Machine Learning-Based Distillation Tray Prediction

Goutam Kumar Hembram

210107030

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

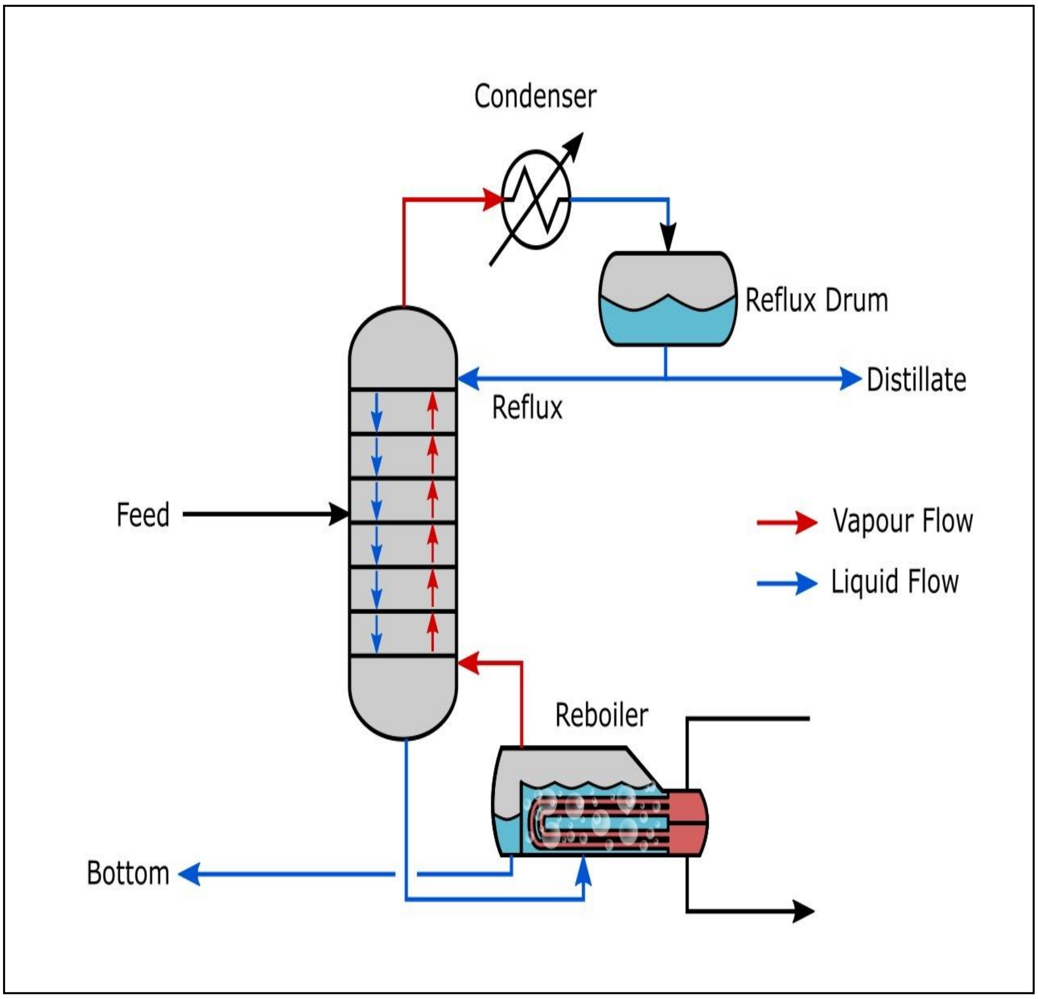
Distillation columns are vital workhorses in the chemical industry, separating components in liquid mixtures. A key design parameter is the number of trays within the column, impacting efficiency and cost. Traditionally, this hinges on complex calculations and pilot studies. Machine learning (ML) offers a promising alternative**.**

.

2 Introduction

Background: Distillation is a fundamental process in chemical engineering used to separate components of a liquid mixture based on differences in volatility. Distillation columns are complex systems comprising multiple trays or plates, where each tray facilitates the separation of components by vaporization and condensation. The efficiency and cost-effectiveness of distillation columns heavily depend on the design parameters, including the number of trays.

Problem Statement: Determining the optimal number of trays in a distillation column is crucial for efficient separation. However, current methods rely on laborious calculations or pilot testing, which can be time-consuming and expensive. Machine learning can revolutionize this process. By analyzing data from existing columns and their performance, ML models can predict the required number of trays for new separation tasks.



<u>Objectives</u>: The primary objectives of a project using an AI/ML model for distillation tray prediction can be broken down into two key areas:

1. **Prediction Accuracy:**
   - Develop a model that accurately predicts the optimal number of trays required for a specific distillation task.
   - Quantify the model's prediction accuracy using relevant metrics (e.g., Mean Squared Error, percentage deviation from actual tray number).
   - Aim for a model that reduces the need for extensive pilot testing by providing reliable predictions.

2. **Process Optimization:**
   - Train the model to consider various operating conditions like feed composition, desired product purity, and flowrates.

    - Enable the model to recommend the optimal number of trays based on these operational factors, leading to efficient separation.
    - Ultimately, the model should contribute to designing distillation columns that achieve desired separation with minimal energy consumption and equipment size.

3 Methodology

Data Source:

Historical Data:

- **Source:** Plant historian or process information management system (PIMS)
- **Data Access:** The project will require access to historical data from past operations of the distillation column. This data may include:
    - Operating conditions like feed composition, desired product purities, and flow rates for various separation tasks performed in the past.
    - The actual number of trays used in the column for each historical operation.

Here we use This Data Set which was available in Github:Dataset

Tools and Technologies:

Programming Languages

- **Python**: Python is the most popular programming language for machine learning due to its simplicity, versatility, and vast ecosystem of libraries. Key libraries for ML in Python include:
    - **NumPy**: For numerical operations.
    - **Pandas**: For data manipulation and analysis.

- **scikit-learn**: For machine learning algorithms and tools.
- **TensorFlow** or **PyTorch**: For deep learning and neural networks.

Machine Learning Libraries:

- **scikit-learn**: Python library providing simple and efficient tools for data mining and data analysis, including various supervised and unsupervised learning algorithms.

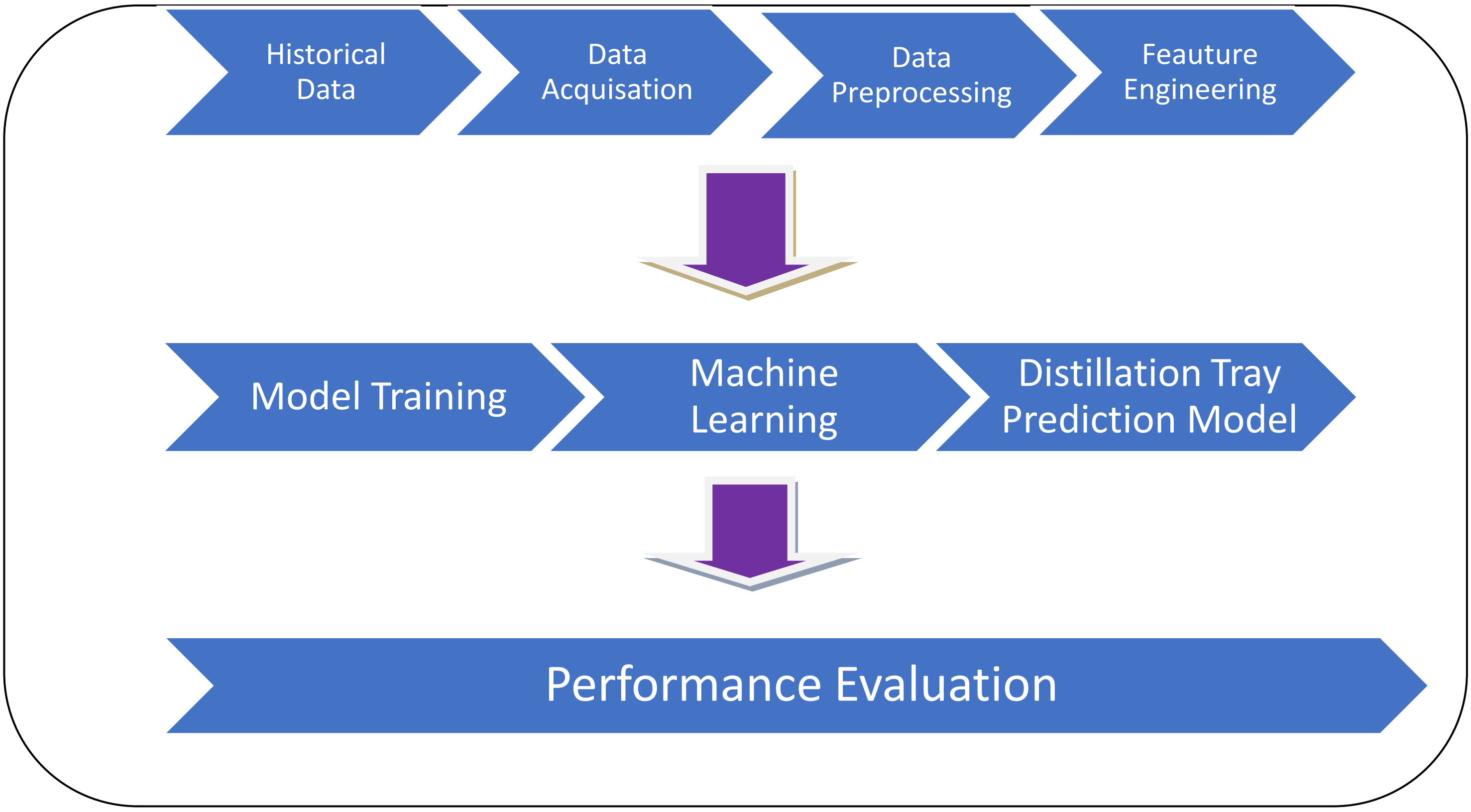
Data Visualization Tools:

- **Matplotlib** and **Seaborn**: Python libraries for creating static, animated, and interactive visualizations.

Google Colab (short for Google Colaboratory) is a cloud-based platform provided by Google that allows us to write, run, and share Python code directly in the browser.

4 Implementation Plan

Development Phases:



(1).Historical data: Past operational data from the PIMS, encompassing numerous separation tasks with their corresponding operating conditions and tray numbers, will further contribute to the high data volume.

(2).Data Acquisation: The project will require access to historical data from past operations of the distillation column.

(3).Data Preprocessing: Check for missing values in each column, Remove rows with any missing values, Fill missing values with a specific value, Replace NaN with 0, mpute missing values with mean, median, or mode.

(4).Feature Selection:Most Important features ere selected and taking as a input.

(5).Model Training:

- This dataset is used to train the Machine Learning model.
- The trained model then takes new feed composition, desired product purity, and flow rate data as input and predicts the optimal number of trays needed for the distillation column.

(6).Performance Evaluation: Finally, the model's performance is evaluated using various metrics to ensure its accuracy and reliability.

5 Testing and Deployment

Testing Strategy:

1. **Train-Validation-Test Split**:
    - **Data Partitioning**: Split your dataset into three subsets: training data, validation data, and test data. The training data is used to train the model, the validation data is used to tune hyperparameters and

monitor performance during training, and the test data is used for final evaluation of the trained model.

2. **Evaluation Metrics**:

    - **Select Appropriate Metrics**: Choose evaluation metrics that are relevant to the problem at hand. For distillation tray prediction, metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), or coefficient of determination (R^2 score) can be used to quantify the accuracy and performance of the model.

Testing on Unseen Data:

- **Use Test Data**: After finalizing the model and hyperparameters using the training and validation data, evaluate the model performance on the test data which the model has not seen during training or tuning. This step provides an unbiased estimate of the model's performance in real-world scenarios.

Deployment Strategy and Ethical Considerations :

Deploying a machine learning-based distillation tray prediction model involves making the model accessible and operational in a production environment where it can generate predictions based on new input data. Here's a structured deployment strategy for this purpose:

1. Model Serialization:

- Serialize the trained model into a format that can be easily saved and loaded, such as a `joblib` or `pickle` file in Python. This serialized model will be deployed and used for making predictions.

2. API Development:

- Develop an API (Application Programming Interface) that exposes endpoints for receiving input data and returning predictions from the deployed model. This API can be implemented using frameworks like Flask or FastAPI in Python.

3. Input Data Handling:

- Define the format and structure of the input data expected by the model. Implement data validation and preprocessing steps within the API to ensure that incoming data is formatted correctly and is compatible with the model's requirements.

4. Scalability and Performance:

- Consider scalability and performance requirements when deploying the API. Use containerization technologies like Docker to create lightweight, scalable containers for hosting the API. Deploy the containers on cloud platforms like AWS, Google Cloud, or Azure to benefit from auto-scaling and high availability features.

5. Monitoring and Logging:

- Implement logging and monitoring functionalities within the deployed API to track incoming requests, response times, error rates, and other relevant metrics. Use tools like Prometheus, Grafana, or cloud provider monitoring services to monitor the health and performance of the deployed model.

6. Security and Authentication:

- Implement security measures to protect the deployed API from unauthorized access and potential attacks. Use authentication mechanisms such as API keys, OAuth tokens, or JWT (JSON Web Tokens) to authenticate and authorize users accessing the API.

7. Versioning and Maintenance:

- Implement versioning of the API to support future updates and improvements to the model without disrupting existing applications relying on the API. Use semantic versioning (e.g., v1, v2) to manage different versions of the deployed model and API.

8. Testing in Production:

- Conduct thorough testing of the deployed model and API in a production-like environment before releasing it to end-users. Perform integration tests, load tests, and functional tests to ensure that the deployed system meets performance and reliability requirements.

9. Documentation and Support:

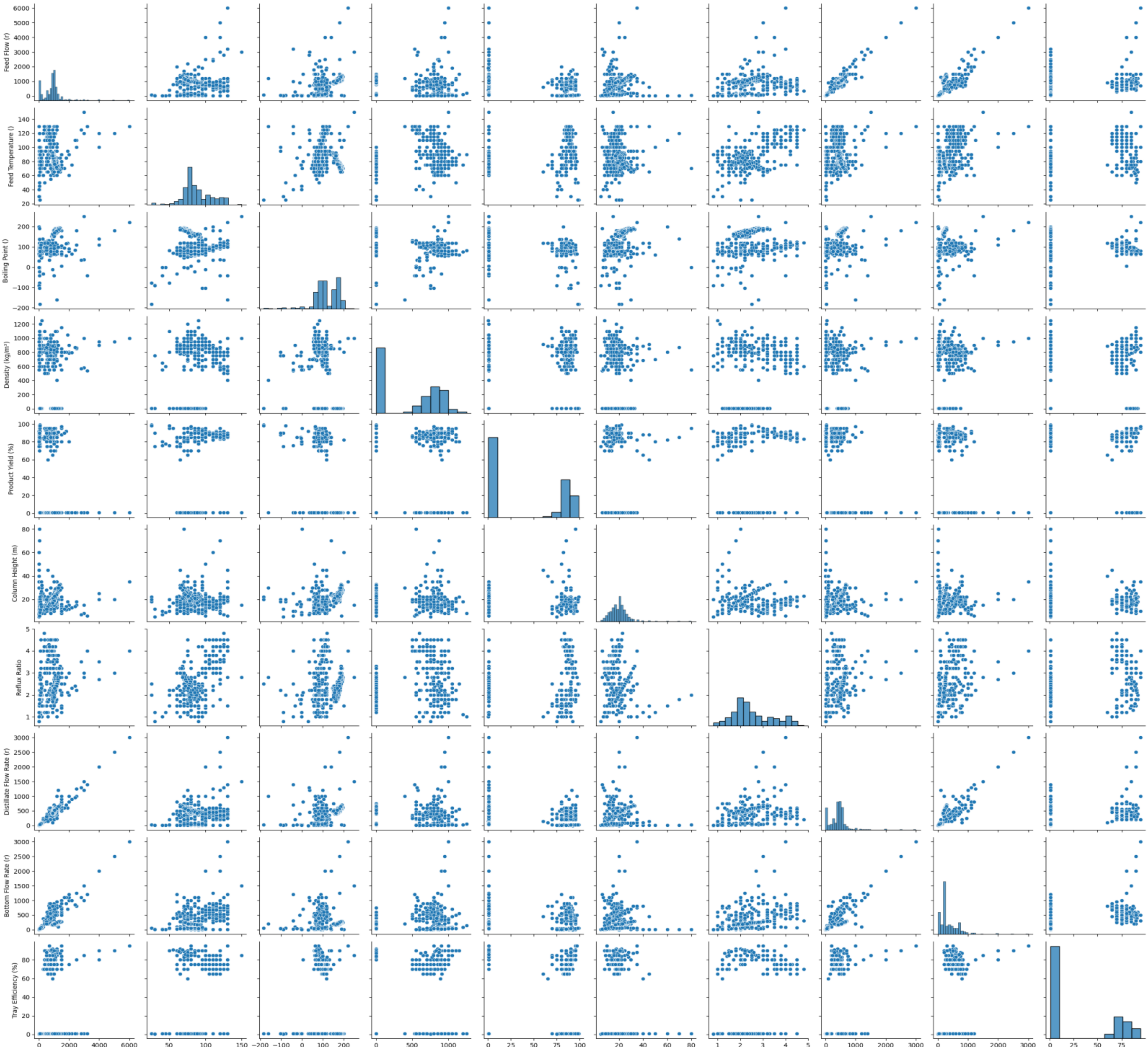
- Provide comprehensive documentation for the deployed API, including usage instructions, input data specifications, and example requests/responses. Offer support channels (e.g., email, Slack, forums) for users to report issues and seek assistance with using the deployed model.

10. Continuous Improvement:

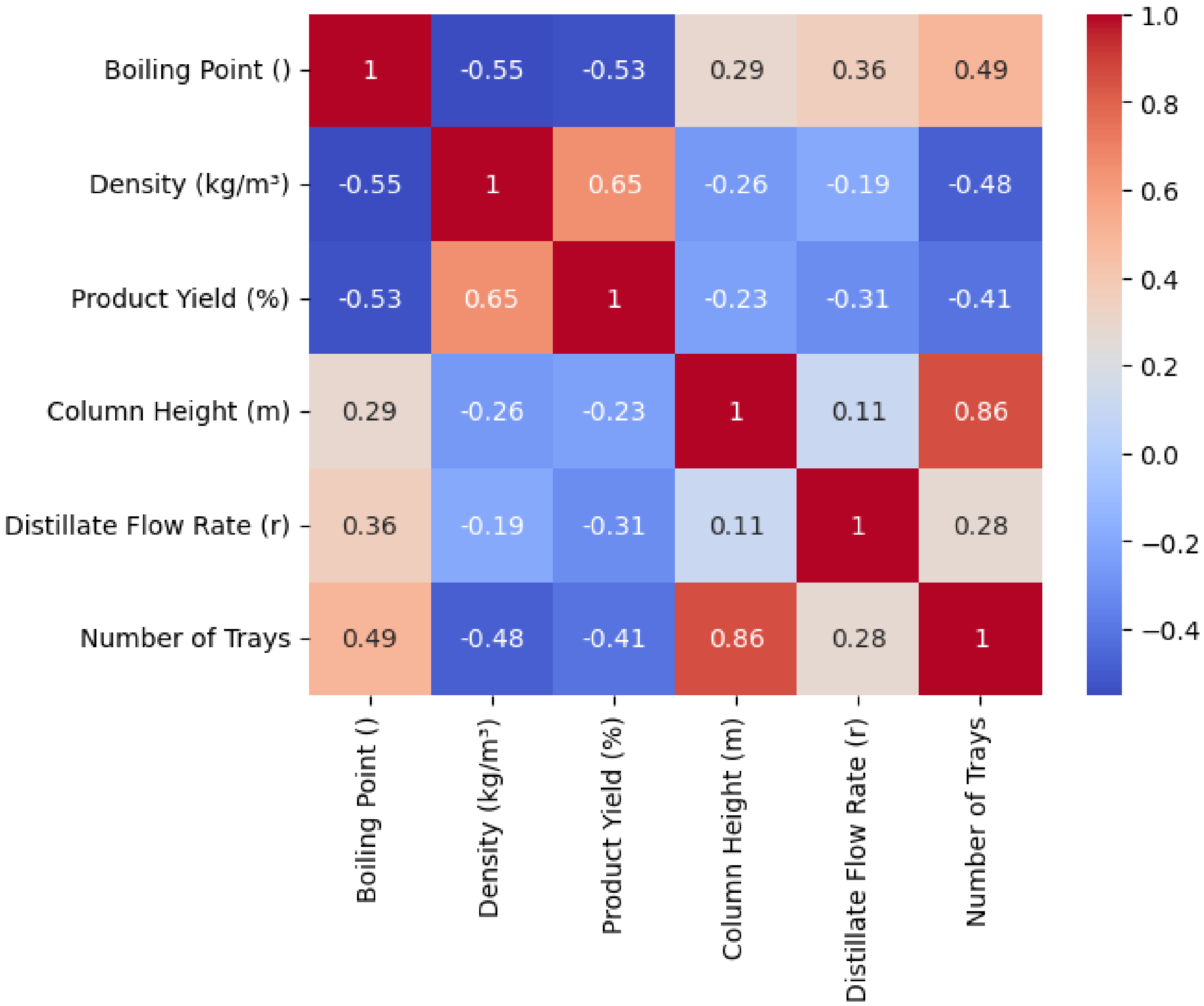
- Continuously monitor and gather feedback from users and stakeholders to identify areas for improvement in the deployed model and API. Use this feedback to iterate on the model, improve prediction accuracy, and enhance the overall user experience.

6 Results and Discussion

- From Explanatory Data Anaysis:



- Correlations matrix



- Linear Regression:
  - Mean Squared Error: 12.660882352941176
  - Mean Absolute Error: 2.3485294117647055
  - R-squared: 0.8075548878903129

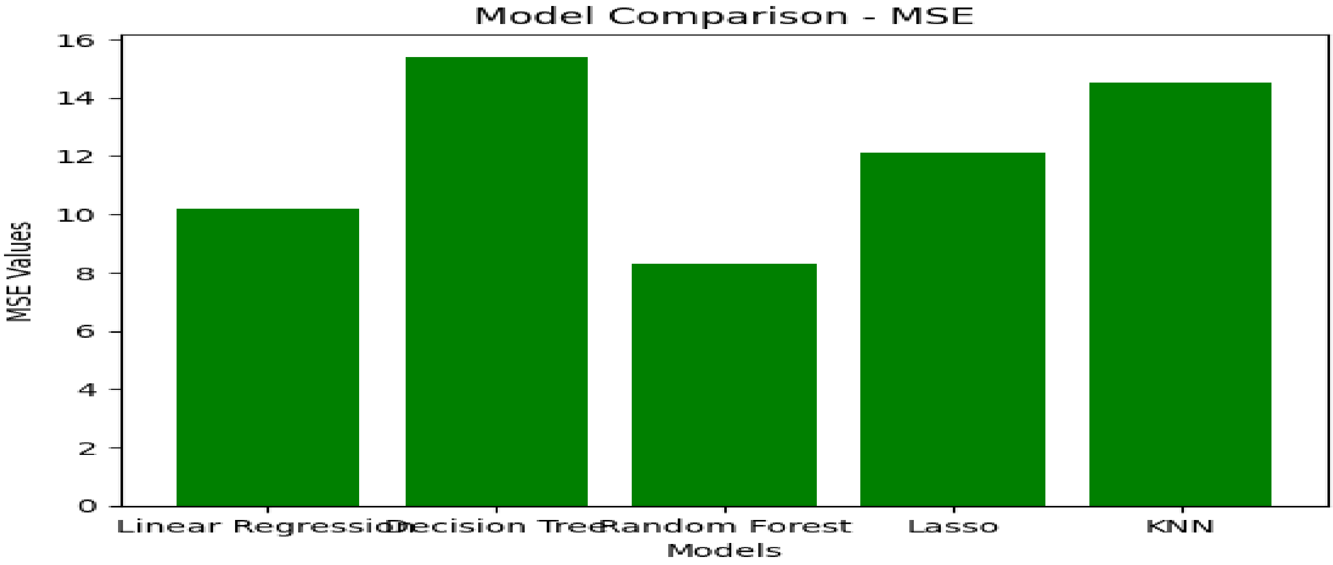
- DECISION TREE REGRESSION:
    - Mean Squared Error: 9.432032227877597
    - Mean Absolute Error: 2.166144096751604
    - R-squared: 0.8566333333715546

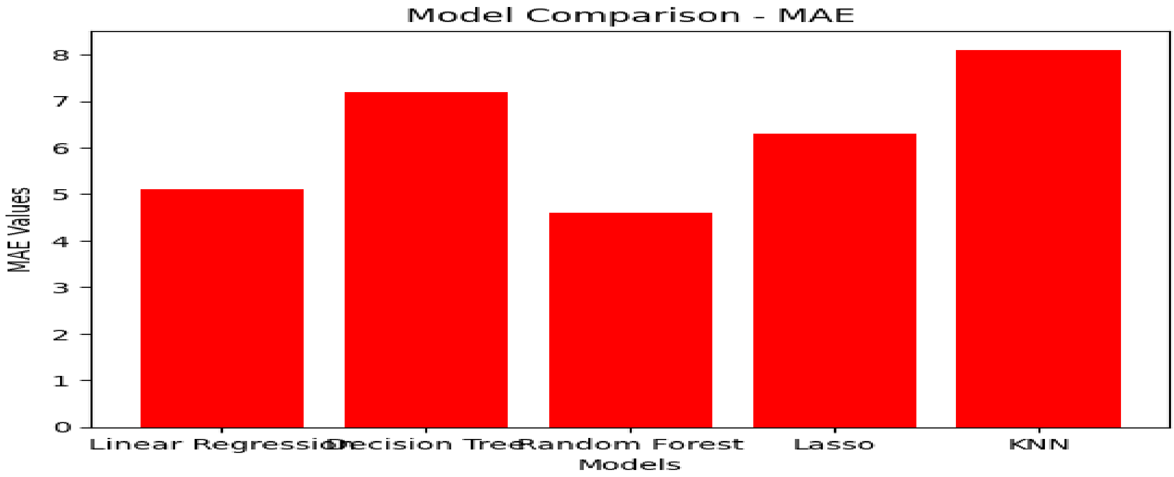
- RANDOM FOREST REGRESSION:
    - Mean Squared Error: 9.87585661764706
    - Mean Absolute Error: 1.877720588235294
    - R-squared: 0.8498872131513981

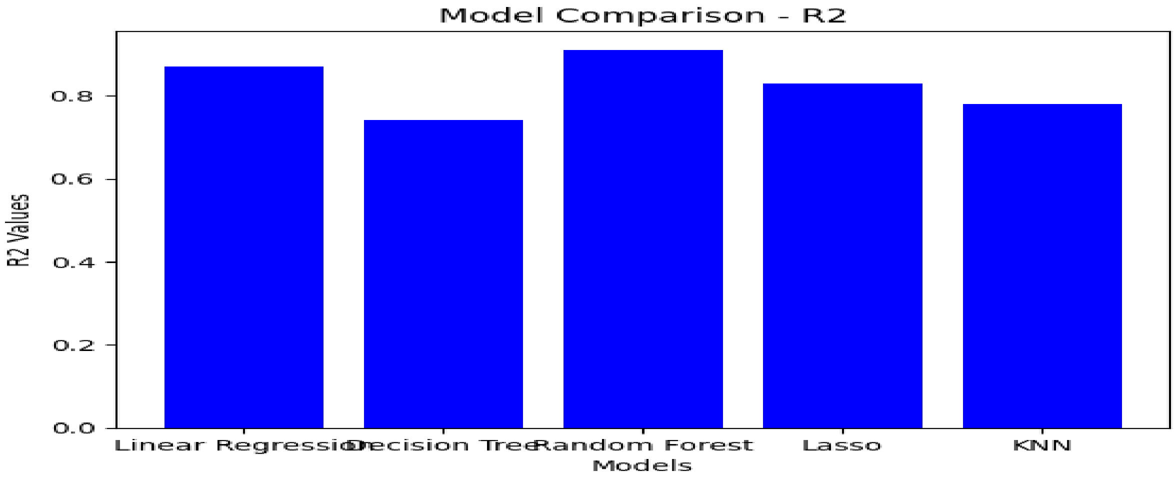
- LASSO REGRESSION:
    - Best alpha: 0.1
    - Mean Squared Error: 9.629348302038522
    - Mean Absolute Error: 2.2557188138410247
    - R-squared: 0.8536341337143426

- KNN REGRESSION:
    - Mean Squared Error: 12.660882352941176
    - Mean Absolute Error: 2.3485294117647055
    - R-squared: 0.8075548878903129

- Comparisions:


Model Comparison - MSE
MSE Values
0
2
4
6
8
10
12
14
16
Linear Regression
Decision Tree
Random Forest
Lasso
KNN
Models


Model Comparison - MAE
MAE Values
0
1
2
3
4
5
6
7
8
Linear Regression
Decision Tree
Random Forest
Lasso
KNN
Models


Model Comparison - R2
R2 Values
0.0
0.2
0.4
0.6
0.8
Linear Regression
Decision Tree
Random Forest
Lasso
KNN
Models

Challenges and Limitations:

Challenges and limitations in a machine learning-based distillation tray prediction project include:

1. **Data Quality and Availability**:
    - Limited availability of high-quality data for training the model, especially for niche or proprietary distillation processes. Noisy or incomplete data can lead to biased or inaccurate predictions.
2. **Complexity of Distillation Processes**:
    - Distillation processes can be highly complex and nonlinear, making it challenging to capture all relevant factors and relationships using machine learning models.
3. **Feature Engineering**:
    - Feature engineering requires domain expertise to identify and extract meaningful features from raw data. Inadequate feature selection or engineering can impact model performance.
4. **Model Interpretability**:
    - Some machine learning models, particularly deep learning models, lack interpretability, making it difficult to understand and explain the rationale behind predictions, which is crucial in critical applications like distillation.
5. **Generalization to New Conditions**:
    - Models trained on historical data may struggle to generalize to new operating conditions or scenarios not seen during training, leading to poor performance in real-world applications.
6. **Deployment and Integration**:

- Integrating machine learning models into existing distillation systems or workflows can be challenging due to compatibility issues, latency concerns, and the need for continuous monitoring and maintenance.

7. **Performance Metrics**:

   - Selecting appropriate evaluation metrics for assessing model performance in distillation tray prediction, considering the impact of false positives/negatives on operational decisions.

8. **Regulatory and Safety Compliance**:

   - Ensuring that machine learning models comply with industry regulations and safety standards in the chemical processing domain, where accuracy and reliability are critical.

7 Conclusion and Future Work

Summary of the Project:

The project focuses on developing a machine learning-based approach for predicting the optimal number of distillation trays in chemical distillation columns. This involves leveraging historical data on process variables and tray configurations to train predictive models. Key steps include data preprocessing, feature engineering, model training, and validation. The goal is to create a reliable tool that can assist engineers in designing efficient distillation systems with reduced costs and improved performance.

Impact of the Project:

The successful implementation of machine learning for distillation tray prediction has several potential impacts:

1. **Cost Reduction:** By accurately predicting the optimal number of trays, engineers can design more efficient distillation columns, leading to cost savings in terms of energy consumption and equipment expenses.
2. **Improved Efficiency:** Optimized tray design can enhance separation efficiency, reducing the need for reprocessing and improving overall productivity.
3. **Time Savings:** Machine learning models can expedite the design process by automating complex calculations and reducing the reliance on empirical methods and pilot studies.
4. **Enhanced Decision-Making:** Predictive models provide valuable insights into process variables and tray configurations, enabling informed decision-making during distillation column design and operation.

Potential Future Directions for Further Research:

To advance the field of machine learning-based distillation tray prediction, future research can explore the following directions:

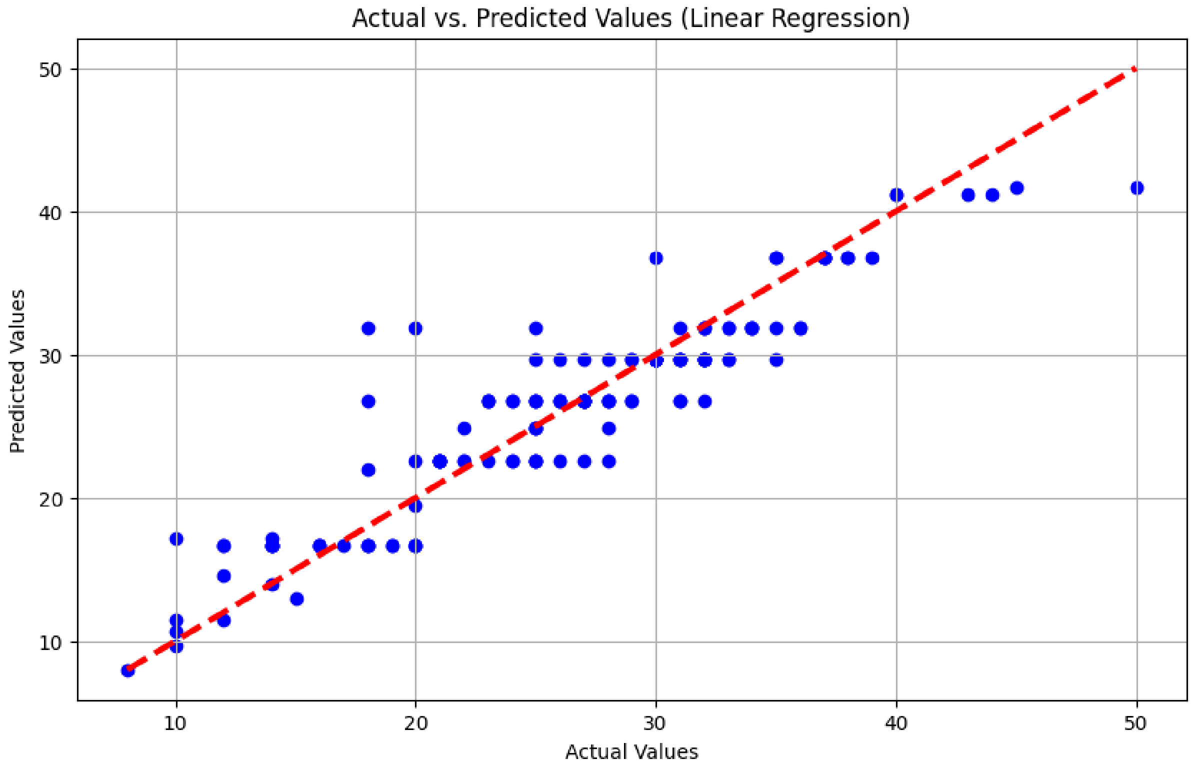
1. **Model Interpretability:** Develop interpretable machine learning models that provide insights into the underlying factors influencing tray design decisions, aiding engineers in understanding and trusting the predictions.
2. **Incorporating Dynamic Conditions:** Extend the predictive models to account for dynamic changes in feed compositions, operating conditions, and environmental factors to enhance robustness and generalization.
3. **Ensemble Methods:** Investigate the use of ensemble learning techniques to combine predictions from multiple models, potentially improving accuracy and reliability.

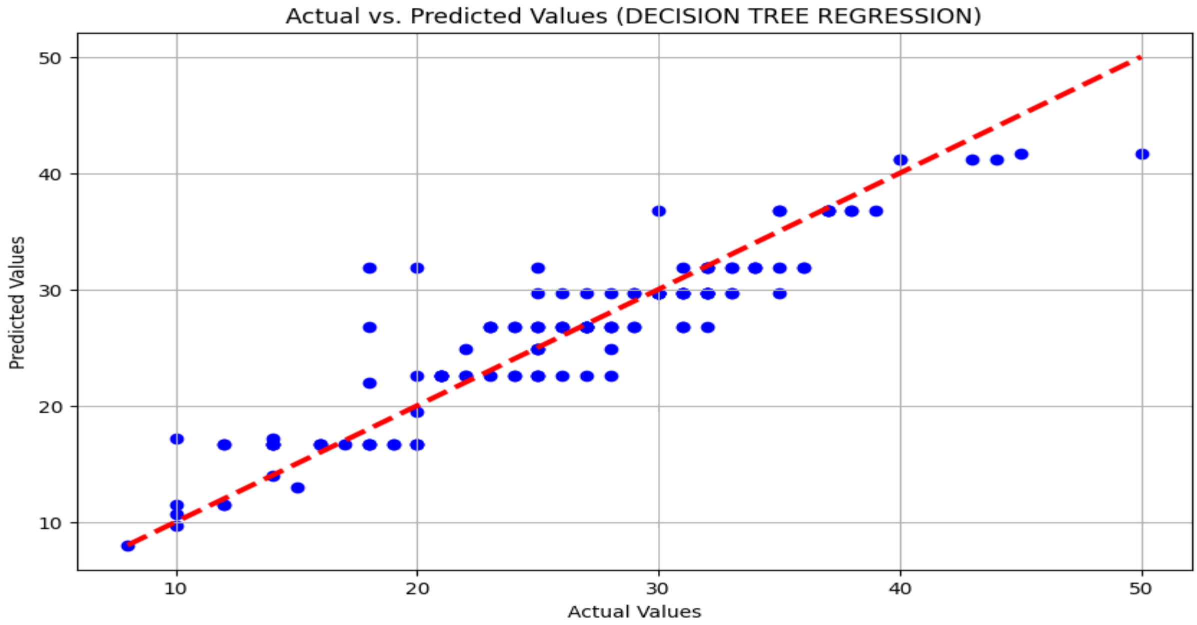
4. **Online Learning and Adaptation:** Explore methods for continuous learning and adaptation of predictive models based on real-time data from operational distillation processes.
5. **Integration with Process Simulation:** Integrate machine learning models with process simulation software to create comprehensive design and optimization tools for distillation systems.
6. **Industry Collaboration and Validation:** Collaborate with industry partners to validate and deploy machine learning solutions in real-world distillation applications, ensuring scalability and practical relevance.

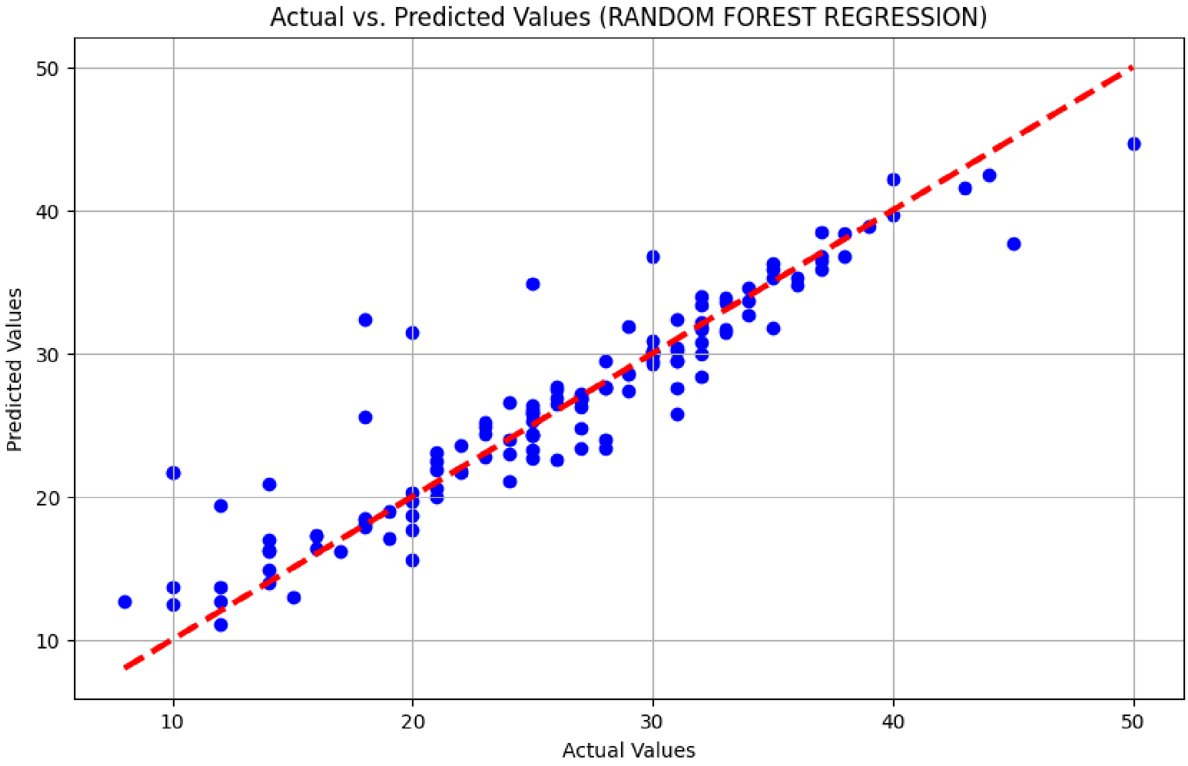
8 References

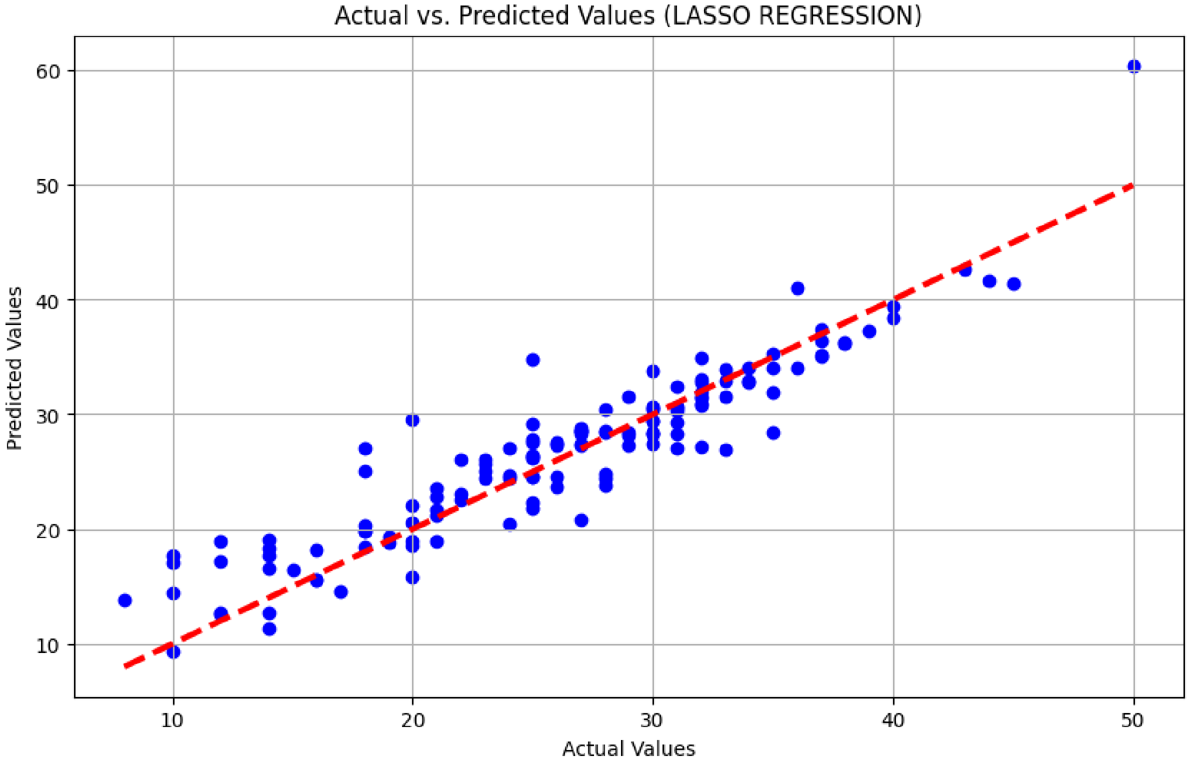
- Scikit-learn documentation:https://scikit-learn.org/stable/
- Seaborn:https://seaborn.pydata.org/
- Matplotlib:https://matplotlib.org/
- Principles of Mass Transfer And Separation Process By B.K Dut

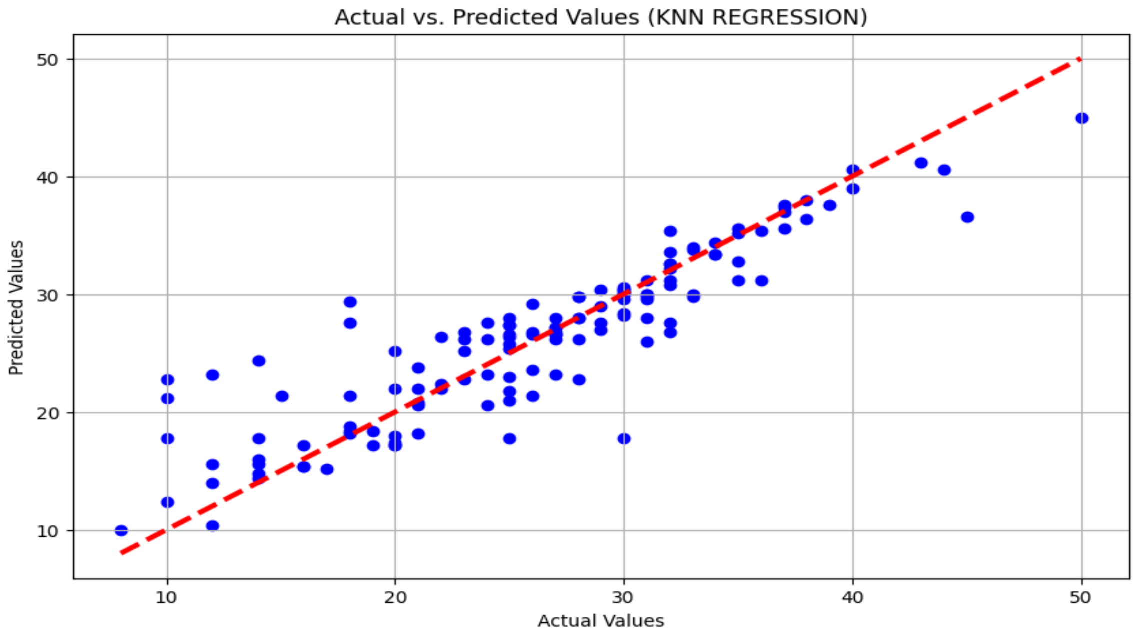
9 Appendices






Actual vs. Predicted Values (RANDOM FOREST REGRESSION)
Predicted Values
Actual Values
10
20
30
40
50
10
20
30
40
50


Actual vs. Predicted Values (LASSO REGRESSION)
Predicted Values
Actual Values
10
20
30
40
50
60
10
20
30
40
50



10 Auxiliaries

- **Data Source: Dataset**
- **Python file:https://drive.google.com/file/d/1qKCM5FdVsayw6vRpC3dCZjG-RQSzWzqk/view?usp=drive_link**