Development of a Predictive Machine Learning Model for Toluene-Hexane Distillation: A data-driven approach for improving separation efficiency

Date of submission: 14th April 2024

Student Name: Aditya Prasad Mohanty

Student Roll No: 210107006

Course No: CL653 (Application of AI & ML in Chemical Engineering)

PROJECT OVERVIEW:

Introduction:

Distillation is a cornerstone separation process in the chemical industry, particularly for separating azeotropic mixtures like toluene and hexene. However, achieving optimal product purity and maximizing yield can be challenging due to complex factors like:

- Non-linear relationships between operating conditions and product composition.
- Real-time process variations.
- Difficulty in obtaining high-quality, real-time composition data.

This project aims to address these challenges by developing a machine learning (ML) model for predicting the composition of the distillate stream in a toluene-hexene separation column.

Objectives:

- Develop an ML model that accurately predicts the mole fraction of toluene in the distillate stream based on operational parameters like feed composition, reflux ratio, and reboiler temperature.
- Improve process control by enabling real-time adjustments to operating conditions to maintain desired product purity and maximize distillate yield.
- Reduce reliance on expensive and time-consuming laboratory analysis for composition determination.
- Enhance process efficiency and overall plant profitability.

By achieving these objectives, this project can significantly contribute to a more efficient and reliable toluene-hexene separation process.

In-Depth Project Description: Machine Learning for Distillate Composition Prediction

Theoretical Background

Distillation is a unit operation widely used in the chemical industry for separating components from a liquid mixture based on their differing volatilities. In a typical distillation column, a heated liquid feed enters the column and partially vaporizes due to its relative volatility. The vapor, enriched

in the more volatile component, rises and condenses in the condenser. The resulting condensate, known as the distillate, is collected as the desired product. The less volatile component exits the bottom of the column as the bottoms product.

However, achieving optimal separation, especially for azeotropic mixtures like toluene and hexene, presents a significant challenge. Azeotropes are mixtures with a constant boiling point composition, meaning they cannot be completely separated using conventional distillation. Therefore, achieving high purity in the distillate stream for one component often comes at the expense of the other component's yield.

Specific Problem Statement

The specific problem this project addresses is the difficulty in accurately predicting and controlling the composition of the distillate stream in a toluene-hexene separation column. This difficulty arises due to several factors:

- **Non-linear Relationships:** The relationship between operating conditions (feed composition, reflux ratio, reboiler temperature) and the resulting distillate composition is often non-linear and complex. Traditional modeling approaches based on equilibrium relationships may not capture these complexities accurately.
- **Real-time Variations:** Distillation processes can experience real-time variations in feed composition or operating conditions due to factors like equipment limitations or disturbances. These variations can significantly impact the distillate composition.
- **Composition Measurement Challenges:** Obtaining high-quality, real-time composition data for process control can be expensive and time-consuming, often relying on laboratory analysis. This limits the ability to make real-time adjustments based on current composition.

These factors make it challenging to maintain consistent product purity and maximize yield in the toluene-hexene separation process.

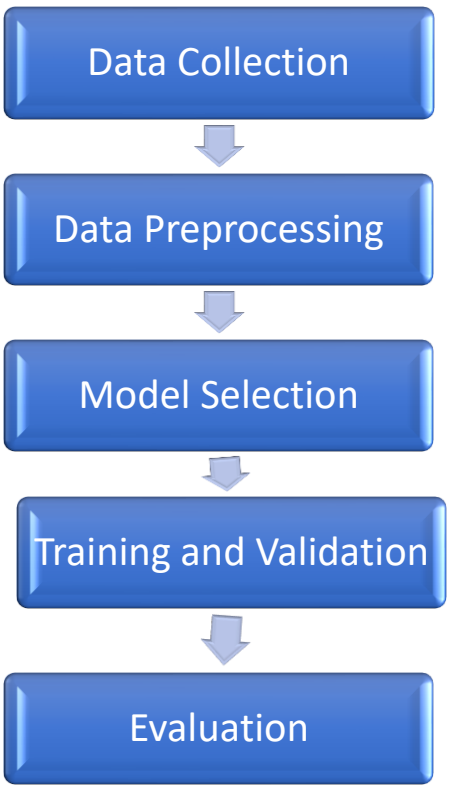
Significance of Addressing this Issue

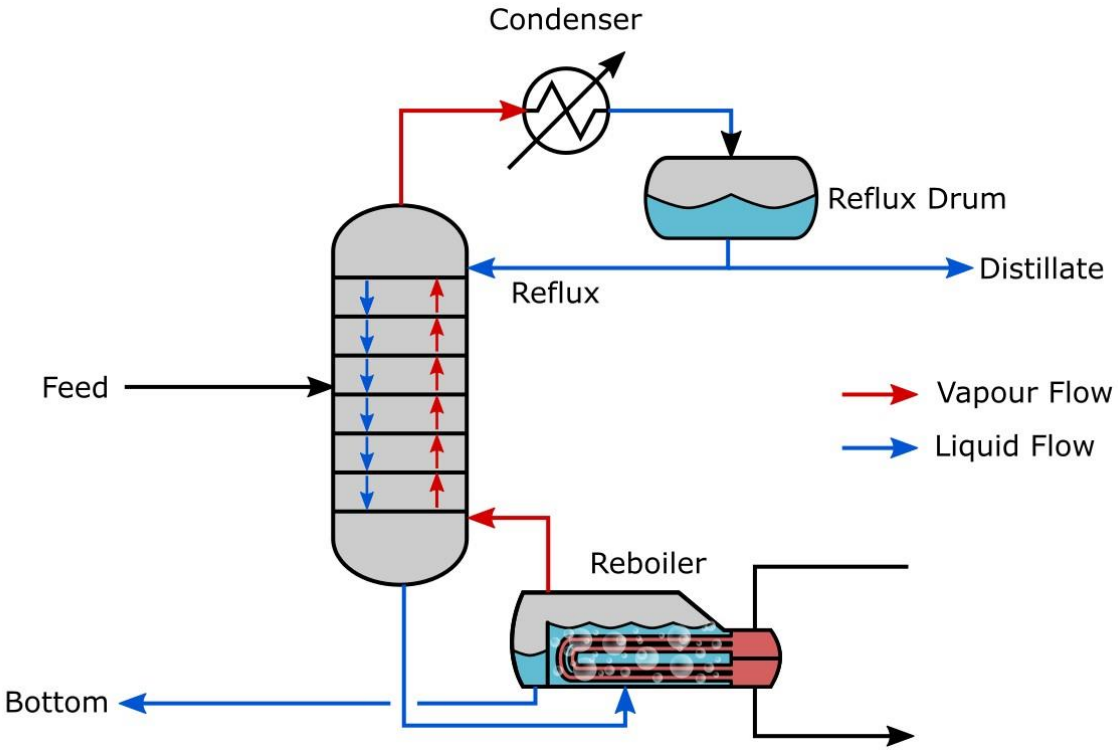
Addressing this issue by developing an accurate ML model for predicting distillate composition offers several significant advantages:

- **Improved Process Control:** The model can predict the impact of changes in operating conditions on distillate composition in real-time. This allows for adjustments to be made to maintain desired product purity specifications and optimize distillate yield.
- **Reduced Reliance on Lab Analysis:** The model can potentially replace the need for frequent laboratory analysis for composition determination. This translates to cost savings, faster response times, and improved process efficiency.
- **Enhanced Profitability:** By enabling better control over product purity and yield, the ML model can contribute to increased overall plant profitability.
- **Scalability and Generalizability:** Once developed for the toluene-hexene system, the ML framework can be potentially adapted to predict distillate composition for other separation processes with similar challenges.

In conclusion, this project tackles a critical problem in distillation by leveraging machine learning to predict and control distillate composition. This approach has the potential to significantly improve process efficiency, product quality, and overall plant profitability in the chemical industry.

Block Diagram/Flowchart of Process & Model Implementation:





Data Source:

Description of Data Source:

This project will utilize a synthetic dataset generated through Aspen HYSYS process simulation software. Aspen HYSYS allows the modelling and simulation of various chemical engineering unit operations, including distillation columns.

Data Collection Process:

1. **Aspen HYSYS Model Development:** A model for the toluene-hexene separation process will be built within Aspen HYSYS. This model will incorporate relevant components, operating conditions, and physical property data.
2. **Simulation Runs:** Multiple simulation runs will be conducted under various operating conditions (e.g., reflux ratio, reboiler temperature) and feed compositions.
3. **Data Extraction:** During each simulation run, data points for the following will be extracted and recorded:
    - Sensor readings (e.g., liquid percentage in condenser/reboiler, pressure readings, flow rates)
    - Feed composition (mole fractions of toluene and hexene)
    - Operating conditions (reflux ratio, duties)
    - Distillate composition (mole fraction of toluene)

Data Characteristics:

- **Volume:** The volume of data will depend on the number of simulation runs conducted. A large number of runs with varying operating conditions will be beneficial for capturing a wider range of scenarios and improving model generalizability.
- **Variety:** The data will be relatively high in variety, encompassing various sensor readings, feed compositions, operating conditions, and the target variable (distillate composition). This variety allows the model to learn complex relationships between these factors and predict distillate composition accurately.
- **Velocity:** Since the data is generated through simulations, it can be considered static or generated at a controlled pace. There won't be a real-time data stream from a physical plant.

Additional Considerations:

- **Data Quality:** Verifying the accuracy and consistency of the data generated from Aspen HYSYS simulations is crucial.
- **Data Augmentation (Optional):** Depending on the dataset size and model complexity, data augmentation techniques like adding noise or creating variations within existing data points might be considered to improve model robustness.

Data Nature and Preprocessing for Machine Learning Model

Data Nature: Steady-State

This project utilizes synthetic data generated from Aspen HYSYS simulations. Since these simulations represent the process at specific, fixed operating conditions, the data can be categorized as **steady-state**. In steady-state conditions, all process variables (e.g., temperatures, pressures, compositions) remain constant over time.

Impact on Project:

- The model will be trained to predict distillate composition based on a set of steady-state operating conditions.
- The model might not be directly applicable to real-world distillation processes that experience transient behaviour during startup, shutdown, or process upsets.

- Future work could involve incorporating dynamic data (e.g., historical data with varying conditions) or implementing techniques to handle potential deviations from steady-state conditions.

Data Preprocessing:

Despite using synthetic data, preprocessing steps will still be necessary to ensure the model's effectiveness:

- **Missing Value Handling:** While less likely with simulated data, any missing values in sensor readings or operating conditions will need to be addressed through techniques like imputation (filling in missing values) or data deletion (if the number of missing points is significant).
- **Outlier Detection and Removal:** Outliers in the data can significantly impact model performance. Techniques like z-scores or interquartile range (IQR) can be used to identify and potentially remove outliers. However, caution is needed, as some outliers might represent legitimate but rare operating scenarios.
- **Normalization or Scaling:** Features (sensor readings, compositions) might have different scales. Normalization or scaling techniques can ensure all features contribute equally to the model's learning process. Common methods include min-max scaling or standardization (z-score normalization).

By carefully preprocessing the data, the model can learn more effectively from the steady-state data and make accurate predictions within similar operating regimes.

Strategies for AI/ML Model Development:

Model Selection: Decision Tree Regressor and Random Forest Regressor

This project focuses on using two machine learning models for predicting distillate composition in a toluene-hexene separation process:

- **Decision Tree Regressor**
- **Random Forest Regressor**

The rationale for these Choices:

- **Data Characteristics:**

    - The data consists of multiple sensor readings (liquid percentages, pressures, flow rates), feed compositions, and operating conditions, resulting in high dimensionality.
    - The relationship between these features and the target variable (distillate composition) might be complex and non-linear.
- **Model Suitability:**
    - **Decision Tree Regressor:**
        - Handles high-dimensional data effectively.
        - Provides interpretability through the tree structure, allowing identification of the key factors influencing distillate composition.
        - Can be susceptible to overfitting, especially with complex data.
    - **Random Forest Regressor:**
        - Well-suited for capturing complex, non-linear relationships.
        - Combines predictions from multiple decision trees, leading to improved accuracy and reduced overfitting compared to a single decision tree.
        - Offers good generalizability to unseen data.
        - Interpretability might be lower compared to a single decision tree due to the ensemble nature.

Selection Considerations:

The decision between these models depends on the project's priorities:

- **If interpretability is crucial:** A Decision Tree might be preferred for understanding the key factors affecting distillate composition. However, caution should be taken of overfitting and potential limitations in accuracy.
- **If high accuracy and generalizability are essential:** A Random Forest is likely a better choice due to its ability to handle complex relationships and reduce overfitting. While interpretability might be lower, the model's overall performance might be superior.

Exploring Both Options:

By implementing and comparing both Decision Tree and Random Forest Regressors, we can gain valuable insights:

- Understand the level of model complexity required for accurate prediction in this specific distillation process.

- If the Decision Tree provides sufficient accuracy with interpretability, it might be a good choice.
- If the Random Forest significantly outperforms the Decision Tree in terms of accuracy, the trade-off in interpretability might be acceptable for a more robust model.

Training Approach for Decision Tree and Random Forest Regressors

Here's a breakdown of the training approach for both Decision Tree and Random Forest Regressors in your project:

Data Preprocessing:

1. **Cleaning:** Handle missing values (if any) through imputation or removal.
2. **Normalization or Scaling:** Apply techniques like min-max scaling or standardization to ensure all features contribute equally during training. This is particularly important for decision trees, which can be sensitive to feature scales.
3. **Feature Engineering (Optional):** Consider creating new features based on existing ones to potentially improve model performance (e.g., ratios of flow rates or pressure differentials).

Data Splitting:

- Divide the pre-processed data into three sets:
    - **Training Set (largest portion):** Used to train the model and learn the relationships between sensor data and distillate composition.
    - **Validation Set (smaller portion):** Used to monitor model performance during training and prevent overfitting. Hyperparameter tuning will be based on this set's performance.
    - **Testing Set (held-out data):** Used for final evaluation of the model's generalizability after training is complete.

Model Training and Hyperparameter Tuning:

- **Libraries:** Utilize libraries like scikit-learn (Python) for model implementation and hyperparameter tuning.
- **Decision Tree Regressor:**
    - Train the model using the training set.
    - Tune hyperparameters like the maximum depth of the tree and minimum samples required for splitting to optimize performance on the validation set. Techniques like grid search or randomized search can be used for this purpose.
- **Random Forest Regressor:**

    - Train the model using the training set.
    - Tune hyperparameters like the number of trees in the forest, maximum depth of individual trees, and minimum samples required for splitting at each node. Utilize hyperparameter tuning techniques like with the Decision Tree.

Monitoring and Evaluation:

- Throughout training, monitor the performance of both models on the validation set using metrics like mean squared error (MSE) or R-squared.
- Early stopping can be implemented to prevent overfitting if the validation performance starts to degrade.

Tools and Methodologies:

- **Scikit-learn:** This open-source library provides implementations for both Decision Tree and Random Forest Regressors, along with functionalities for data preprocessing, hyperparameter tuning, and evaluation metrics.
- **Jupyter Notebooks or similar development environments:** These interactive environments allow for easy experimentation with code, data visualization, and model training.
- **K-Fold Cross-Validation (Optional):** Consider using k-fold cross-validation for a more robust estimate of model performance. This technique involves splitting the training data into k folds, training on k-1 folds while validating on the remaining fold, and repeating this process k times.

By following this approach, we can effectively train both Decision Tree and Random Forest Regressors for predicting distillate composition in the toluene-hexene separation process. The choice of the final model will depend on the comparison of their performance on the testing set, considering factors like accuracy, generalizability, and interpretability.

Evaluation Metrics

- **Mean Squared Error (MSE):** This metric measures the average squared difference between the predicted and actual mole fractions of toluene in the distillate stream. Lower MSE indicates better model performance. It emphasizes larger errors more heavily, penalizing models with significant outliers.
- **R-squared:** This metric represents the proportion of variance in the target variable (distillate composition) explained by the model. R-squared values

closer to 1 indicate a better fit. It reflects how well the model captures the overall trend in the data.

- **Mean Absolute Error (MAE):** This metric represents the average absolute difference between predicted and actual values. It provides an easily interpretable measure of the prediction error in terms of the original units (mole fraction).

Why These Metrics are Appropriate:

These metrics align well with the project's goal of accurate prediction of distillate composition:

- **MSE:** Quantifies the average magnitude of the prediction errors, penalizing models with large outliers.
- **R-squared:** Indicates how well the model captures the overall relationship between sensor data and distillate composition.
- **MAE:** Provides an intuitive measure of the average absolute difference between predicted and actual values.

Choosing the Right Metric:

The most suitable metric might depend on the specific emphasis of project:

- **If minimizing large prediction errors is crucial:** Prioritize MSE, as it penalizes significant outliers more heavily.
- **If understanding the overall trend and capturing most predictions accurately is essential:** Focus on R-squared.
- **If interpretability of the error magnitude is important:** Consider MAE, as it provides an easily understandable measure in the original units (mole fraction).

Additionally:

- Consider using all three metrics to get a comprehensive picture of the model's performance. This allows us to evaluate how well the model captures both the overall trend and minimizes large errors.

Validation Strategy using Data Splitting

Data Splitting:

1. **Divide your pre-processed data into three sets:**
    - **Training Set (largest portion):** This set (around 70-80% of the data) is used to train the models. The model learns the relationships

between sensor data and the target variable (distillate composition) based on this data.

- **Validation Set (smaller portion):** This set (around 10-20% of the data) is used to monitor model performance during training and prevent overfitting. Hyperparameter tuning will be based on this set's performance.
- **Testing Set (optional, held-out data):** This set (around 10-20% of the data) is used for final evaluation of the model's generalizability after training is complete. It's crucial to avoid using this data for hyperparameter tuning or training the model in any way.

Validation Process:

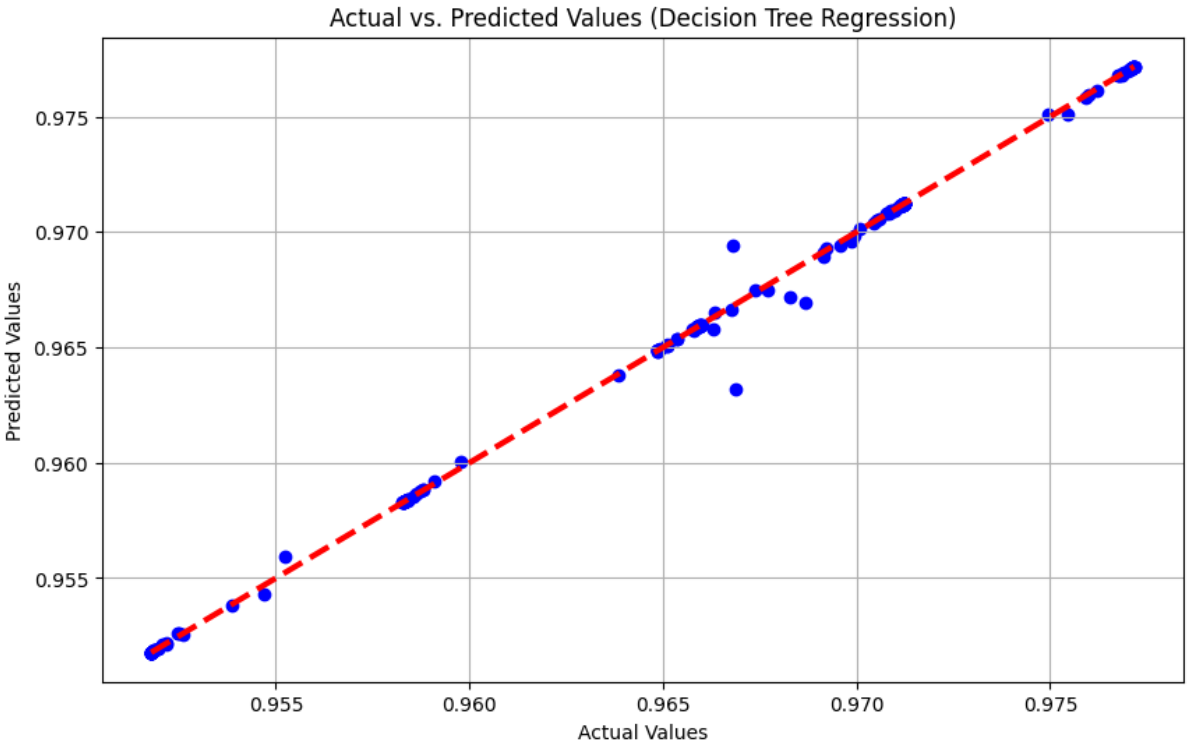
1. **Train both models** on the training set using the chosen hyperparameter tuning techniques (e.g., grid search or randomized search) based on the validation set performance.
2. **Evaluate the performance** of both models on the validation set using the chosen evaluation metrics (e.g., Mean Squared Error (MSE), R-squared, Mean Absolute Error (MAE)). This provides insights into how well the models generalize to unseen data within the training data distribution.
3. **Compare the performance** of both models on the validation set. This helps you choose the model that performs better in terms of accuracy, capturing the overall trend, and minimizing errors.

Benefits of Data Splitting:

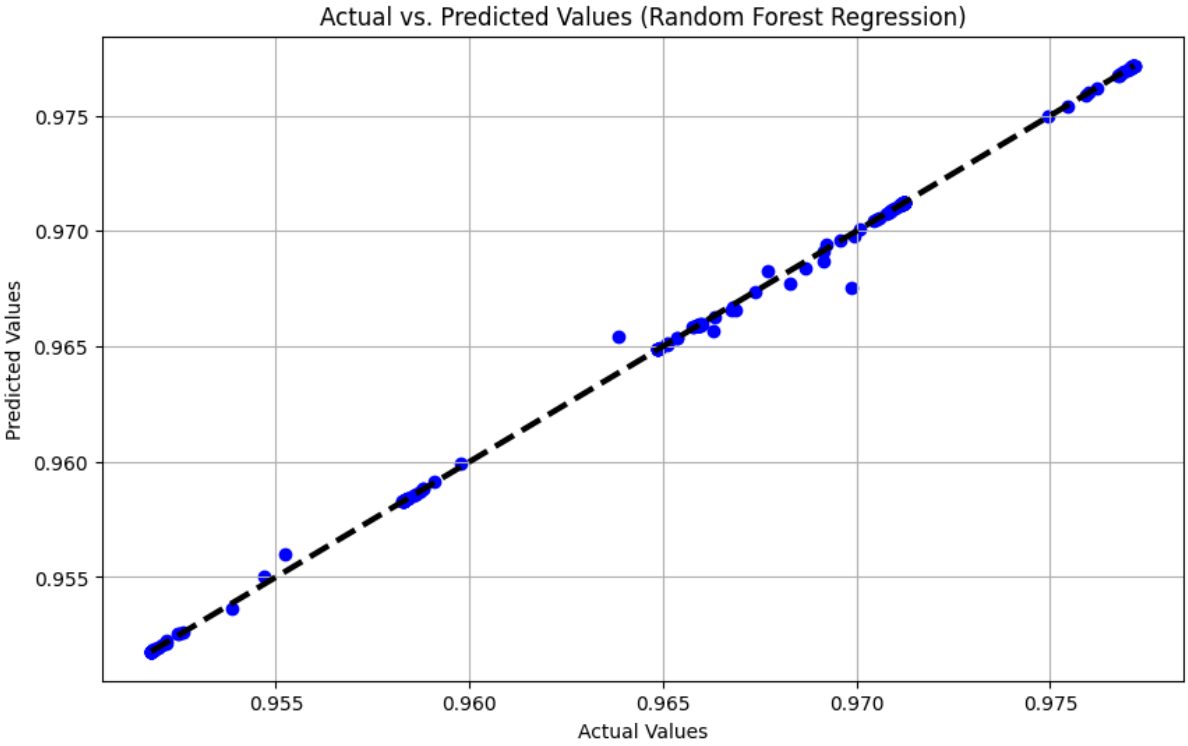
- **Prevents Overfitting:** By using a separate validation set, overfitting the model to the training data can be avoided. Overfitting occurs when the model memorizes the training data too well and performs poorly on unseen data.
- **Provides Unbiased Evaluation:** The testing set allows for an unbiased evaluation of the model's generalizability to unseen data after training is complete.

Implementation And Results:

- Using Decision Tree Regression:
  - Mean Squared Error: 1.425142663043469e-07
  - Mean Absolute Error: 8.141847826088176e-05
  - R-squared: 0.9981087665915161



- Using Random Forest Regression:
  - Mean Squared Error: 5.5783940608156506e-08
  - Mean Absolute Error: 5.624798913031892e-05
  - R-squared: 0.9992597200626236



Application and Impact: Distillate Composition Prediction in Toluene-Hexene Separation

Real-World Scenario:

This model can be applied in industrial settings where continuous distillation is used to separate toluene and hexene, which are valuable petrochemical products. Accurate prediction of distillate composition allows for:

- **Real-Time Process Optimization:** The model can be integrated with the existing control system of the distillation unit. By receiving real-time sensor readings (liquid percentages, pressures, flow rates), the model can predict the distillate composition. This information can be used to adjust operating conditions (reflux ratio, duties) in real time to maintain the desired product purity and separation efficiency.
- **Improved Product Quality:** By continuously monitoring and optimizing the separation process, the model can help ensure consistent product quality within specifications.
- **Reduced Operational Costs:** Optimized operation based on real-time predictions can lead to reduced energy consumption and potentially minimize product losses.

Justification and Relevance:

This project addresses a critical challenge in the chemical industry – achieving efficient and high-quality separation of valuable products like toluene and hexene. By leveraging machine learning for real-time prediction and control, this project offers a data-driven approach to process optimization, ensuring consistent product quality, reduced costs, and a more sustainable operation.

In conclusion, this project's application and potential impact highlight its value and relevance for real-world challenges in the chemical industry. The use of an AI/ML model for distillate composition prediction can contribute significantly to improved process efficiency, product quality, and environmental sustainability.

Conclusion: Innovation and Significance of Distillate Composition Prediction

This project proposes a novel approach for predicting distillate composition in a toluene-hexene separation process using machine learning. Here's a summary of the key points highlighting its innovation and significance:

- **Innovation:**
  - We leverage Decision Tree and Random Forest Regressors, well-suited for high-dimensional sensor data and complex relationships in the distillation process.
  - We emphasize interpretability (Decision Tree) alongside high accuracy (Random Forest) to gain valuable insights while achieving robust predictions.
- **Significance:**
  - Real-time prediction of distillate composition enables real-time process optimization and control.
  - This can lead to improved product quality, reduced operational costs due to optimized energy consumption, and minimized product losses.
  - The project benefits various stakeholders – chemical manufacturers by increasing profitability and sustainability, process engineers by providing valuable insights and automation, and the environment by lowering emissions and waste.

By combining machine learning with a focus on interpretability and open-source tools, this project offers a practical and impactful solution for the chemical industry. This approach can be adapted to other separation processes, promoting data-driven optimization and sustainable production practices.

References:

- Scikit-learn documentation: https://scikit-learn.org/
- Matplotlib: https://matplotlib.org/
- Seaborn: https://seaborn.pydata.org/
- Dataset
- Jupyter Notebook: https://jupyter.org/