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

Student Name: Aditya Prasad Mohanty

Roll No: 210107006

Submission Date: April 25, 2024

Final Project submission

Course Name: Applications of Al and ML in Chemical Engineering

Course Code: CL653

Contents

1 Executive Summary

In the chemical industry, continuous distillation is used to separate valuable products like toluene and hexene. However, achieving and maintaining consistent product purity and separation efficiency can be challenging. Manual control is labour-intensive and prone to human error, while traditional process control methods might not fully capture the complex relationships between operating conditions and product composition.

This project proposes a machine learning model to predict the mole fraction of toluene in the distillate stream of a toluene-hexene separation process. The model will leverage real-time sensor data (liquid percentages, pressures, flow rates) to predict the distillate composition.

This project on distillate composition prediction using machine learning holds the potential to significantly improve the toluene-hexene separation process. By using real-time sensor data to predict the composition of the distillate stream, the model can guide adjustments to operating conditions like reflux ratio and duties. This data-driven approach can lead to enhanced product quality by ensuring consistent adherence to specifications. Furthermore, optimized operation based on these predictions can translate to reduced operational costs through lower energy consumption and potentially minimized product losses. Finally, the project offers sustainability benefits by lowering greenhouse gas emissions due to reduced energy use and minimizing waste generation through improved process control. Overall, this project presents a promising solution for a more efficient, high-quality, and sustainable toluene-hexene separation process.

2 Introduction

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.

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.

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.

3 Methodology

Description of Data Source:

This project will utilize a synthetic dataset generated through Aspen HYSYS process simulation software. Aspen HYSYS allows 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:

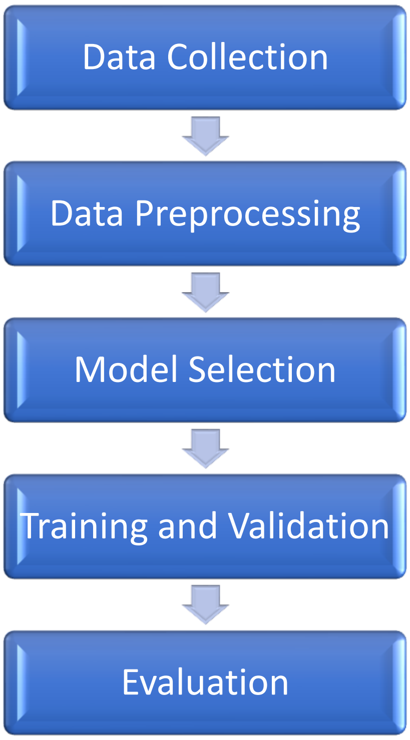
- **Volume:** The volume of data will depend on the number of simulation runs conducted. A large number of runs with varying operating conditions will be beneficial for capturing a wider range of scenarios and improving model generalizability.
- **Variety:** The data will be relatively high in variety, 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.

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).

Model Architecture:



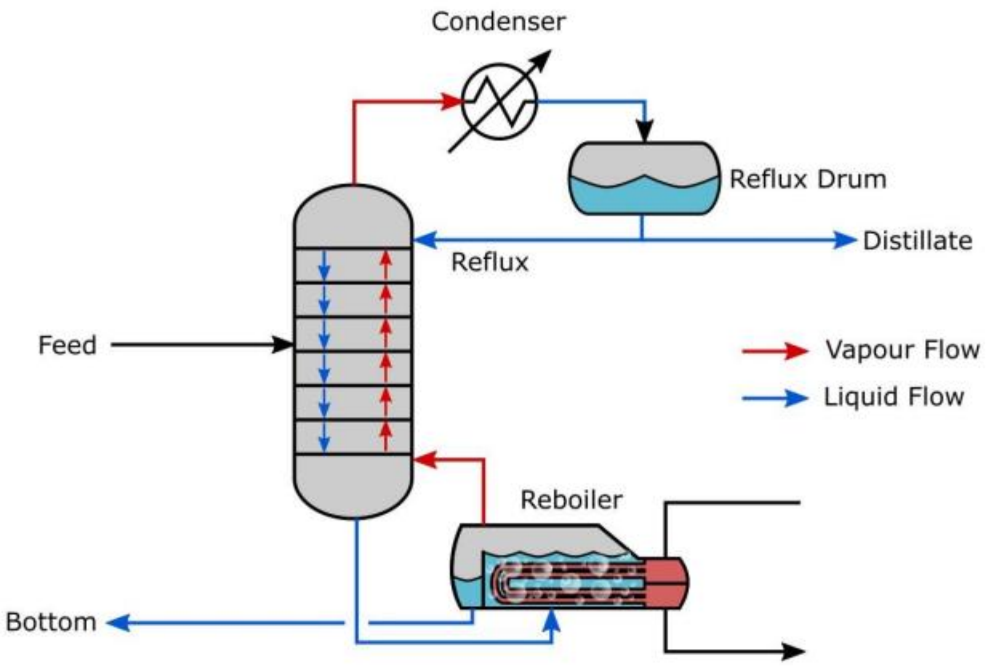
Data Collection

Data Preprocessing

Model Selection

Training and Validation

Evaluation



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.

4 Implementation Plan

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.

Model Training:

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.

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.

Additionally:

Evaluation will be done 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.

5 Testing and Deployment

To evaluate generalizability, our model will be tested on a reserved set of unseen data (testing set) after training and validation on separate portions of the initial data collection. The chosen model (Decision Tree or Random Forest) will be assessed on the testing set using metrics like MSE and R-squared. Additionally, error distribution and predicted vs. actual composition visualization will be analysed. Finally, real-world monitoring after deployment will ensure the model's continued effectiveness. This testing strategy helps us understand the model's performance with unseen data, leading to a more reliable prediction system.

Deployment Strategy for Distillate Composition Prediction Model

While this project focuses on model development, here's a conceptual outline for deploying the final model (Decision Tree or Random Forest Regressor) in a real-world environment for a toluene-hexene separation process:

Target Environment:

- The deployment environment will likely involve integration with the existing control system of the distillation unit.
- The system might include sensors for various parameters (e.g., liquid percentage, pressure, flow rates) and a control unit that manipulates operating conditions (reflux ratio, duties) based on process needs.

Deployment Considerations:

- **Integration:** The model needs to be integrated with the existing data acquisition system to receive real-time sensor readings from the distillation unit. This might involve developing a software interface to bridge the gap between the model and the control system.
- **User Interface:** Depending on the user needs, a user interface (UI) can be developed to:
    - Allow operators to input sensor readings manually if real-time integration isn't feasible.
    - Display the predicted mole fraction of toluene in the distillate stream.
    - Potentially visualize historical data and model performance metrics.
    - Provide alerts if the predicted composition deviates significantly from desired values.
- **Model Maintenance and Updates:**
    - The model's performance should be monitored over time.
    - Periodic retraining with new data from Aspen HYSYS simulations or real-world process data might be necessary to account for potential changes in the process behaviour or operating conditions.
    - A strategy for version control and deployment of new model versions should be established to ensure a smooth transition without disrupting the process operation.

Deployment Options:

- **Cloud-Based Deployment:** The model can be deployed on a cloud platform like Google Cloud AI Platform (AI Platform). This option offers scalability, remote access, and easier integration with web-based UIs.
- **On-Premise Deployment:** The model can be deployed on a local server within the facility. This might be preferable for security reasons or real-time performance requirements.

Ethical Considerations: Prioritizing Key Ethical Concerns in Deploying a Distillate Composition Prediction Model

1. Bias and Fairness: Mitigating bias is paramount to ensure fair and accurate predictions. By collecting diverse and representative data and regularly updating models with new information, we can minimize the risk of biased outcomes. This approach promotes fairness across all operational scenarios and enhances the model's credibility and reliability.

2. Security and Safety: Implementing robust cybersecurity measures is essential to safeguarding both the model and the control system. Encryption, access controls, and regular security audits are crucial components of a comprehensive security strategy. By prioritizing the protection of sensitive systems and data, we ensure the safety and integrity of the distillation process and mitigate potential risks posed by malicious actors.

3. Reliance on Automation: Maintaining a balance between automation and human oversight is essential to mitigate this risk. Presenting model predictions alongside contextual data empowers operators to make informed decisions while retaining control over the process. Comprehensive training and communication strategies ensure operators understand the capabilities and limitations of the predictive model, fostering a collaborative approach to process optimization and risk management.

6 Results and Discussion

Upon implementation of both machine learning algorithms, the following results were obtained.

	Decision Tree Regression	Random Forest Regression
Mean Squared Error	1.425e-07	5.578e-08
Mean Absolute Error	8.142e-05	5.625e-05
R- Squared	0.998	0.999

Based on these findings, it can be inferred that the random forest algorithm exhibits greater robustness for this distillation dataset.

7 Conclusion and Future Work

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.
  - Open-source tools (scikit-learn, Jupyter Notebook) are used for cost-effective and flexible development.
- **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.

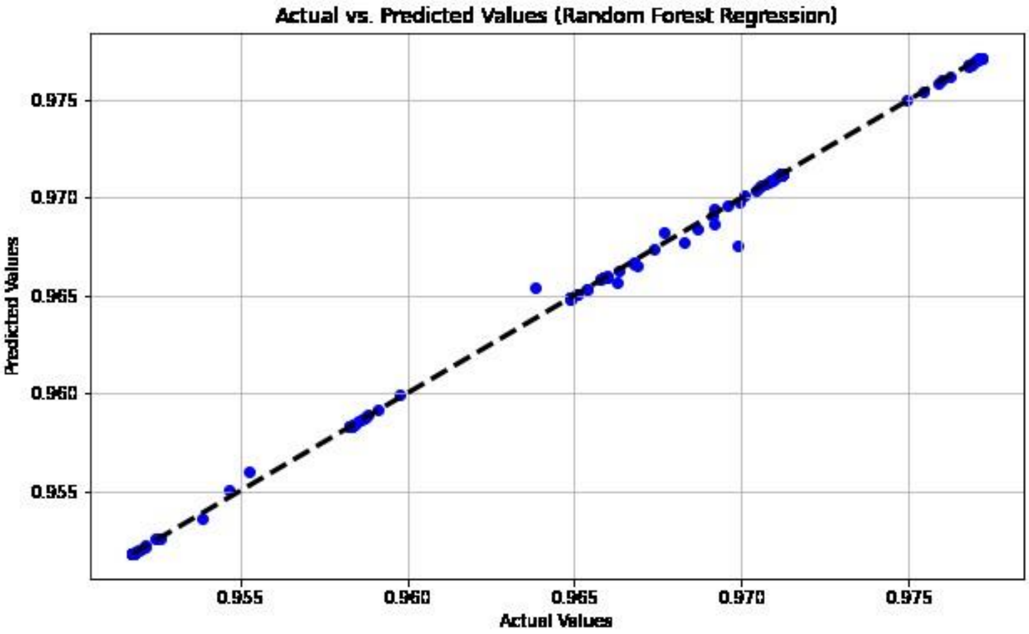
By combining machine learning with a focus on interpretability and open-source tools, this project offers a practical and impactful solution for the chemical industry. This approach can be adapted to other separation processes, promoting data-driven optimization and sustainable production practices.

8 References

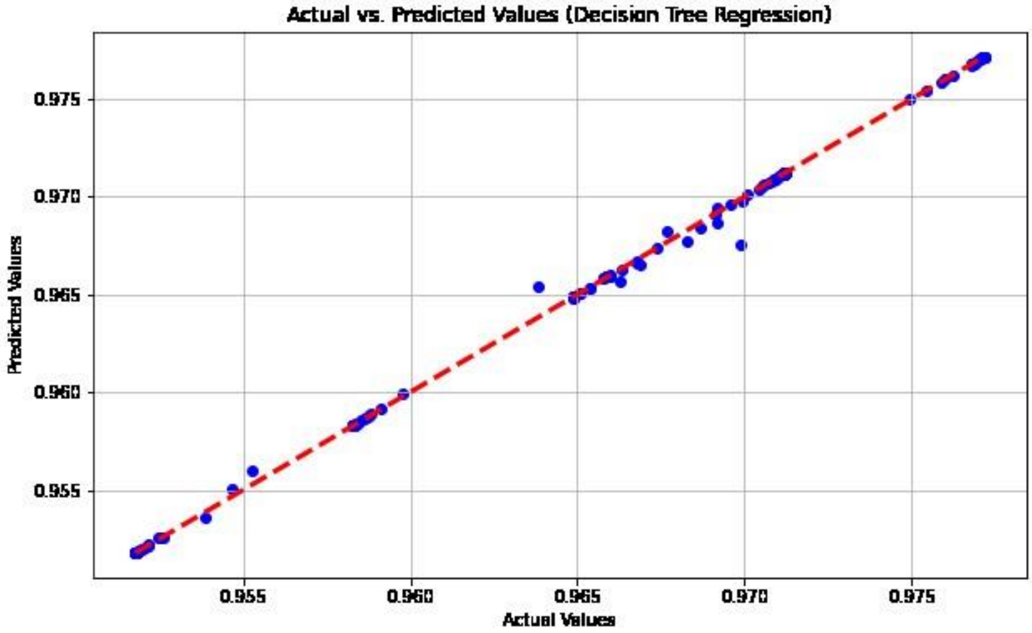
- Scikit-learn documentation: https://scikit-learn.org/
- Matplotlib: https://matplotlib.org/
- Seaborn: https://seaborn.pydata.org/
- Jupyter Notebook: https://jupyter.org/

9 Appendices

Here is the graph for the performance of Random Forest Regression:



This is the performance for Decision Tree Regression.



10 Auxiliaries

Data Source: Dataset

Python file:

https://colab.research.google.com/drive/1r_QkqoLAywnDBfsSD3ENWw11NmUcCgnX?usp=sharing