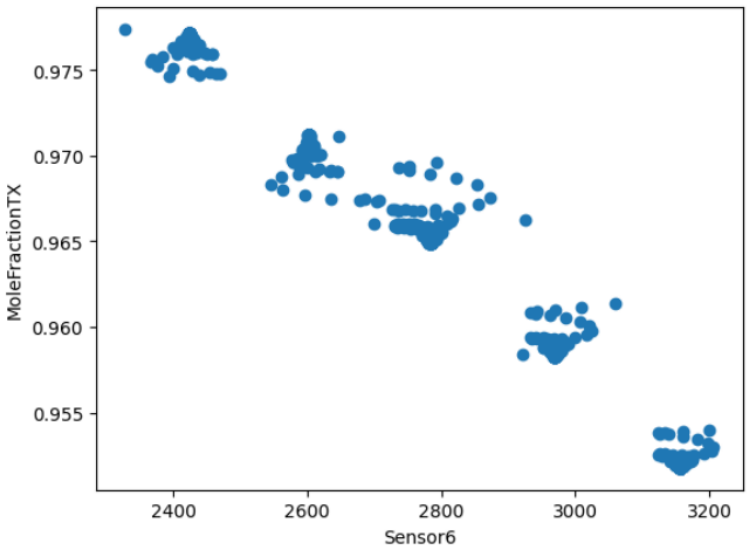


The plot shown in the above figure is mole fraction of TX (on y axis) versus sensor 4 data (on x axis) the conclusion point from this plot is the variation in mole fraction as the value of sensor data increased from 3000 to 4600 the value of mole fraction decreased from 0.975 to

0.955 so we have to tune the value of sensor 4 so that we can achieve the maximum value of mole fraction.

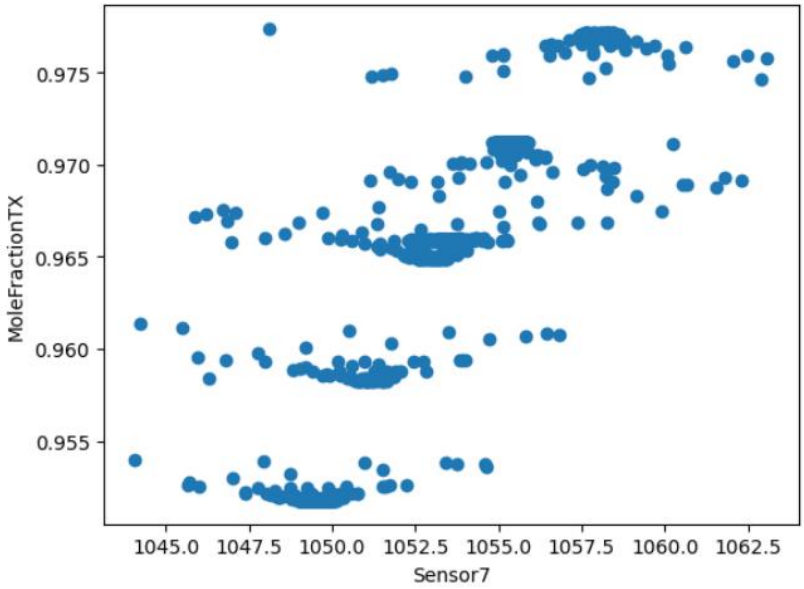


The plot shown in the above figure is mole fraction of TX (on y axis) versus sensor 5 data (on x axis) the conclusion point from this plot is the variation in mole fraction as the value of sensor data increased from 6000 to 8000 the value of mole fraction decreased from 0.975 to 0.955 so we have to tune the value of sensor 5 so that we can achieve the maximum value of mole fraction.

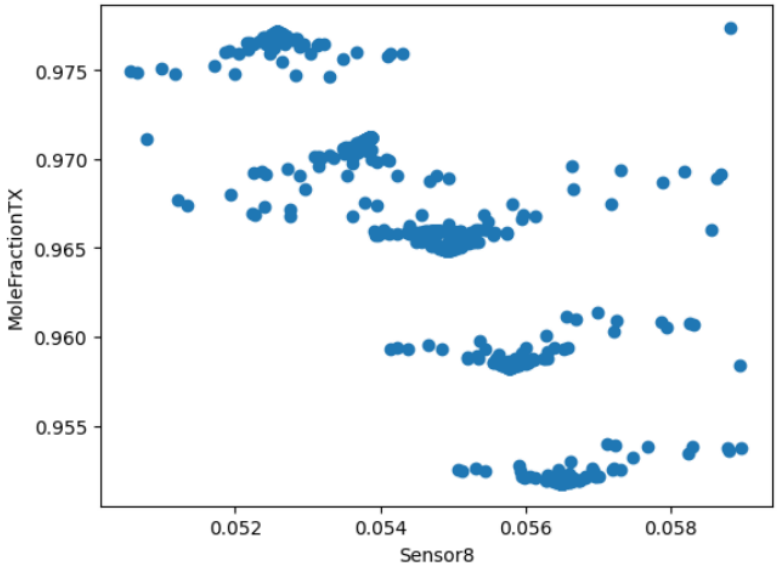


The plot shown in the above figure is mole fraction of TX (on y axis) versus sensor 6 data (on x axis) the conclusion point from this plot is the variation in mole fraction as the value of sensor data increased from 2400 to 3200 the value of mole fraction decreased from 0.975 to

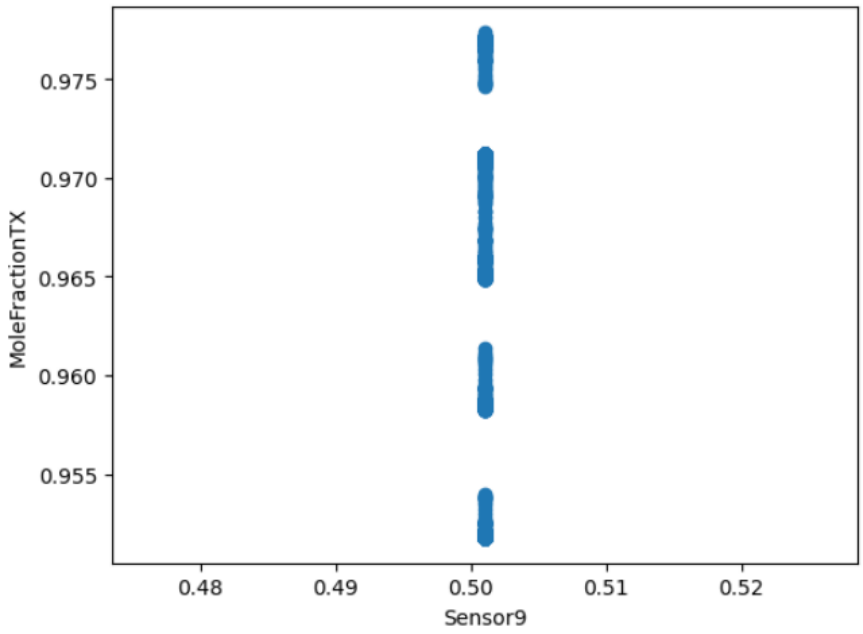
0.955 so we have to tune the value of sensor 6 so that we can achieve the maximum value of mole fraction.



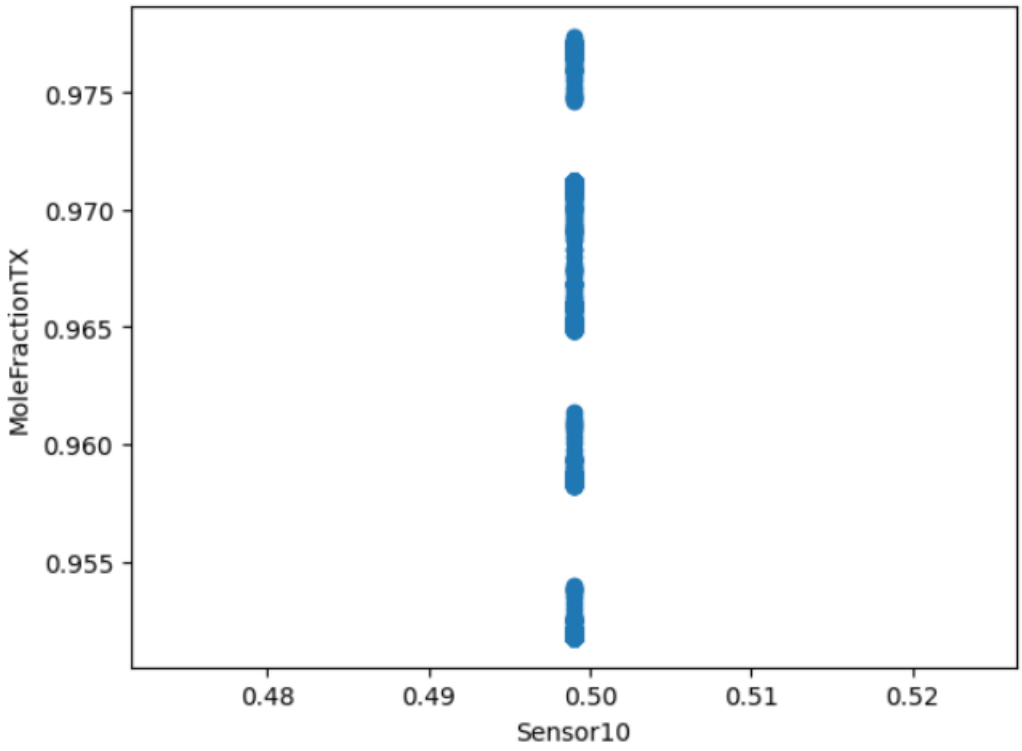
The plot shown in the above figure is mole fraction of TX (on y axis) versus sensor 7 data (on x axis) the conclusion point from this plot is the variation in mole fraction as the value of sensor data increased from 1045 to 1062.5 the value of mole fraction increased from 0.955 to 0.975 so we have to tune the value of sensor 7 so that we can achieve the maximum value of mole fraction.



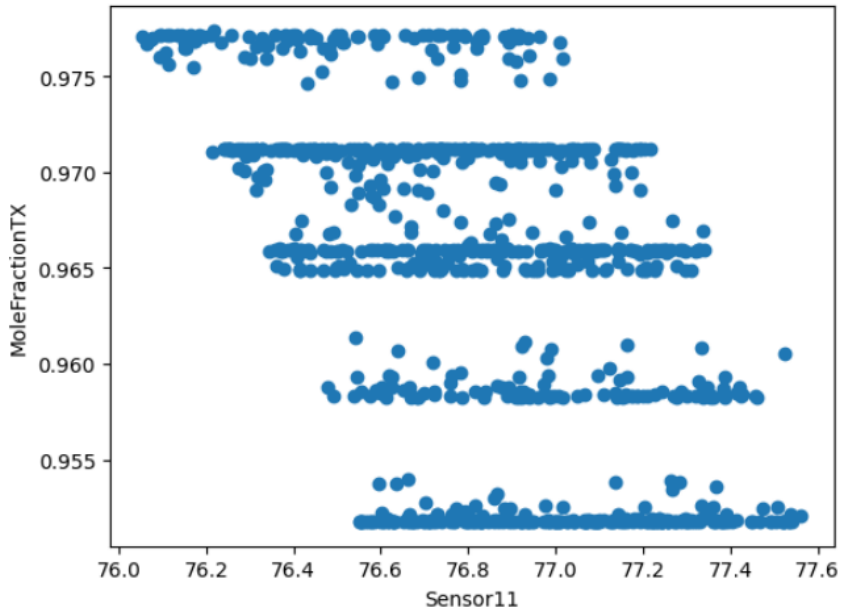
The plot shown in the above figure is mole fraction of TX (on y axis) versus sensor 8 data (on x axis) the conclusion point from this plot is the variation in mole fraction as the value of sensor data increased from 0.052 to 0.058 the value of mole fraction decreased from 0.975 to 0.955 so we have to tune the value of sensor 8 so that we can achieve the maximum value of mole fraction.



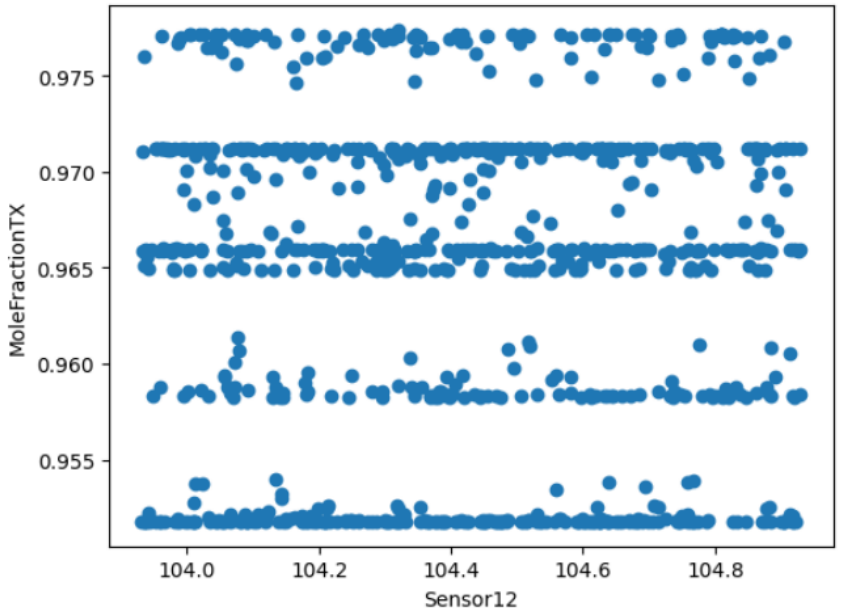
The plot shown in the above figure is mole fraction of TX (on y axis) versus sensor 9 data (on x axis) the conclusion point from this plot is the variation in the mole fraction doesn't depend on the sensor 9.



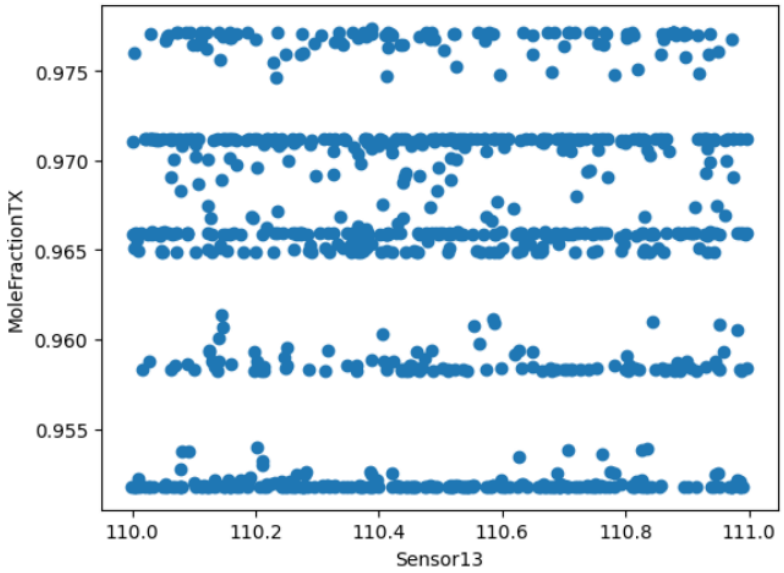
The plot shown in the above figure is mole fraction of TX (on y axis) versus sensor 10 data (on x axis) the conclusion point from this plot is the variation in the mole fraction doesn't depend on the sensor 10.



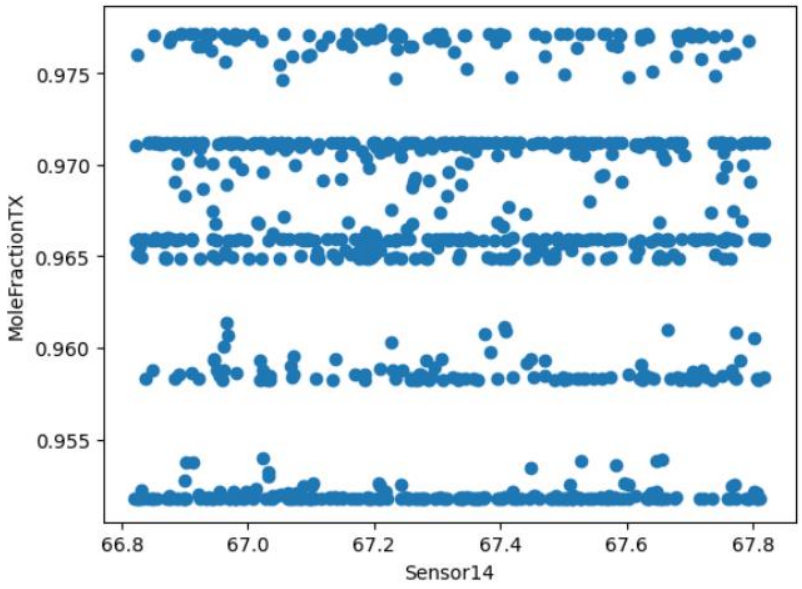
The plot shown in the above figure is mole fraction of TX (on y axis) versus sensor 11 data (on x axis) the conclusion point from this plot is the variation in mole fraction as the value of sensor data increased from 76 to 77.6 the value of mole fraction decreased from 0.975 to 0.955 so we have to tune the value of sensor 11 so that we can achieve the maximum value of mole fraction.



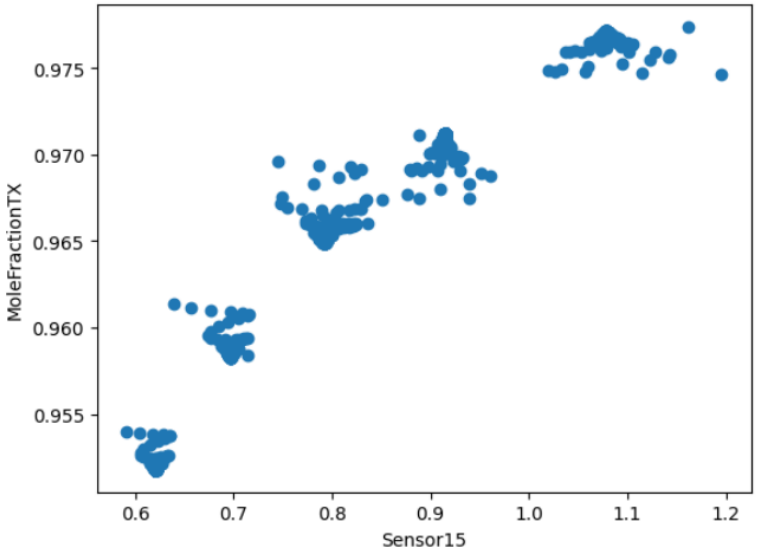
The plot shown in the above figure is mole fraction of TX (on y axis) versus sensor 12 data (on x axis) the conclusion point from this plot is the variation in mole fraction does not depend much on sensor data since the of is 0.975 when the data value 104 and same 0.975 when the data value is 104.8, we can conclude that sensor 12 data contribute in the value of mole fraction but doesn't show much variation when disturbed.



The plot shown in the above figure is mole fraction of TX (on y axis) versus sensor 13 data (on x axis) the conclusion point from this plot is the variation in mole fraction does not depend much on sensor data since the of is 0.975 when the data value 110 and same 0.975 when the data value is 111.0, we can conclude that sensor 13 data contribute in the value of mole fraction but doesn't show much variation when disturbed.



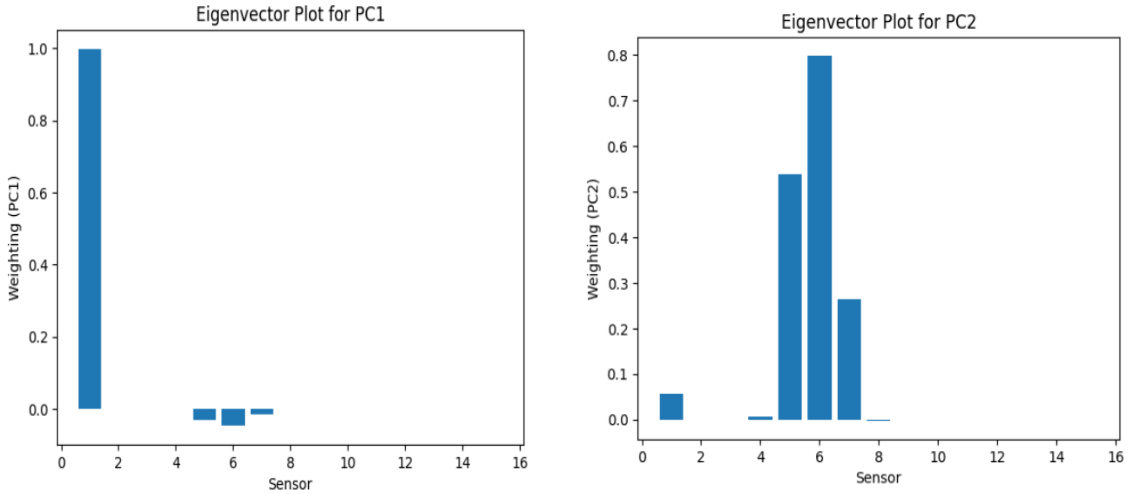
The plot shown in the above figure is mole fraction of TX (on y axis) versus sensor 14 data (on x axis) the conclusion point from this plot is the variation in mole fraction does not depend much on sensor data since the of is 0.975 when the data value 66.8 and same 0.975 when the data value is 67.8, we can conclude that sensor 14 data contribute in the value of mole fraction but doesn't show much variation when disturbed.



The plot shown in the above figure is mole fraction of TX (on y axis) versus sensor 15 data (on x axis) the conclusion point from this plot is the variation in mole fraction as the value of sensor data increased from 0.6 to 1.2 the value of mole fraction increased from 0.955 to 0.975 so we have to tune the value of sensor 15 so that we can achieve the maximum value of mole fraction.

Results obtained from PCA Analysis

The PCA analysis is performed for this sensor data (number of principle components =2)



Conclusion for PC1

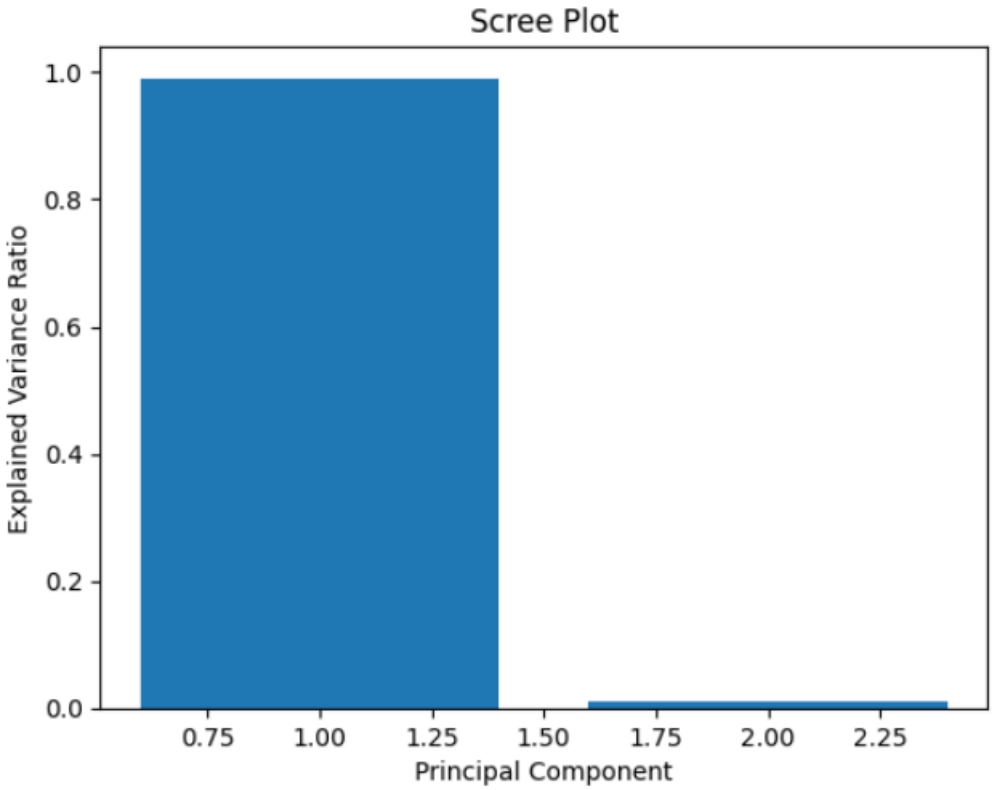
the PCA model, the weightings provided for each sensor in PC1 illustrate the degree to which each sensor contributes to the construction of the first principal component. PC1 captures the direction of maximum variance in the dataset, and the weightings indicate the strength and

direction of the relationship between each sensor and PC1. For instance, Sensor 1 has a weighting of 1.00, signifying a strong positive correlation with PC1. This indicates that variations in Sensor 1 are highly aligned with the primary pattern captured by PC1. Conversely, sensors with weightings close to 0, such as Sensor 2 through Sensor 15, have minimal influence on PC1. Negative weightings, such as those for Sensor 5, Sensor 6, and Sensor 7, suggest an inverse relationship with PC1. In practical terms, these weightings offer valuable insights into which sensors are most significant in explaining the variability observed in the data. For instance, Sensor 1's high weighting suggests that it may be a critical contributor to the underlying patterns or trends in the dataset, while sensors with negligible weightings may have less relevance in explaining the variability captured by PC1. Understanding these weightings is crucial for interpreting the PCA results effectively and identifying key features or variables that drive the observed patterns in the data. This information can inform decision-making processes, such as feature selection, anomaly detection, or understanding underlying relationships between variables, ultimately enhancing the quality of analysis and decision support.

Conclusion for PC2

In PCA, the weightings for each sensor in PC2 show how much they contribute to making the second main pattern in the data. PC2 shows the biggest differences left after we've already explained the main differences with PC1. For instance, if Sensor 5, Sensor 6, and Sensor 7 have high weightings like 0.54, 0.80, and 0.26, it means they're closely related to PC2. If their values go up, PC2 goes up too. But sensors with weightings close to 0, like Sensor 2 or Sensor 3, don't affect PC2 much. Understanding these weightings helps us see which sensors are important for different patterns in the data. This is useful for making decisions, like choosing which sensors are most important, finding unusual data points, or understanding how sensors relate to each other. Sharing this information helps others understand the data better and make better decisions. In a thousand words, the weightings assigned to each sensor in the second principal component (PC2) within the context of Principal Component Analysis (PCA) serve as crucial indicators of their respective contributions to the overall structure and variability observed in the dataset. PCA is a powerful statistical technique used to transform high-dimensional data into a lower-dimensional representation, capturing the most important patterns and relationships among variables.PC2 represents the direction of the maximum remaining variance after accounting for the primary pattern captured by PC1. It offers orthogonal information to PC1, meaning that it captures a different aspect of the data's variability. The weightings associated with each sensor in PC2 indicate both the strength and direction of their relationship with this secondary pattern. For instance, if a sensor has a high positive weighting in PC2, such as Sensor 5, Sensor 6, and Sensor 7 with weightings of 0.54, 0.80, and 0.26 respectively, it suggests a strong positive correlation between the variations in these sensors and the pattern represented by PC2. In simpler terms, when the values of these sensors increase, PC2 tends to increase as well, indicating a consistent relationship between these sensors and the underlying pattern captured by PC2.Conversely, sensors with weightings close to 0 in PC2, such as Sensor 2, Sensor 3, Sensor 8, and others, have minimal influence on this component. This implies that variations in these sensors are not strongly aligned with the secondary pattern represented by PC2.Understanding these weightings is essential for interpreting the PCA results effectively. It helps highlight the significance of each sensor in capturing the variability observed in the data. By identifying which sensors contribute most strongly to specific patterns, PCA facilitates decision-making processes such as feature selection, anomaly detection, or understanding the underlying relationships between

variables. For example, knowing that sensors like Sensor 5, Sensor 6, and Sensor 7 are closely related to PC2 can inform decisions about prioritizing these sensors for further analysis or monitoring. Conversely, sensors with negligible weightings in PC2 may be considered less relevant for understanding the secondary patterns in the data.

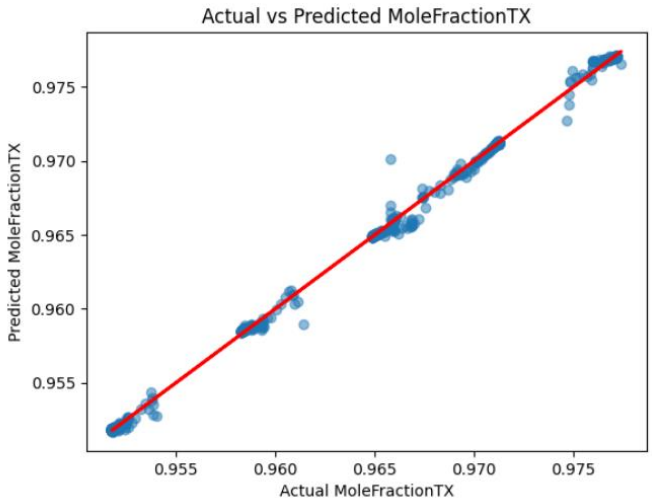


The scree plot shows the PC1 accounted as 93 percent and PC2 accounted 7%.

The model equation obtained from the PCA analysis is Model equation:

y = 0.96 + 0.00PC1 + -0.00PC2

plot for actual value versus predicted value



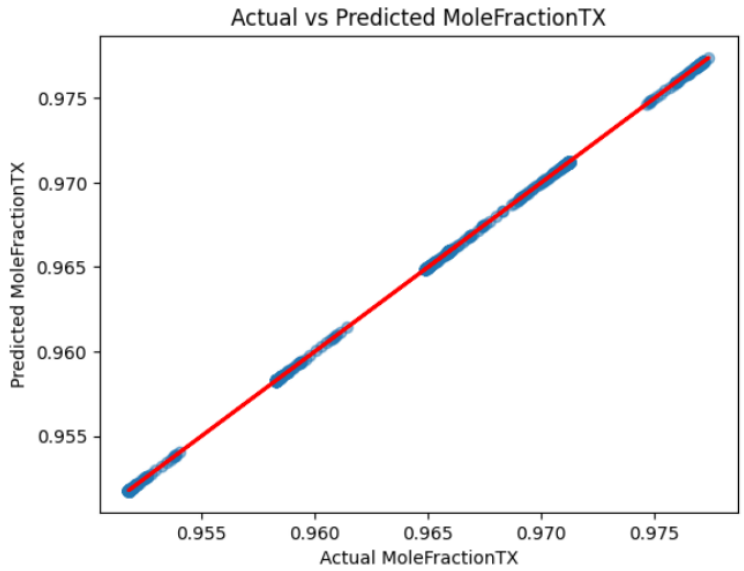
The provided scatter plot is between actual values against predicted values, with the x-axis representing the actual values and the y-axis denoting the predicted values. The plot visually showcases the alignment between the model's predictions and the true values, evidenced by a substantial concentration of data points clustering around the trend line. This clustering indicates that the model's predictions closely track the actual trends observed in the dataset. Furthermore, the relatively low Mean Absolute Error (MAE) of **0.00013578684012382198** underscores the accuracy of the model's predictions. The MAE serves as a metric to quantify the average magnitude of errors between the predicted and actual values. In this instance, the small value of MAE signifies minimal discrepancies between the predicted and actual values, suggesting that the model's predictions exhibit high fidelity to the true outcomes. The observation of numerous data points aligning closely with the trend line implies that the model captures the underlying patterns and relationships present in the data effectively. This alignment signifies that the model's predictions exhibit a strong correlation with the actual observations, bolstering confidence in its predictive capabilities. The scatter plot's depiction of the model's predictions in the correct direction, along with the clustering of data points around the trend line, underscores the model's robustness and accuracy. This alignment between predicted and actual values, coupled with the low MAE, affirms the model's efficacy in making reliable predictions. Overall, the plot serves as compelling evidence of the model's ability to accurately forecast outcomes based on the input data.

The best polynomial equation and plot the equation

Polynomial equation:

$y = 57.68 + 0.00PC1 + 0.00PC2 + 0.00PC1^2 - 0.00PC1 \times PC2 - 0.00PC2^2 + 0.00PC1^3 + 0.00PC1^2 \times 2PC2 - 0.00PC1 \times PC2^2 + 0.00PC2^3$

Plot for actual value versus predicted value



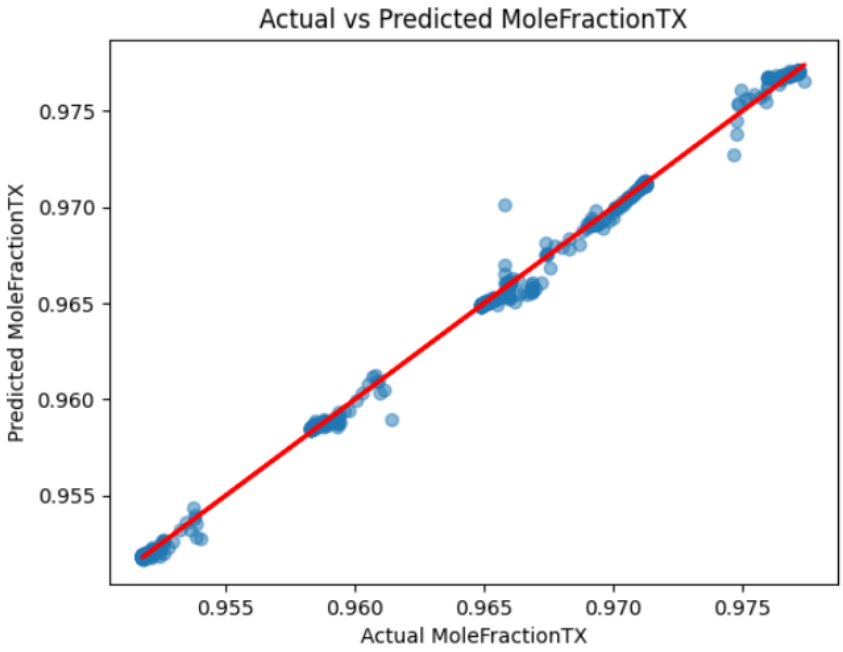
The scatter plot depicts the relationship between actual values (x-axis) and predicted values (y-axis), demonstrating the model's predictive performance. Notably, the majority of data points align closely with the trend line, indicating a strong correspondence between predicted and actual values. This alignment suggests that the model accurately captures the underlying trends present in the data, reinforcing its reliability in forecasting outcomes. The Mean Absolute Error (MAE) of 3.4606268532467245e-06 quantifies the average magnitude of errors between predicted and actual values. Such a small MAE value indicates minimal discrepancies between the model's

predictions and the true outcomes, underscoring the high level of accuracy achieved by the model. This low error rate reinforces confidence in the model's ability to make precise predictions. The clustering of data points around the trend line further supports the model's effectiveness in capturing the inherent patterns and relationships within the dataset. This clustering indicates that the model's predictions closely track the actual trends observed in the data, reinforcing its predictive capabilities and suggesting a strong correlation between predicted and actual values. Overall, the scatter plot provides compelling evidence of the model's proficiency in accurately forecasting outcomes. The alignment between predicted and actual values, coupled with the low Mean Absolute Error, serves as a testament to the model's robustness and reliability. These findings affirm the model's efficacy and underscore its potential for making accurate predictions based on the input data, thereby facilitating informed decision-making processes. But we say that the polynomial model predicts the value more accurately and we can conclude this as over fitted model.

Multilinear model for predicting the Mole Fraction of TX

$MoleFractionTX = -0.04x_1 + 0.01x_2 - 0.00x_3 - 0.00x_4 - 0.00x_5 - 0.00x_6 - 0.00x_7 + 0.18x_8 - 0.00x_9 - 0.00x_{10} + -0.02x_{11} + 0.01x_{12} + 0.01x_{13} + 0.03x_{14} - 0.02x_{15}$

Where x_o indicates the sensor data o.



The provided multilinear regression equation offers insights into the relationship between the MoleFractionTX and the various sensor data inputs (x_1 to x_{15}). Each coefficient in the equation represents the weightage assigned to the corresponding sensor data in predicting the MoleFractionTX. Starting with Sensor 1, its coefficient of -0.04 indicates a negative influence on the predicted MoleFractionTX, implying that increases in Sensor 1 data lead to decreases in the predicted MoleFractionTX. Conversely, Sensor 2's coefficient of 0.01 suggests a positive correlation, meaning higher values of Sensor 2 are associated with higher predict MoleFractionTX. Sensors 3 to 7 have coefficients close to zero, indicating minimal impact on the predicted MoleFractionTX. On the other hand, Sensor 8's coefficient of 0.18 signifies a strong positive influence, suggesting that higher values of Sensor 8 correspond to higher predicted MoleFractionTX. Sensors 9 and 10, with coefficients close to zero, have negligible influence, while Sensor 11's coefficient of -0.02 indicates a negative correlation, similar to Sensor 1.

Sensors 12 and 13 have coefficients of 0.01, indicating a positive impact on the predicted MoleFractionTX. Sensor 14's coefficient of 0.03 suggests a moderate positive influence, and Sensor 15, like Sensor 11, has a negative coefficient (-0.02), indicating an inverse relationship with the predicted MoleFractionTX. In summary, understanding the weightages assigned to each sensor data input is essential for interpreting the model's predictions. Sensors with higher absolute coefficients have a stronger influence on the predicted MoleFractionTX, while those with coefficients close to zero have minimal impact. This detailed analysis provides valuable insights into the individual contributions of each sensor to the prediction of MoleFractionTX, aiding in the interpretation and application of the multilinear regression model.

Purpose and Use Case

The purpose of distillation data fusion, especially employing techniques like multilinear, polynomial, and PCA curve fitting, is to enhance our understanding and optimization of distillation processes. By combining data from various sources and fitting it into mathematical models, we aim to capture the complex relationships within the distillation process more accurately. This allows us to make informed decisions regarding process control, optimization, and troubleshooting. By fusing data from sensors monitoring temperature, pressure, flow rates, and composition, and applying multilinear, polynomial, and PCA curve fitting techniques, operators can gain insights into the behaviour of the distillation columns. This can help identify potential inefficiencies, predict equipment failures, optimize operating conditions, and ultimately improve product quality and yield. By analysing distillation data using advanced modelling techniques, researchers can optimize purification processes, improve product consistency, and reduce production costs. This can lead to more efficient manufacturing processes and higher-quality. Overall, the purpose of distillation data fusion with multilinear, polynomial, and PCA curve fitting is to leverage AI/ML techniques to extract valuable insights from complex distillation data, leading to improved process efficiency, product quality, and operational performance across various industries.

Why to choose ml model over mathematical model: -

Choosing between a machine learning (ML) model and a mathematical model for predicting the mole fraction of distillate involves considering various factors such as data complexity, interpretability, accuracy requirements, and the nature of the problem at hand.

Machine learning models, particularly deep learning models, are adept at capturing intricate patterns and relationships in complex data. This makes them particularly useful when the data on distillation processes exhibits nonlinearity or involves interactions between numerous variables. Unlike traditional mathematical models, which may struggle to accommodate such complexities, ML models can effectively learn from the data and make accurate predictions.

Moreover, ML models offer flexibility and adaptability, allowing them to adjust to changing data patterns over time without requiring manual adjustments. This is particularly advantageous in dynamic distillation processes where conditions may evolve unpredictably, as ML models can continuously learn and update their predictions based on new information.

Handling high-dimensional data is another area where ML models shine. With techniques like feature selection and dimensionality reduction, ML models can extract meaningful information from datasets with a large number of variables or features. This ability is crucial in distillation processes where numerous factors may influence the mole fraction of distillate. **In terms of prediction accuracy, ML models, when properly trained and validated, can often outperform traditional mathematical models, especially when dealing with complex or noisy data**. Their ability to learn intricate patterns from the data enables them to make more accurate predictions, which is essential for optimizing distillation processes and achieving desired outcomes. Scalability is another advantage of ML models, as they can handle large datasets with ease. This scalability makes them suitable for applications where there is abundant data on distillation processes, allowing for comprehensive analysis and optimization.

Additionally, ML models offer automatic feature extraction, potentially eliminating the need for manual feature engineering, which can be time-consuming and domain-specific. By automatically learning relevant features from the data, ML models streamline the modeling process and reduce the burden on domain experts.

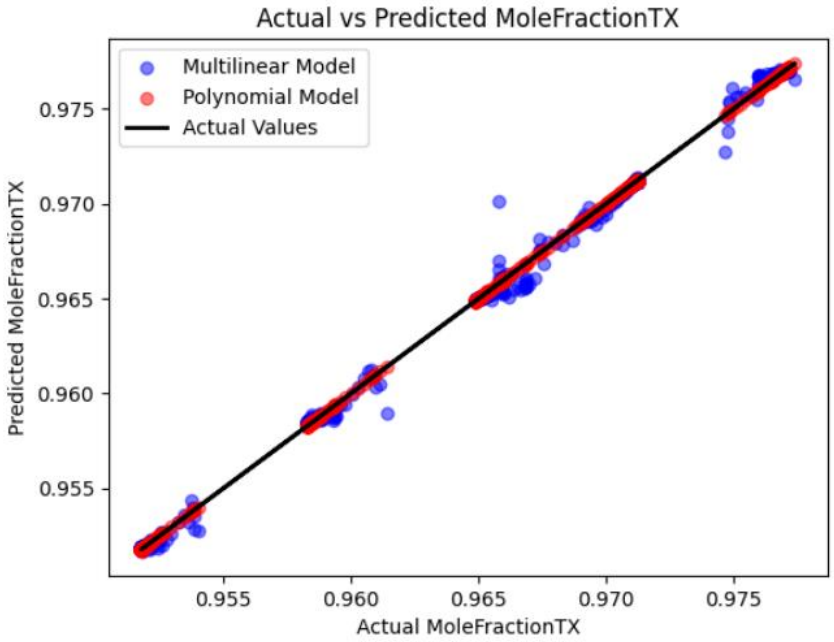
However, it's important to consider the trade-offs associated with ML models. They often require more data for training and can be computationally intensive, which may pose challenges in resource-constrained environments. Furthermore, ML models may lack the interpretability of traditional mathematical models, making it difficult to understand the underlying principles driving the predictions. In situations where interpretability is crucial or where a deep understanding of the physical processes is required, a mathematical model might be preferred despite potentially lower predictive performance.

Overall, the choice between an ML model and a mathematical model for predicting the mole fraction of distillate depends on the specific requirements of the problem, including the complexity of the data, the need for interpretability, and the desired level of prediction accuracy. By carefully evaluating these factors, practitioners can select the most appropriate modelling approach to optimize distillation processes and achieve desired outcomes.

Since the data which we get from the sensor is dynamics in nature and many complex variable data such as temperature, pressure, flowrate. If we choose mathematical model there will be a difficulty to capture the noisy data since the temperature and pressure is difficult to maintain constant so capture this fluctuations machine learning model is used to predict the distillate mole fraction

Conclusion

In conclusion, distillation data fusion, especially when utilizing techniques like multilinear, polynomial, and PCA curve fitting, offers immense potential for improving distillation processes across various industries. By integrating data from different sources and applying advanced mathematical models, we can gain a deeper understanding of the complex relationships within distillation systems. This enhanced understanding enables us to make more informed decisions regarding process control, optimization, and troubleshooting, ultimately leading to improved efficiency, product quality, and operational performance. where it facilitates better process management, predictive maintenance, and cost reduction. As we continue to leverage AI/ML techniques for distillation data fusion, we anticipate further advancements in process optimization and innovation, driving continuous improvements in industrial practices and product quality. the continued advancement of AI/ML techniques holds the promise of further improving distillation data fusion. As algorithms become more sophisticated and computing power increases, Furthermore, the application of advanced mathematical models such as multilinear, polynomial, and PCA curve fitting enables the extraction of valuable insights from complex distillation data. These models can capture nonlinear relationships, interactions between variables, and underlying patterns in the data, providing a more accurate representation of the distillation process. By fitting mathematical curves to the data, these techniques can uncover hidden correlations and dependencies, aiding in the identification of optimization opportunities and the prediction of future behaviour. we can expect even greater insights to be derived from distillation data, leading to more efficient processes, higher-quality products, and enhanced competitiveness for industries across the globe. By embracing these technological advancements and fostering collaboration between academia, industry, and technology providers, we can unlock the full potential of distillation data fusion and usher in a new era of innovation and excellence in distillation processes.



Based on the analysis, it appears that while the multilinear regression model provides reasonable predictions, its accuracy in aligning with the actual values is

suboptimal. Conversely, the polynomial model's predictions exhibit significantly higher accuracy compared to the actual values. However, it's important to consider the possibility of overfitting in the polynomial model. Overfitting occurs when a model learns to capture noise or random fluctuations in the training data rather than the underlying patterns. In the case of the polynomial model, its superior accuracy in predicting the target variable may be indicative of overfitting. This means that while the model performs exceptionally well on the training data, it may struggle to generalize to unseen data, leading to poor performance when applied to new datasets. On the other hand, although the multilinear regression model may not perform as well in terms of accuracy, it may offer better generalization to new data. Its simpler structure and lower risk of overfitting make it more robust and reliable in capturing the underlying relationships between the predictors and the target variable. In summary, while the polynomial model may demonstrate superior accuracy in predicting the target variable, its potential for overfitting raises concerns about its reliability in real-world applications. Conversely, the multilinear regression model, despite its lower accuracy, may offer better generalization and stability, making it a more preferable choice for practical use. Therefore, in this scenario, the multilinear regression model may be considered better than the polynomial model due to its potential for improved generalization and avoidance of overfitting.

References

1. Kacprzynski, G. J., Yang, F., & Georgakis, C. (2018). Multivariate statistical process monitoring using polynomial regression models
2. Shah, S. L., & Ponnambalam, K. (2019). Principal component analysis-based process monitoring
3. Xue, Z., Li, Y., Sun, L., & Lin, J. (2019). Study on modelling and control of a distillation process based on principal component analysis.
4. The data is collected from this site https://www.kaggle.com/datasets/amarhaiqal/aspen-hysys-distillation-column-data/discussion/485534