Optimizing Distillation Column Operations: Integration and Maintenance of AI Models

Rajat Singhal

210107100

Submission Date: April 26, 2024

Final Project submission

Course Name : Applications of Al and ML in chemical engineering

Course Code: CL653

Contents

1 Executive Summary

This project aims to develop a robust AI/ML model for simulating distillation column performance under industrial conditions, addressing challenges posed by noise, outliers, and missing data. Distillation columns play a critical role in chemical processes, but their efficiency can be hindered by various industrial factors. By incorporating these factors into a mathematical model, we seek to accurately simulate real-world distillation column operation.

Problem Statement:

Distillation columns are essential components in chemical processing industries, but their performance can be suboptimal due to challenges such as noise, outliers, and missing data. This project aims to develop an AI/ML model that can effectively predict key parameters for distillation column optimization, specifically targeting ethanol concentration as a critical performance metric.

Proposed Solution:

The proposed solution involves leveraging AI and machine learning techniques to develop a predictive model capable of simulating distillation column operation. By utilizing essential parameters such as pressure, temperature, flow rates, and ethanol concentration from a provided dataset, we aim to optimize column performance and address industrial challenges effectively.

Methodologies:

Data Preprocessing Pipeline: We will start by removing redundant columns and handling missing values to ensure data integrity. Outliers will be addressed using appropriate techniques to enhance model robustness. Input data normalization using Min-Max scaling will be applied to mitigate scale differences across features.

Model Development: We have selected **Linear Regression (LR), Support Vector Regression (SVR),** and **Random Forest Regression** as primary models. These models

will be trained using the pre-processed data to predict ethanol concentration and optimize distillation column operation.

Expected Outcomes:

The project aims to identify the most accurate regression algorithm for predicting ethyl concentration in a distillation column, providing valuable insights into the relationships between input features and ethyl concentration. These insights aid in optimizing processes, contributing significantly to predictive modelling advancements in chemical engineering. Ultimately, this work facilitates better decision-making and enhances process efficiency within chemical engineering applications.

2 Introduction

Background:

Distillation columns are fundamental units in chemical engineering processes, serving to separate liquid mixtures into their component parts based on differences in volatility. The efficiency and effectiveness of distillation columns have significant implications for production costs, energy consumption, and overall product quality within chemical manufacturing. Traditional modeling approaches often fall short in accurately representing real-world conditions due to oversimplification and neglect of industrial complexities such as noise, outliers, and missing data.

Problem Statement:

The specific problem addressed in this project is the need for a robust AI/ML model capable of accurately simulating distillation column operation under realistic industrial conditions. Existing models frequently struggle to account for the variability and uncertainties present in industrial settings, leading to suboptimal performance and operational inefficiencies. By incorporating atypical data features like noise, outliers, and missing data, we aim to develop a model that accurately reflects the complexities of distillation column operation and enables more informed decision-making in process optimization.

The significance of addressing this issue lies in its direct impact on the efficiency and profitability of chemical processes. Improved simulation models facilitate enhanced product quality and reduced resource consumption, allowing engineers to anticipate and mitigate operational challenges, resulting in smoother production processes and fewer disruptions. By

developing a model capable of handling industrial complexities, we contribute to advancing the field of AI/ML in chemical engineering, opening doors to more reliable and practical applications in industrial settings.

Objectives:

The main objectives of this project are as follows:

1. Develop a robust AI/ML model capable of simulating distillation column performance under realistic industrial conditions.
2. Address challenges posed by noise, outliers, and missing data through appropriate data preprocessing techniques.
3. Optimize distillation column operation by predicting key parameters such as ethanol concentration using the developed model.
4. Validate the effectiveness of the model for soft-sensors validation and feature selection algorithm validation, enhancing its reliability and applicability in industrial settings.

3 Methodology

Data Source: The dataset used in this project comprises 4408 entries capturing various input and output parameters for a distillation column. The data sources include literature references and datasets sourced from Kaggle, an open-source platform for sharing datasets and projects. Ethical considerations and data privacy norms were adhered to throughout the acquisition and use of the dataset.

One of the literature sources referenced for this project is a research paper titled "Un algoritmo de selección de variables de enfoque híbrido basado en información mutua para aplicaciones de sensores blandos industriales basados en datos" by Cote-Ballesteros, J. E., Grisales Palacios, V. H., & Rodriguez-Castellanos, J. E. (2022), published in Ciencia E Ingeniería Neogranadina (Vol. 32, No. 1, pp. 59–70). This paper contributed valuable insights into variable selection methods and sensor applications, aligning with the project's focus on distillation column optimization and predictive modeling.

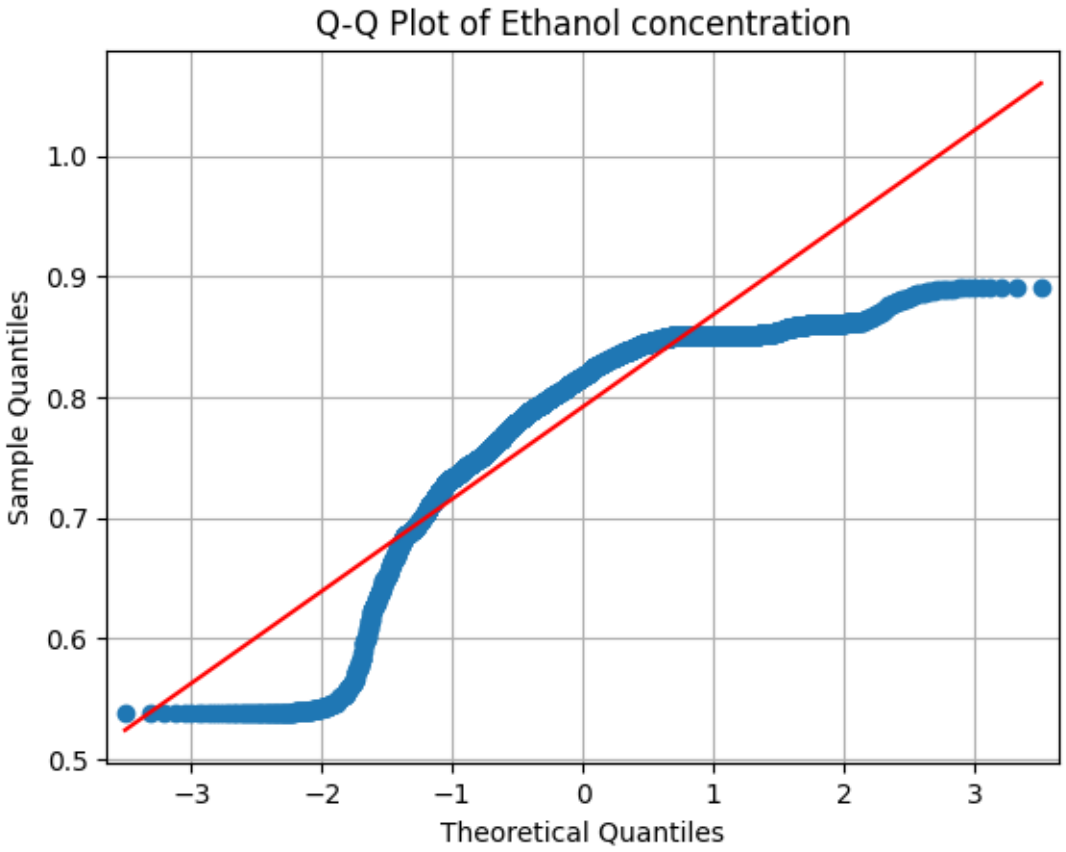
The input parameters in the dataset include crucial operating conditions such as column pressure (measured in bar), temperature at each tray (measured in Kelvin), liquid flowrate, vapor flowrate, distillate flowrate, bottoms flowrate, and feed flowrate (all measured in Kg

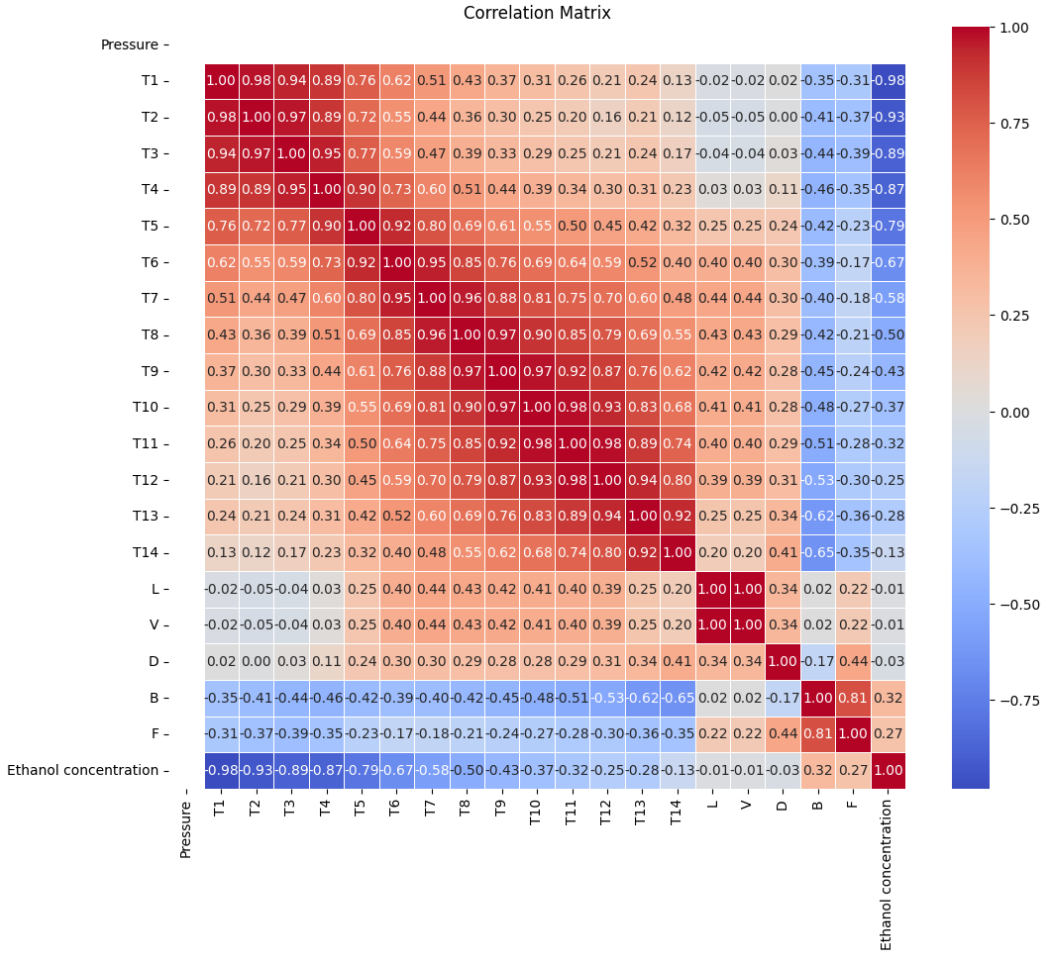
mol/hour). These variables play a pivotal role in distillation column performance and efficiency.

The output parameter of interest is the molar concentration of ethanol, which is essential for ethanol distillation applications in industries such as fuel or beverage production. The molar concentration of ethanol serves as a critical indicator of product quality and is a key variable for optimization in distillation column operations. The dataset's richness and relevance make it a valuable resource for developing and validating AI/ML models to enhance distillation column

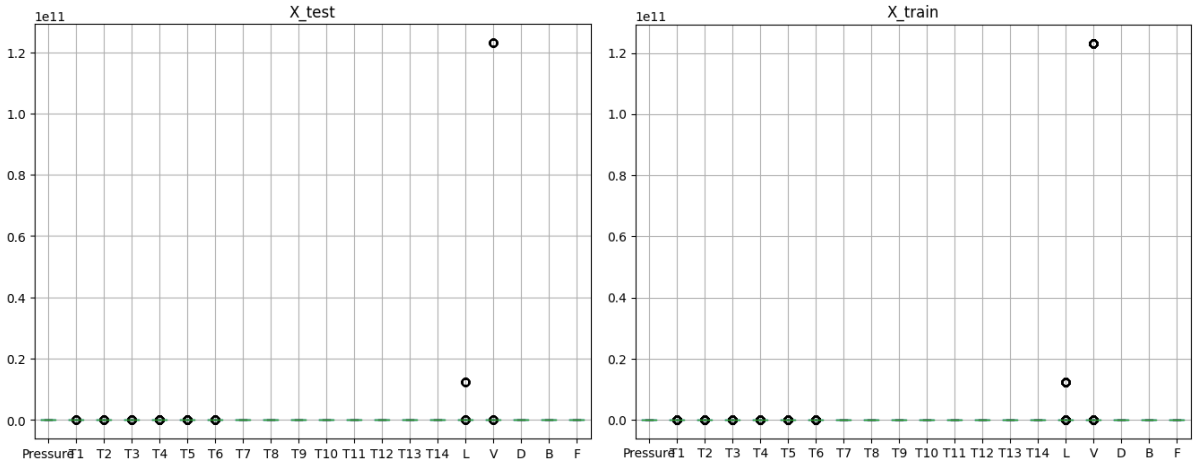
Data Preprocessing:

We initiated our data preprocessing by conducting Exploratory Data Analysis (EDA), during which we utilized techniques such as QQ plot and correlation matrix to analyze the trends in our data and identify the key factors influencing the output variable.

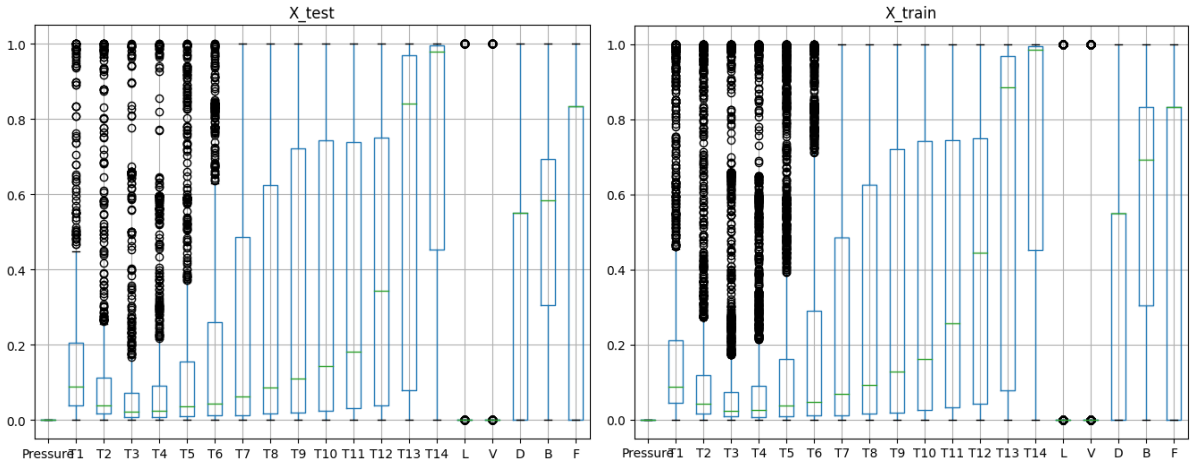




Then we normalized the input data using Min-Max scaling to mitigate scale differences across features. This involved transforming each feature to a common scale, ensuring that no single feature dominated the model training process based solely on its larger magnitude.

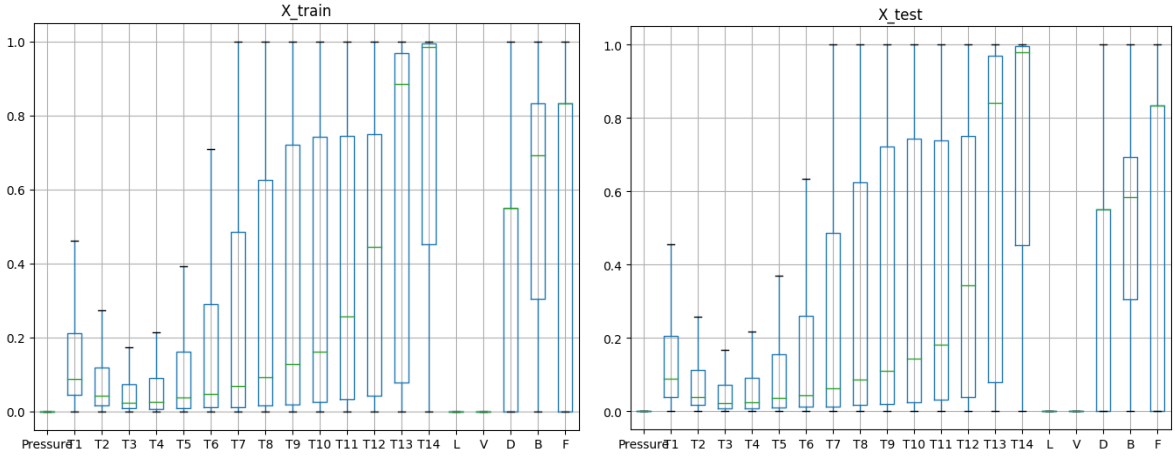


Box Plots before scaling of data



Box Plots after scaling of data

Following this, we addressed existing outliers to enhance the robustness of our machine learning models. Outliers can significantly impact model performance by skewing the data distribution, so we applied appropriate techniques to ensure our models were more resilient to extreme values.



Box Plots after outlier removal of data

We then proceeded to drop redundant columns and checked for null values to ensure the integrity and quality of our dataset. By removing redundant columns, we streamlined our data and focused only on the most relevant features for our modeling tasks. Given that the number of input features was relatively small and manageable, we determined that there would be no significant benefit in employing dimensionality reduction techniques. Dimensionality reduction methods like Principal Component Analysis (PCA) are typically useful when dealing with high-dimensional datasets to reduce computational complexity and improve model

generalization. However, in our case with a limited number of features, preserving all available information was deemed more advantageous for modeling accuracy and interpretability.

By implementing these preprocessing steps in this order, we prepared our dataset in a clean, standardized format suitable for training robust and accurate machine learning models tailored to simulate distillation column operation effectively under industrial conditions.

Model Architecture:

For the distillation column problem, we adopted a diverse ensemble of machine learning models, including Linear Regression (LR), Support Vector Regression (SVR), and Random Forest Regressor(RFR), for their unique capabilities in capturing different aspects of the complex relationships within distillation processes.

Linear Regression (LR): LR was chosen due to its suitability for modeling linear relationships between input features and the target variable. Distillation processes often exhibit straightforward dependencies among parameters such as flow rates, temperatures, and pressures, making LR a suitable choice for capturing these relationships. The transparency and interpretability offered by LR enable a clear understanding of the impact of each input variable on the predicted output.

Support Vector Regression (SVR): SVR was integrated into our model architecture to handle nonlinear relationships that may exist in complex distillation processes. SVR excels in capturing intricate patterns and can efficiently manage high-dimensional feature spaces. Its robustness to outliers and noise makes it well-suited for real-world industrial datasets, where data quality may be compromised. By leveraging SVR, we aimed to enhance prediction accuracy and model robustness in scenarios where linear assumptions may not hold.

Random Forest Regressor (RFR): The Random Forest Regressor was included for its ability to handle complex interactions and nonlinear relationships within the data. Random forests utilize ensemble learning techniques by combining predictions from multiple decision trees, resulting in improved accuracy and resilience to overfitting. This approach is particularly effective for capturing the diverse and intricate dynamics present in distillation column operation.

Reasoning for Model Selection: The selection of LR, SVR, and Random Forest Regressor was strategic to leverage the strengths of each model in addressing different facets of the distillation column simulation problem. This ensemble approach enabled us to account for both linear and nonlinear relationships, handle complex interactions within the data, and ensure robust performance in the presence of outliers and high-dimensional feature spaces. Our model architecture was designed to optimize performance, interpretability, and resilience to real-world challenges encountered in distillation column operations.

Tools and Technologies:

- Pandas & NumPy (Data Manipulation)
- Matplotlib (Visualisation)
- train_test_split, cross_val_score, cross_val_predict, KFold, GridSearchCV (from
- sklearn.model_selection) (Model Selection and Cross Validation)
- mean_absolute_error & r2_score (from sklearn.metrics)
- LinearRegression
- SupportVectorRegression
- Random Forest Regressoion

4 Implementation Plan

Development Phases:

The entire project was structured into eight distinct stages, including problem identification, dataset acquisition, exploratory data analysis, data preprocessing, model selection, model training, model evaluation, and report creation, culminating in final documentation. Each stage was carefully planned to ensure a systematic approach to solving the problem of simulating distillation column operation effectively.

Timeline for the project

Week 1-2: Problem Identification and Dataset Finding

Defined the specific problem statement and project objectives related to simulating distillation column operation.

Researched and identified suitable datasets containing essential parameters such as pressure, temperature, flow rates, and ethanol concentration for distillation column modeling.

Week 3-4: Exploratory Data Analysis (EDA)

Performed exploratory data analysis on the acquired dataset to understand its characteristics.
Visualized key features using histograms, scatter plots, and correlation matrices to identify patterns and relationships in the data.

Week 5-6: Data Preprocessing

Cleansed the dataset by handling missing values, duplicates, and inconsistencies.
Dropped redundant columns that did not contribute significantly to the modeling task.
Addressed outliers using appropriate techniques to enhance data quality and robustness.

Week 7-8: Model Selection

Chose suitable machine learning models for simulating distillation column performance (e.g., Linear Regression, Support Vector Regression, Random Forest Regressor).
Evaluated different models based on their ability to capture linear or nonlinear relationships inherent in distillation processes.

Week 9-10: Model Training and Optimization

Implemented selected machine learning models using the preprocessed dataset.
Trained models using the training dataset and fine-tuned hyperparameters to optimize performance.
Explored feature engineering techniques to enhance model accuracy and interpretability.

Week 11-12: Model Evaluation and Validation

Evaluated model performance using appropriate metrics (e.g., mean squared error, R-squared) on the testing dataset.
Validated model predictions against actual distillation column data to assess reliability.
Conducted sensitivity analysis and robustness testing to ensure model generalization.

Week 13-14: Final Documentation and Reporting

Compiled project findings, including EDA insights, data preprocessing steps, model selection criteria, and evaluation results.
Prepared a comprehensive report summarizing project methodologies, outcomes, and recommendations.
Created visual presentations or dashboards to effectively communicate project findings to stakeholders.

Model Training: In our approach to training the model for simulating distillation column operation, we adopted specific strategies tailored to each machine learning algorithm, encompassing parameter tuning and algorithm selection. Here's an overview of our strategies:

Train-Test Split:
We randomly divided the dataset into a 75:25 train-test split. The train split (75% of the data) was used for model training and parameter tuning, while the test split (25% of the data) remained untouched for final evaluation.

Support Vector Regression (SVR):
For SVR, we utilized the following parameters during model training:
Gamma (γ): Gamma is a parameter of the radial basis function (RBF) kernel in SVR. It defines the influence of a single training example, with low values indicating a large influence radius and high values indicating a narrow influence radius. In our training, we set gamma to 0.87, optimizing the balance between model complexity and generalization.

C (Regularization Parameter): The regularization parameter (C) in SVR controls the trade-off between achieving a low training error and minimizing model complexity to avoid overfitting. We set C to 0.01, favoring a simpler model to prevent excessive sensitivity to noise in the data.

Random Forest Regressor:
The Random Forest Regressor was trained using conventional parameters, leveraging its ability to handle complex interactions and nonlinear relationships within the dataset. This ensemble learning method combines predictions from multiple decision trees to improve accuracy and reduce overfitting.

Linear Regression (LR):

Similarly, Linear Regression was trained with standard parameters suitable for modeling linear relationships between input features and the target variable. LR provides transparency in understanding feature importance and impact on the predicted outcome.

Model Evaluation: In evaluating our machine learning models for simulating distillation column operation, we employed specific metrics and methods to assess performance and reliability. Here's an overview of the evaluation criteria:

Mean Absolute Error (MAE):

MAE measures the average magnitude of errors between predicted and actual values. It provides a straightforward interpretation of model accuracy, where lower MAE values indicate better performance. MAE is less sensitive to outliers compared to other metrics like MSE (Mean Squared Error).

$$\mathrm{MAE} = \frac{\sum_{i=1}^{N} |\hat{Y} - Y_i|}{N}$$

R-squared (R2) Score:

R2 score quantifies the proportion of variance in the target variable that is predictable from the input features. It ranges from 0 to 1, with higher values indicating a better fit of the model to the data. R2 score of 1 indicates a perfect fit, while a score of 0 suggests that the model does not explain the variability in the target variable.

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N} (Y_i - \underline{Y})^2}$$

5 Testing and Deployment

Testing Strategy: The model will be tested against unseen data using a 75:25 train-test split approach. After training the models on the training dataset (75% of the data), the trained models will be evaluated and tested on the unseen test dataset (25% of the data). This allows us to assess the generalization performance of the models on new, unseen data and ensures that the

models are capable of making accurate predictions outside of the training dataset context. The evaluation will involve calculating metrics such as Mean Absolute Error (MAE), R-squared (R2) score, and analyzing model predictions versus actual values to gauge model performance and reliability.

Deployment Strategy: For deploying our distillation column models in a real-world environment, we'll follow a comprehensive strategy to ensure seamless integration, user-friendly access, and ongoing maintenance and updates.

• Firstly, we'll integrate the trained models into the existing process control systems of the distillation column facility. This integration will involve developing APIs or microservices capable of receiving input data from the control system, executing predictions using the deployed models, and returning results for decision-making.

• To meet user interface needs, we'll develop a user-friendly interface enabling operators and engineers to interact with the deployed models. This interface will offer features like inputting new data, visualizing model predictions, and accessing diagnostic tools for performance evaluation. It will prioritize simplicity and efficiency to encourage easy adoption.

• Maintenance and updates are crucial for the continued effectiveness of the deployed models. We'll implement monitoring tools to track model performance and identify any anomalies or degradation in prediction accuracy. Furthermore, we will establish a schedule for periodic model retraining to incorporate new data and adapt to evolving process dynamics. Automated pipelines will streamline model retraining, validation, and deployment processes.

• Version control and documentation will maintain transparency and facilitate collaboration. Thorough documentation of the models, including architecture, training data, and performance metrics, will be kept. Version control systems will track changes and updates to ensure reproducibility and traceability of results.

By adhering to this deployment strategy, we aim to seamlessly integrate our distillation column models into the operational workflow, providing intuitive access to users, and ensuring their continued relevance and reliability over time.

Ethical Considerations: Deploying machine learning models in real-world applications, such as distillation column operations, raises important ethical considerations that must be addressed to ensure responsible and ethical use of technology. Here are some key ethical implications to consider:

- **Bias and Fairness:**

  Machine learning models can inadvertently perpetuate biases present in the training data, leading to unfair or discriminatory outcomes. It's crucial to assess and mitigate biases in the training data and model predictions to ensure fairness and equity in decision-making.

- **Transparency and Accountability:**

  Transparency in model development, deployment, and decision-making processes is essential for building trust with stakeholders. Clear documentation of model architecture, training data, and performance metrics promotes accountability and enables users to understand the limitations and potential biases of the deployed models.

- **Privacy and Data Protection:**

  Deploying models in operational environments involves handling sensitive data, such as process parameters or operational data. It's imperative to prioritize data privacy and implement robust security measures to protect confidential information from unauthorized access or misuse.

- **Human-in-the-Loop:**

  While machine learning models automate decision-making processes, human oversight and intervention remain critical. Implementing a "human-in-the-loop" approach ensures that final decisions consider ethical, legal, and practical implications beyond model predictions.

- **Algorithmic Transparency and Interpretability:**

  Enhancing the interpretability of machine learning models is crucial for understanding how decisions are made. Employing interpretable algorithms or post-hoc techniques to explain model predictions fosters trust and facilitates meaningful human-machine collaboration.

- **Ethical Use Cases and Impact Assessment:**

  Conducting comprehensive impact assessments to evaluate potential ethical risks and unintended consequences of deploying models is essential. Identifying and addressing ethical dilemmas early in the deployment process helps mitigate risks and ensure responsible use of technology.

- **Continuous Monitoring and Evaluation:**

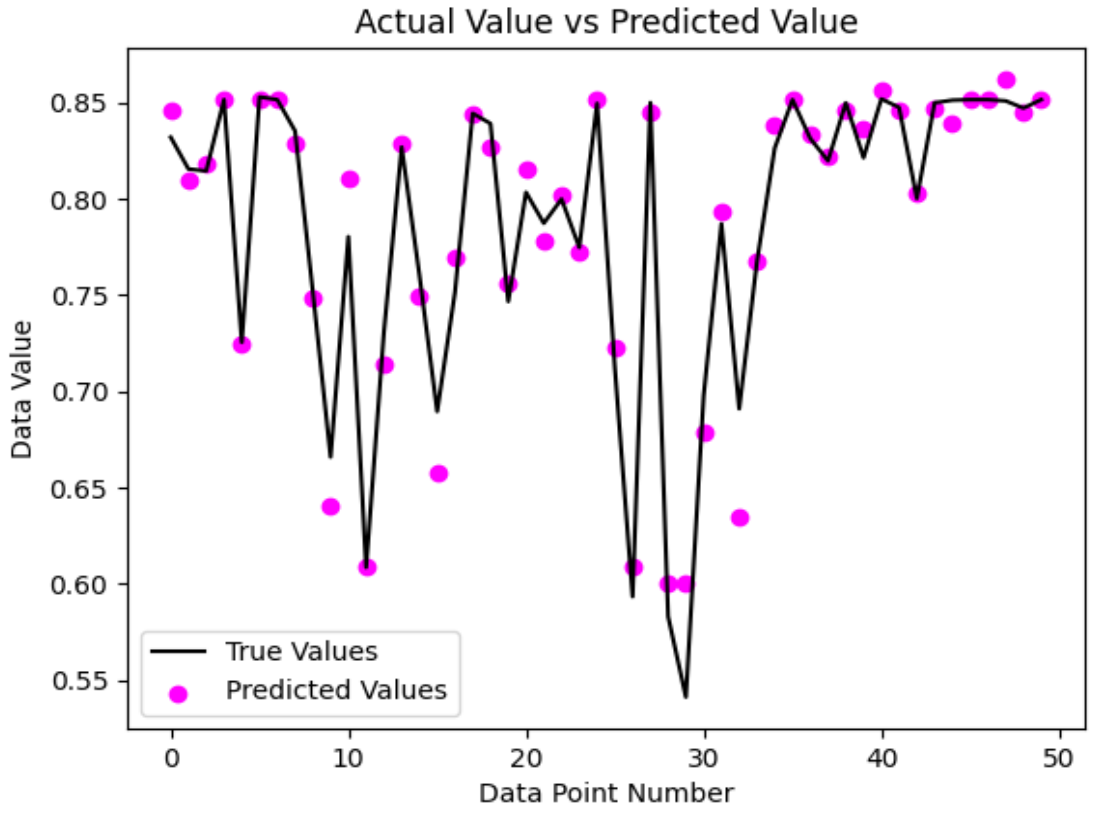
  Implementing ongoing monitoring and evaluation mechanisms to assess model performance and impact over time is essential. Proactively addressing ethical concerns and adapting to changing ethical standards and regulatory requirements fosters responsible model deployment and maintenance.

By addressing these ethical considerations proactively and integrating ethical principles into the deployment strategy, we aim to promote responsible innovation and ensure that our distillation column models contribute positively to operational efficiency while upholding ethical standards and societal values.

6 Results and Discussion:

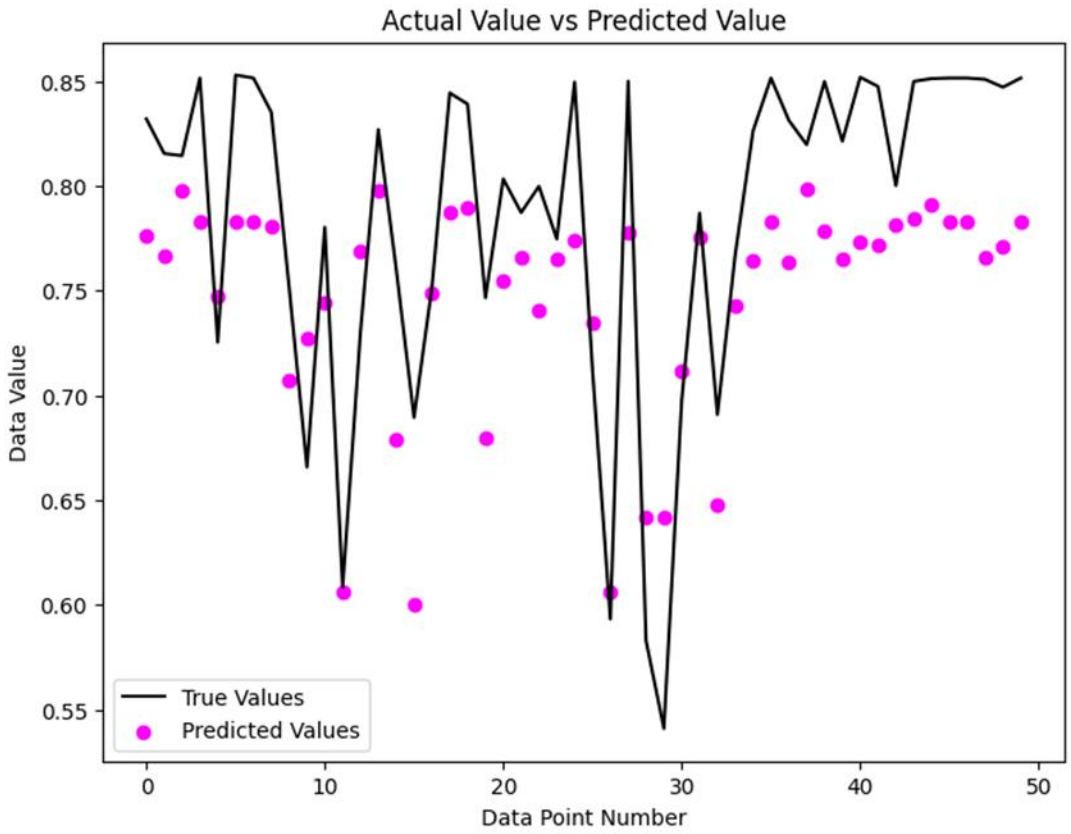
Based on the results of model training and evaluation, the mean absolute error (MAE) and R2 Score provide important insights into the performance of the Linear Regression (LR) and Support Vector Regression (SVR) models.

For the **LR model**, the **MAE** is measured at **0.0103**, indicating a relatively low average absolute error between the predicted and actual values. Additionally, the **R2 Score** for the **LR model** is notably high at **0.9499**, suggesting that the LR model explains approximately **94.99%** of the variance in the target variable, showcasing a strong fit to the data.



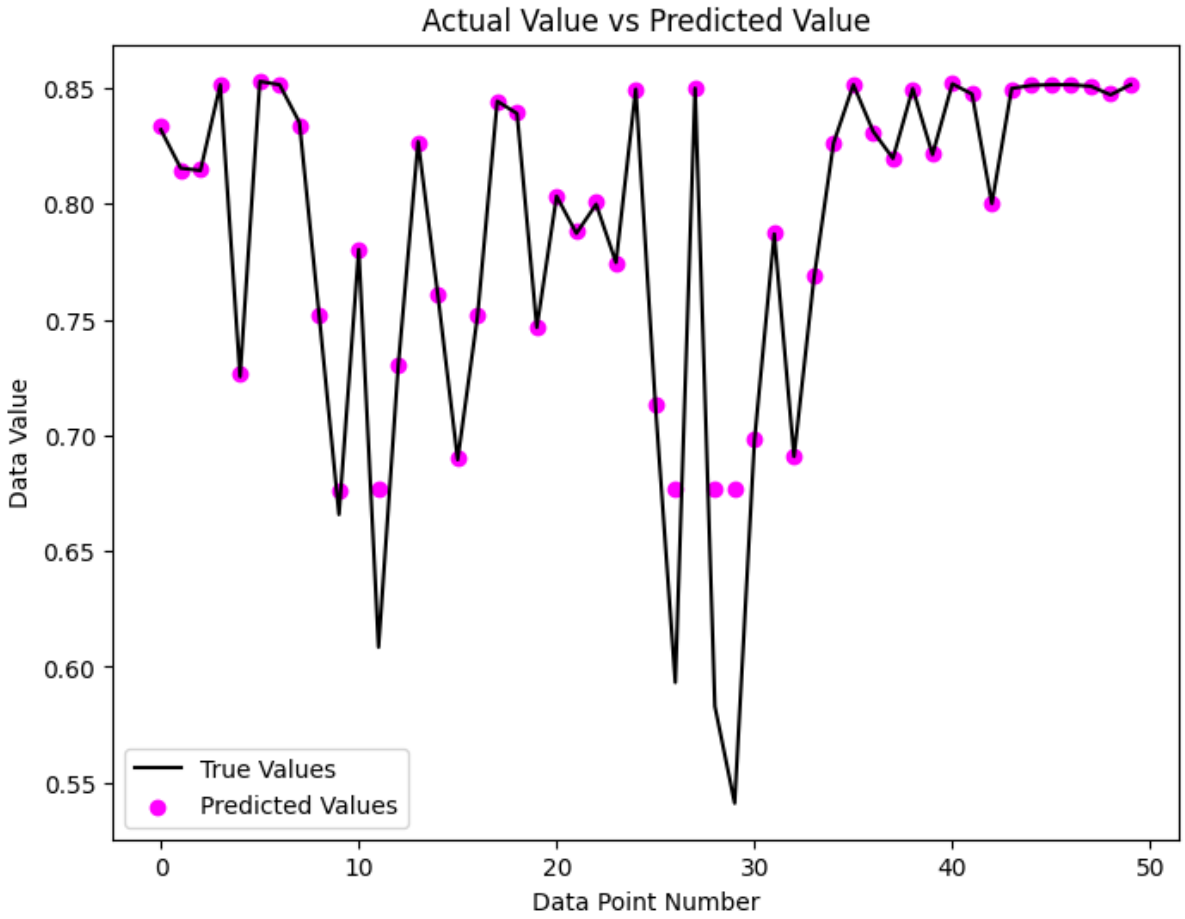
Linear regression model

On the other hand, the **SVR model** exhibits a higher **MAE** of **0.0585**, signifying a larger average absolute error compared to the LR model. Furthermore, the **R^2 Score** for the SVR model is substantially lower at **0.2968**, indicating that the SVR model explains only about **29.68%** of the variance in the target variable, implying a weaker fit to the data.



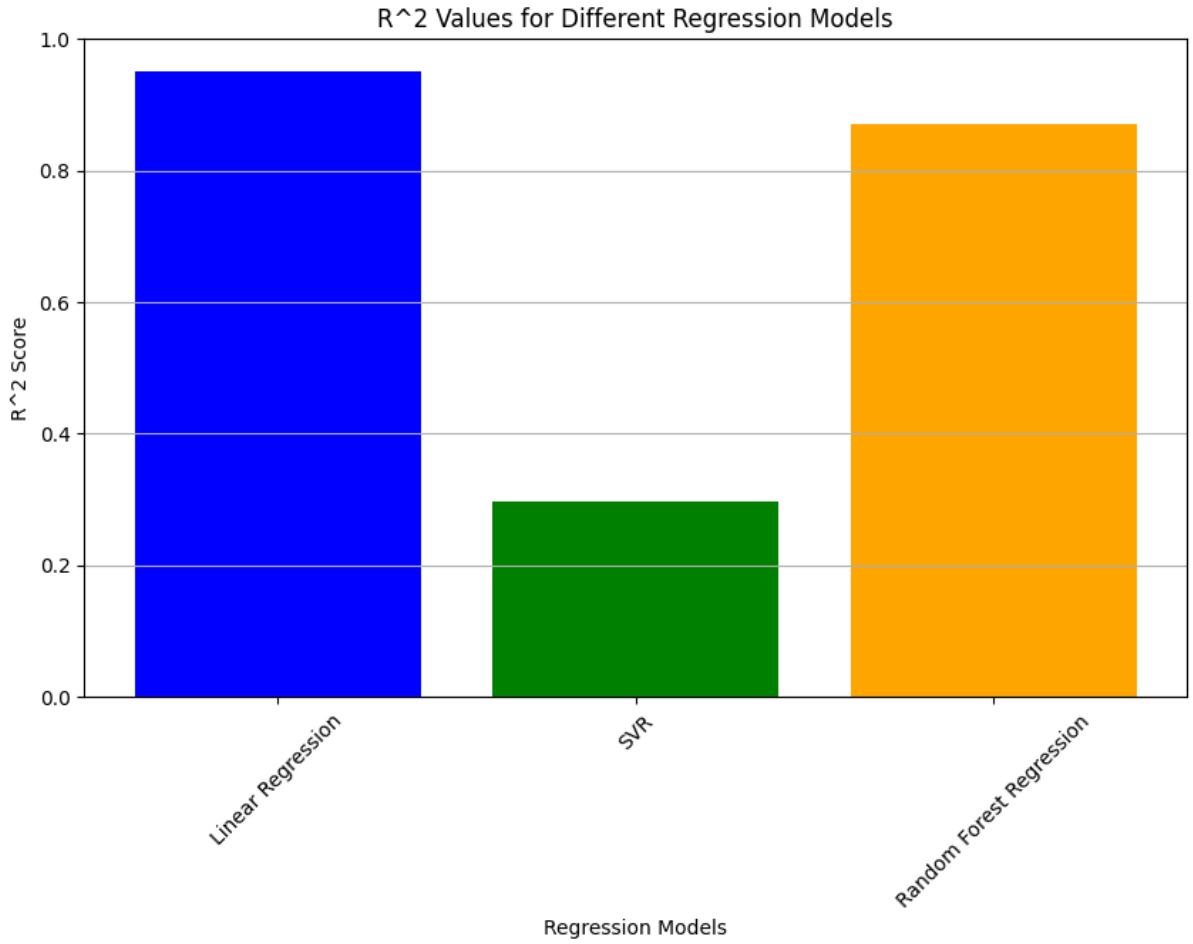
SVR model

The **Random Forest Regressor** yielded promising results with a **Mean Absolute Error (MAE) of 0.0068** and an **R-squared (R2) score of 0.871**, suggesting that the **RFR** model explains approximately **87.1%** of the variance in the target variable, showcasing a strong fit to the data.



RFR model

Findings: The evaluation of different regression models for distillation column simulation reveals notable differences in their predictive performance. The Linear Regression (LR) model demonstrates exceptional accuracy, indicating precise predictions and a strong fit to the data. Its high R-squared (R2) score signifies a robust ability to explain the variance in the target variable. In contrast, the Support Vector Regression (SVR) model shows higher error metrics and a lower R2 Score, indicating poorer performance and weaker explanatory power. The Random Forest Regressor (RFR) showcases impressive results, demonstrating its ability to explain a substantial portion of the variance in the target variable. These insights highlight the effectiveness of LR and RFR models for distillation column simulation, suggesting their suitability for practical implementation in industrial settings.



Challenges and Limitations: During the course of this project, we encountered several challenges related to data preprocessing that influenced the overall approach and outcomes. One notable challenge was handling diverse data types and formats within the provided dataset. This required careful consideration of how different data types (e.g., numerical, categorical) were processed and transformed to ensure compatibility with machine learning algorithms.

Additionally, we faced issues with data represented in exponential forms, which required special handling to prevent numerical instability and maintain data integrity during preprocessing. Transforming exponential data into a suitable format for modeling involved techniques such as logarithmic scaling or custom feature engineering.

Despite our efforts to address these challenges, there are inherent limitations to our proposed solution. Firstly, the reliance on traditional regression models like Linear Regression and Support Vector Regression may restrict the model's ability to capture complex nonlinear relationships inherent in distillation column processes. Furthermore, while Random Forest Regressor showed promising results, it may still encounter challenges with overfitting or scalability in larger datasets or dynamic operational environments.

Moreover, the preprocessing steps undertaken, while essential for data quality and model performance, may not fully account for all variations and complexities present in real-world distillation operations. Future iterations of this project could benefit from exploring advanced modeling techniques, incorporating domain-specific knowledge, and continuously refining data preprocessing strategies to enhance the robustness and applicability of the predictive models.

In summary, while we successfully addressed immediate challenges during data preprocessing and model development, the project's scope and methodology have inherent limitations that underscore the need for ongoing refinement and adaptation to effectively address the complexities of distillation column simulation and optimization.

7 Conclusion and Future Work

The project focuses on optimizing distillation column operations using AI/ML techniques, with a specific emphasis on predicting the molar concentration of ethanol in the distillate. Through rigorous data preprocessing, model selection, and deployment strategies, our work aims to enhance process efficiency, reduce energy consumption, and minimize waste generation in distillation processes.

The impact of our project extends across various industries reliant on distillation, including chemical manufacturing, petroleum refining, and pharmaceutical production. By implementing our model, manufacturers can achieve increased operational efficiency, cost savings, and environmental sustainability through optimized process control and decision-making.

Looking ahead, future research directions could explore advanced modeling techniques, such as deep learning and reinforcement learning, to capture complex nonlinear relationships inherent in distillation processes. Additionally, incorporating real-time sensor data and

feedback mechanisms could enhance model accuracy and adaptability to dynamic operational conditions.

Furthermore, expanding the scope to include multi-objective optimization, considering factors like economic feasibility and environmental impact, could provide a holistic approach to distillation column optimization. Collaboration with industry partners and academia could facilitate the integration of AI/ML technologies into existing process control systems, driving continuous innovation and improvement in industrial distillation practices.

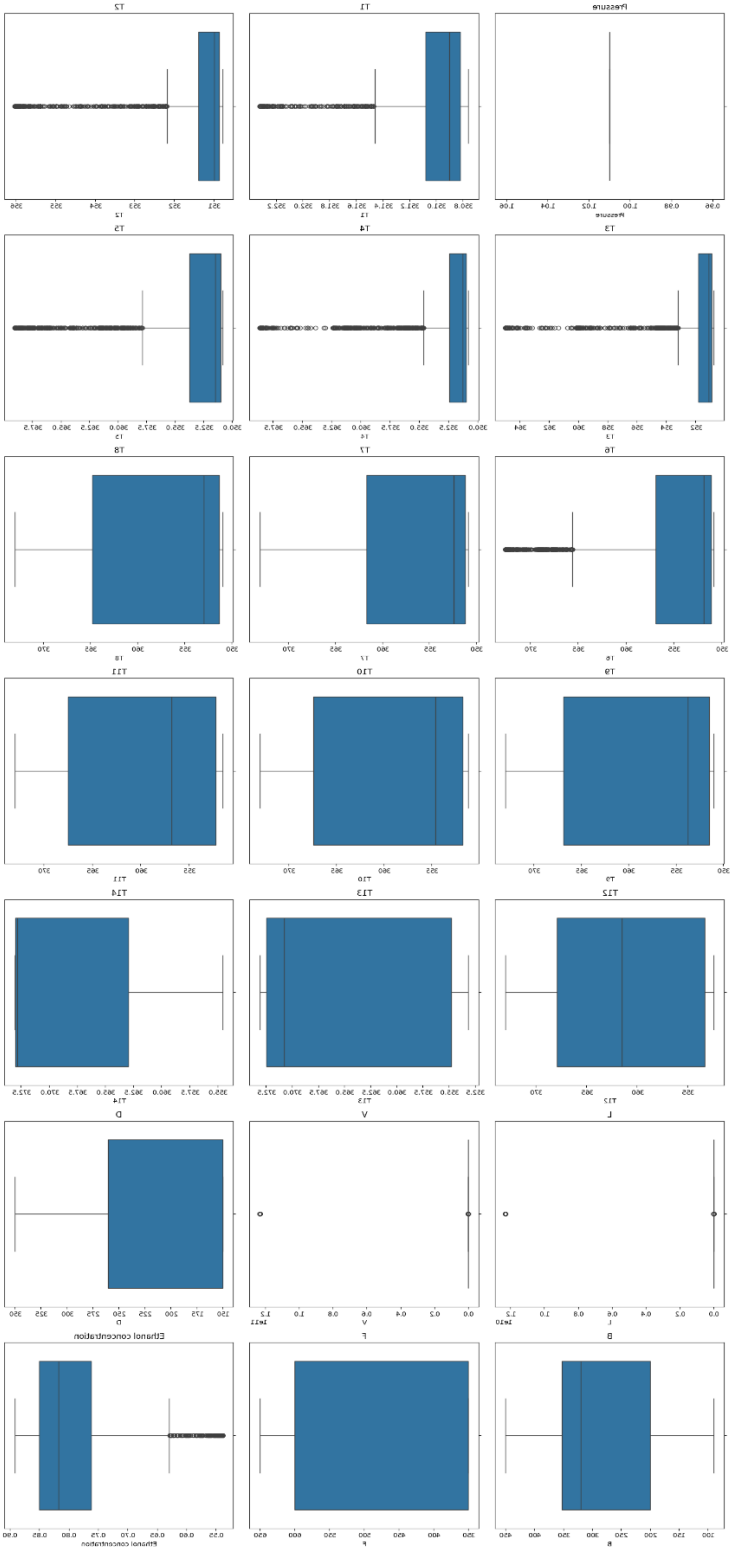
In conclusion, our project exemplifies the transformative potential of AI/ML in chemical engineering, highlighting its role in driving sustainable and cost-effective manufacturing practices. By leveraging data-driven insights and innovative technologies, we aspire to pave the way for future research and development in distillation optimization, ultimately contributing to a more efficient and environmentally conscious industrial landscape.

8 References

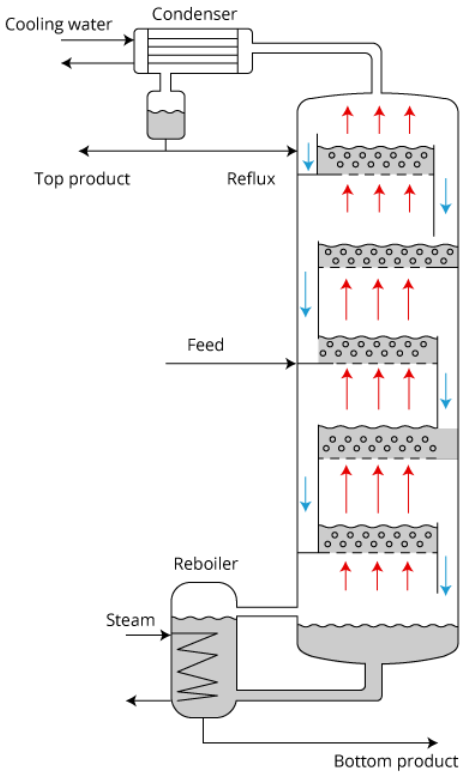
The data for the project has been sourced from Kaggle https://www.kaggle.com/datasets/jorgecote/distillation-column

Cote-Ballesteros, J. E., Grisales Palacios, V. H., & Rodriguez-Castellanos, J. E. (2022). Un algoritmo de selección de variables de enfoque híbrido basado en información mutua para aplicaciones de sensores blandos industriales basados en datos. Ciencia E Ingeniería Neogranadina, 32(1), 59–70. https://doi.org/10.18359/rcin.5644

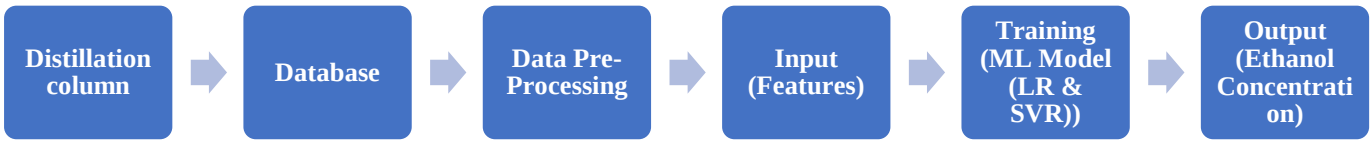
9 Appendices



Box plot of each input variable showing no. of outliers in each of the features.



Process Diagram of a Distillation Column



Flow sheet showing the approach used for model development

10 Auxiliaries

Data Source: [Raw Data]

Python file: [Collab Link]