AI/ML Surrogate Modeling for Methanol–Ethanol Binary Distillation

1. Introduction / Objective

The primary objective of this project is to develop and compare multiple machine learning (ML) surrogate models to predict the performance of a binary distillation column for Methanol-Ethanol separation. Given that Ethanol-Water is an azeotropic mixture, the chemical system for this study was changed to Methanol-Ethanol to ensure a feasible separation. Six different ML algorithms were evaluated: Polynomial Regression, Random Forest, an Artificial Neural Network (ANN), Support Vector Regression (SVR), Gradient Boosting, and XGBoost. The goal is to identify a model that is not only highly accurate but also robust and physically consistent, making it a suitable replacement for computationally intensive DWSIM simulations.

2. Simulation and Data Generation

A steady-state DWSIM simulation was used to generate the dataset. The flowsheet consisted of a distillation column with a total condenser and a kettle-type reboiler, using Raoult's Law as the property package.

Input and Output Variables:
The following table details the ranges of the input variables used to generate the 1200 data points required for the study, along with the outputs that were recorded.

Parameter	Symbol	Range / Value	Role
Reflux Ratio	R	0.8 - 5.0	Input
Feed Mole Fraction (Methanol)	xF	0.2 - 0.95	Input
Feed Flowrate	F	70 - 130 kmol/hr	Input
Number of Stages	N	30	Fixed
Distillate Purity (Methanol)	xD	0.506 - 0.9999	Output
Reboiler Duty	QR	871.1 - 1661.1 kW	Output

3. Exploratory Data Analysis (EDA)

Before model training, an exploratory data analysis was performed.

Feature distributions of inputs and outputs show uniform sampling for inputs and more complex distributions for outputs, especially high frequency at high purities.

The feature distributions in Figure 1 show a uniform sampling of inputs and the resulting distributions of the outputs.

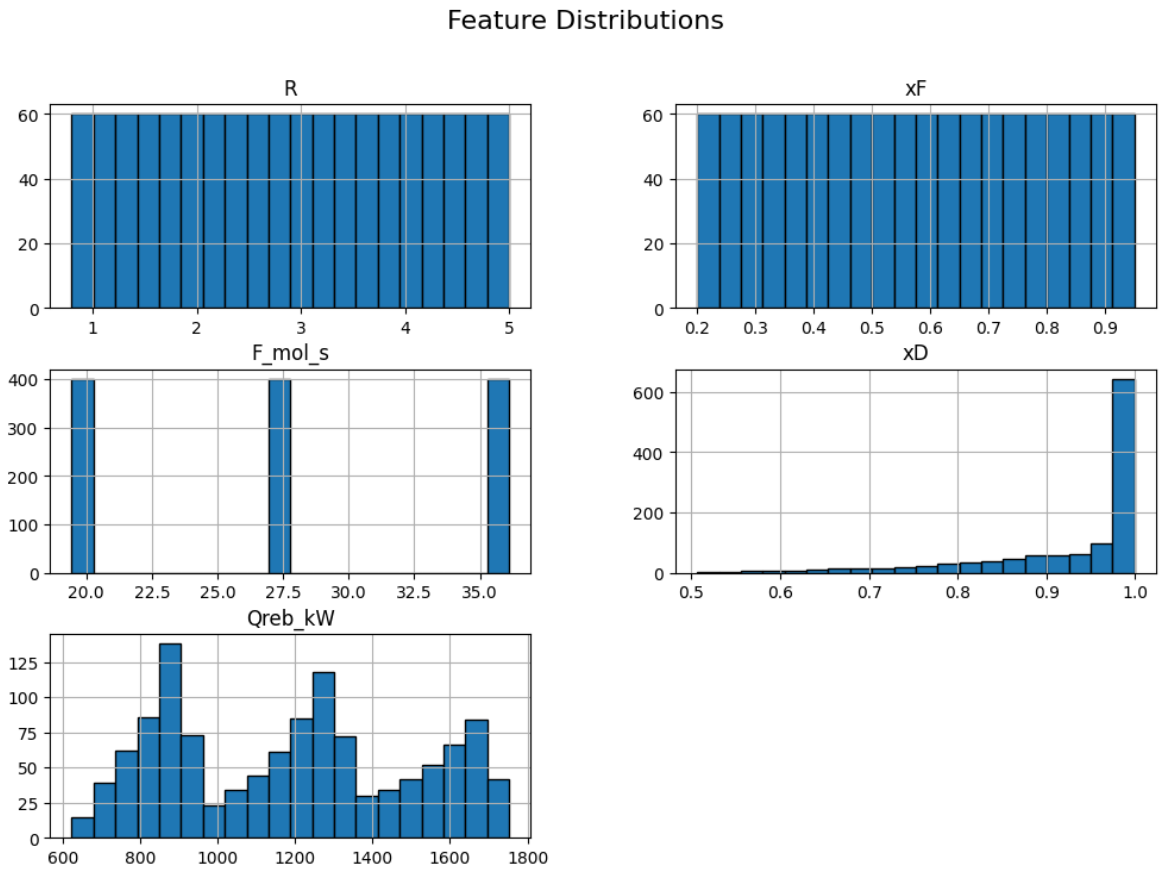
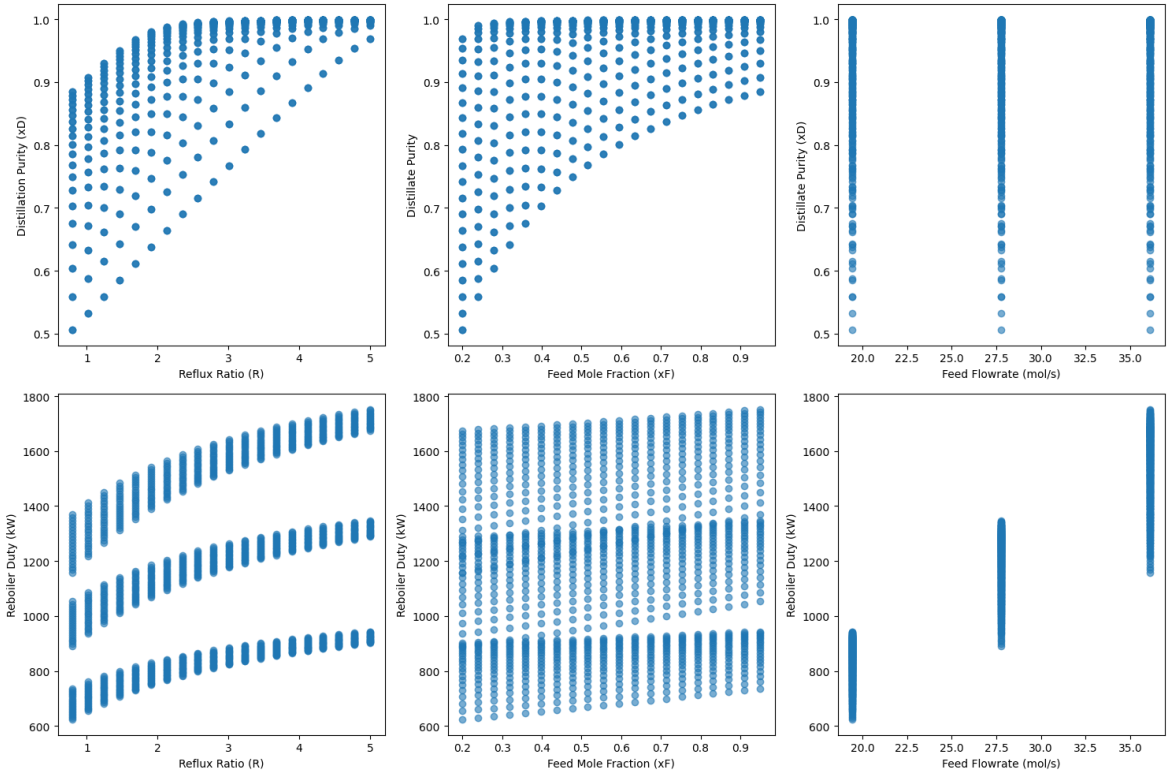


Figure 1

Scatter plots of xD and Qreb_kW vs. process inputs confirm expected non-linear trends

The scatter plots in **Figure 2** and **Figure 3** visualize the strong, non-linear relationships between the inputs and outputs



(Figure 2: Scatter plot - xD VS inputs) and (Figure 3: Scatter plot - Qreb_KW vs Inputs)

Correlation Heatmap confirms strong correlations: xD with reflux ratio and feed mole fraction; Qreb with feed flowrate.

The correlation heatmap in **Figure 4** reveals strong positive correlations between distillate purity (xD) and both reflux ratio (R) and feed composition (xF). These relationships align with the fundamental principles of distillation.

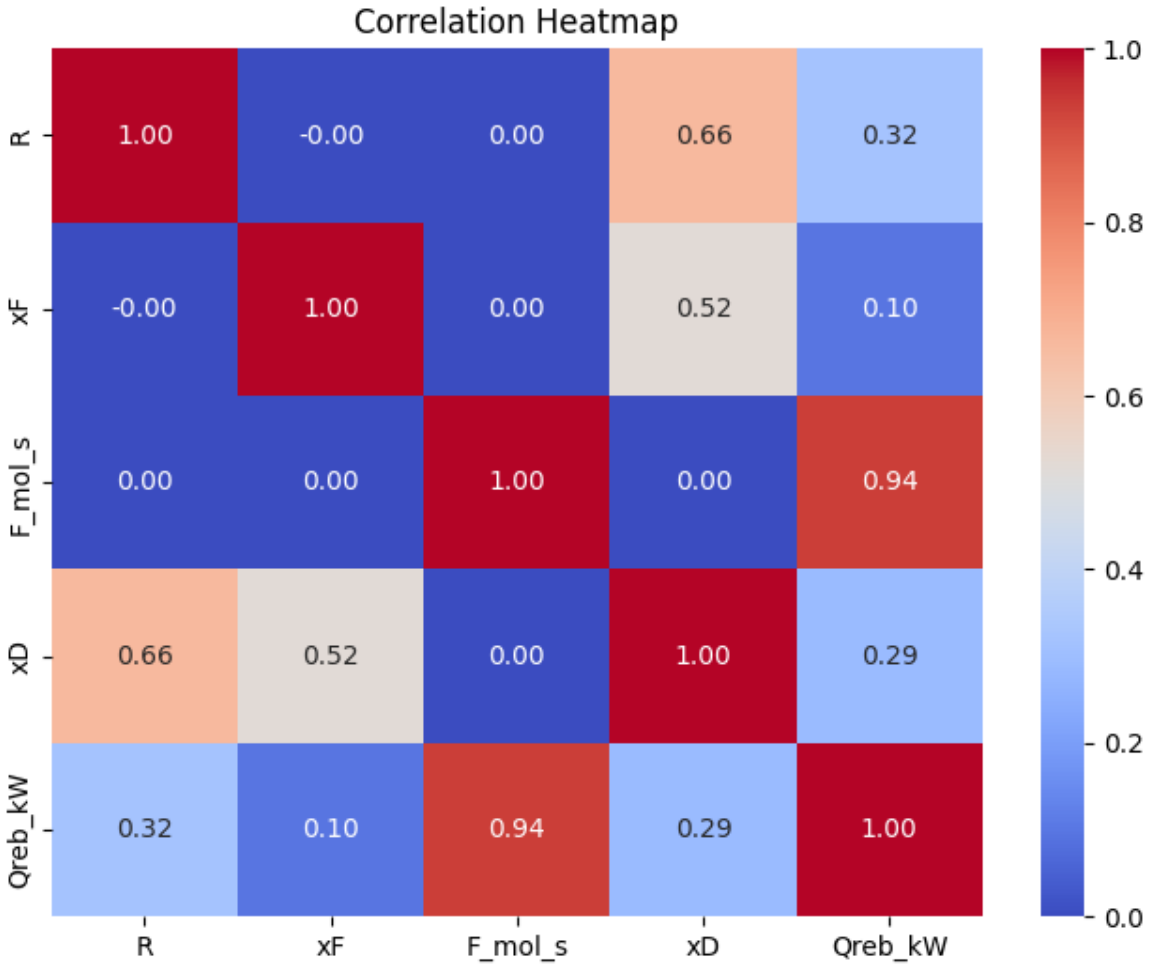


Figure 4

4. Model Evaluation

Six distinct machine learning models were developed to predict distillate purity (xD) and reboiler duty (Qr):

- Polynomial Regression: A baseline model to capture non-linearities.
- Random Forest: An ensemble of decision trees.
- Artificial Neural Network (ANN): A multi-layer perceptron.
- Support Vector Regression (SVR): A kernel-based method effective in high-dimensional spaces.
- Gradient Boosting: An ensemble boosting method.
- XGBoost: A highly optimized implementation of gradient boosting.

All models were trained and tuned on the same dataset. A detailed comparison revealed that while several models performed well, SVR and XGBoost were the top contenders.

5. Results and Evaluation

Performance Metrics

The performance of the top models was evaluated on a held-out test set. The table below shows a comparison between the two best-performing models: SVR and XGBoost.

Target Variable	Model	MAE	RMSE	R2 Score
Distillate Purity (xD)	SVR	0.002459	0.003301	0.999912
	XGBoost	0.000592	0.001045	0.999121
Reboiler Duty (QR, kW)	SVR	1.550805	1.808395	0.999967
	XGBoost	2.923606	4.779384	0.999767

Parity Plot Analysis

Parity plots show predicted vs actual values for different models.

The parity plots in **Figure 5 and Figure 6** provide a clear visual comparison.

- **Support Vector Regression (SVR)** demonstrated exceptional performance. For both xD and QR, its predictions align almost perfectly with the actual values, showing minimal deviation from the 45-degree line. This indicates the highest level of predictive accuracy.
- **XGBoost** also performed at a very high level, though its parity plot for xD may show slightly more scatter compared to SVR.

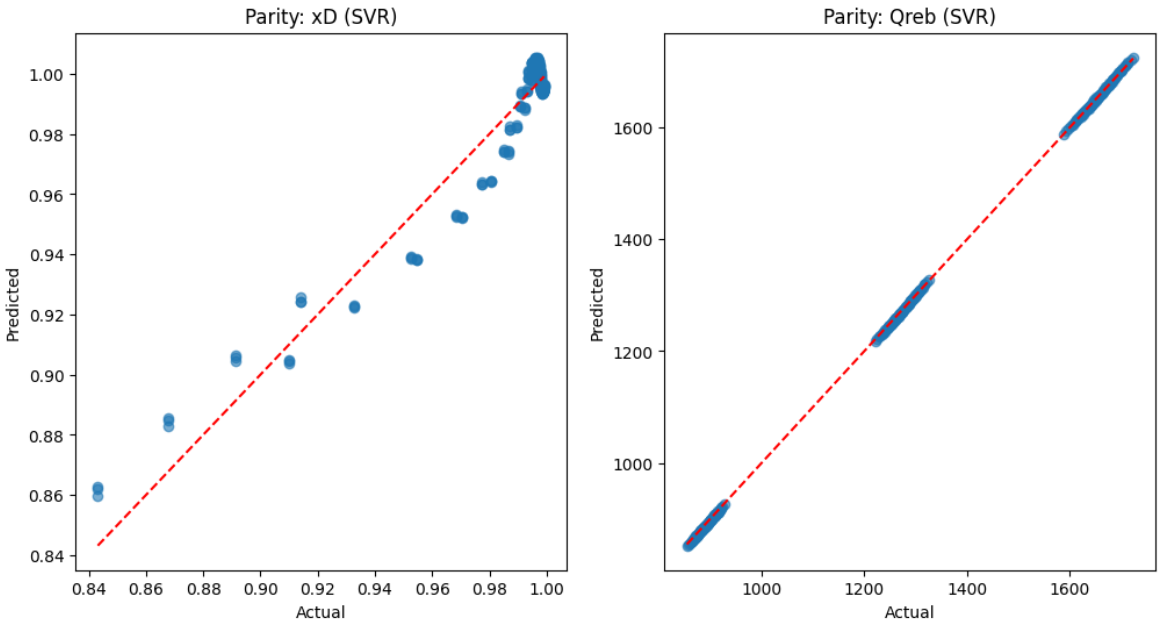


Figure 5

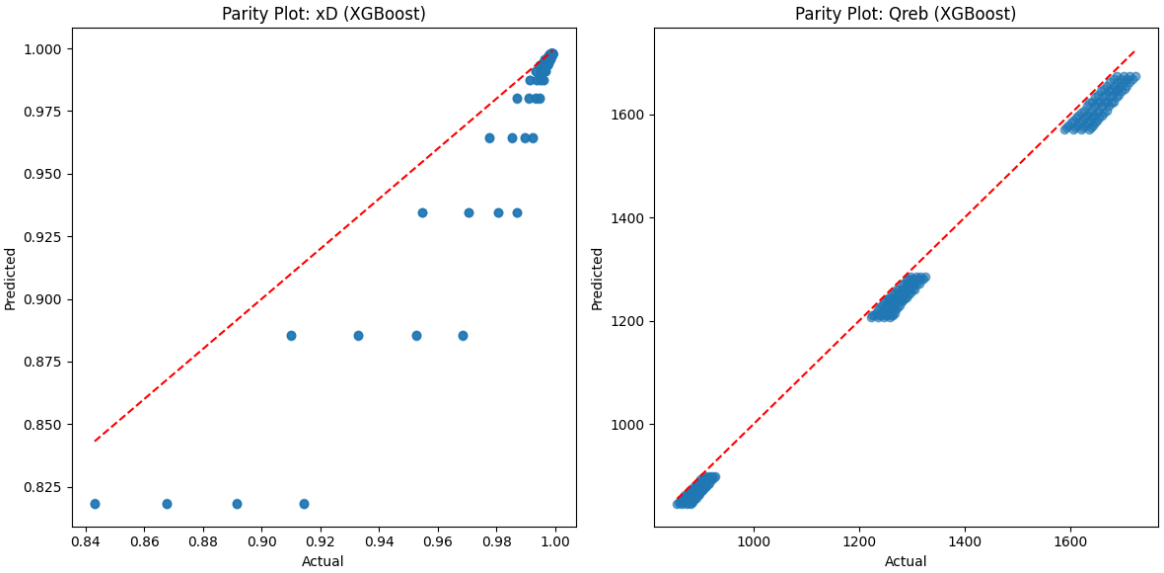


Figure 6

Residual Analysis

Residual analysis for Gradient Boosting shows small, unbiased residuals.

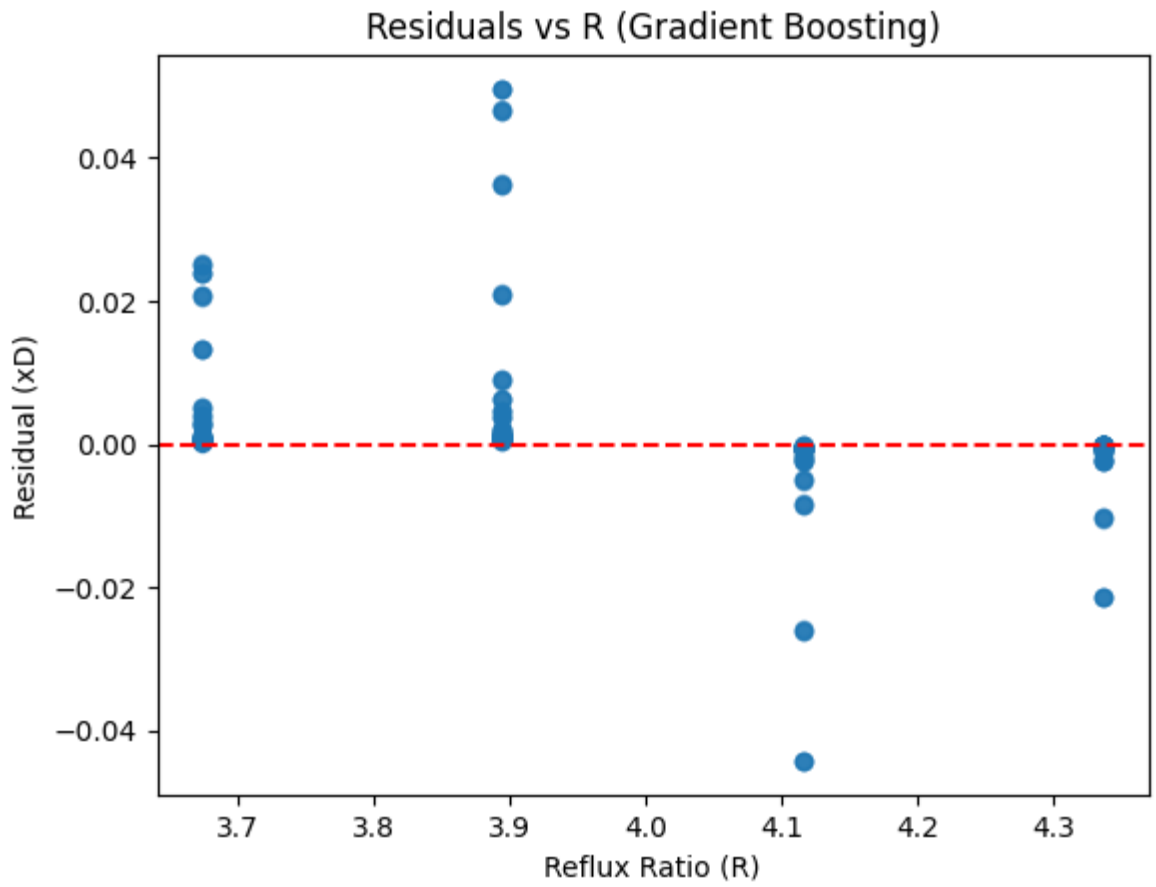


Figure 7

The residual plot in **Figure 7** shows that the errors are small, centered around zero, and unbiased, which is characteristic of a well-fitted model.

6. High-Purity Slice Analysis

A separate analysis was conducted to evaluate the model's performance specifically in the high-purity operating region (where xD $\geq$0.95), which is often the most critical for industrial applications. The Mean Absolute Error (MAE) for the XGBoost model in this slice was calculated.

High Purity Region MAE (xD , QR): [0.0121, 39.47 kW]

The very low error for purity (0.0121) confirms that the model maintains its high accuracy and reliability even when predicting near-perfect separations.

7. Physical Consistency & Diagnostics

The sensitivity analysis in **Figure 8** and **Figure 9** confirms that the XGBoost model's predictions are physically consistent. The plots correctly show that distillate purity (xD) increases with both a higher reflux ratio (R) and a higher mole fraction of methanol in the feed (xF).

Partial dependence plots confirm that surrogate models capture physically meaningful relationships. xD increases with both reflux ratio and feed mole fraction, consistent with distillation principles.

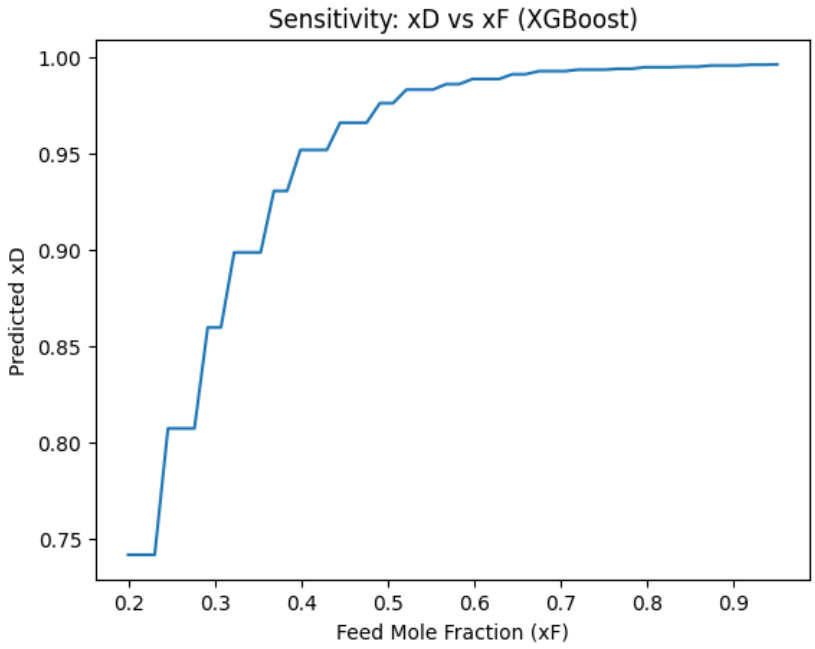


Figure 8

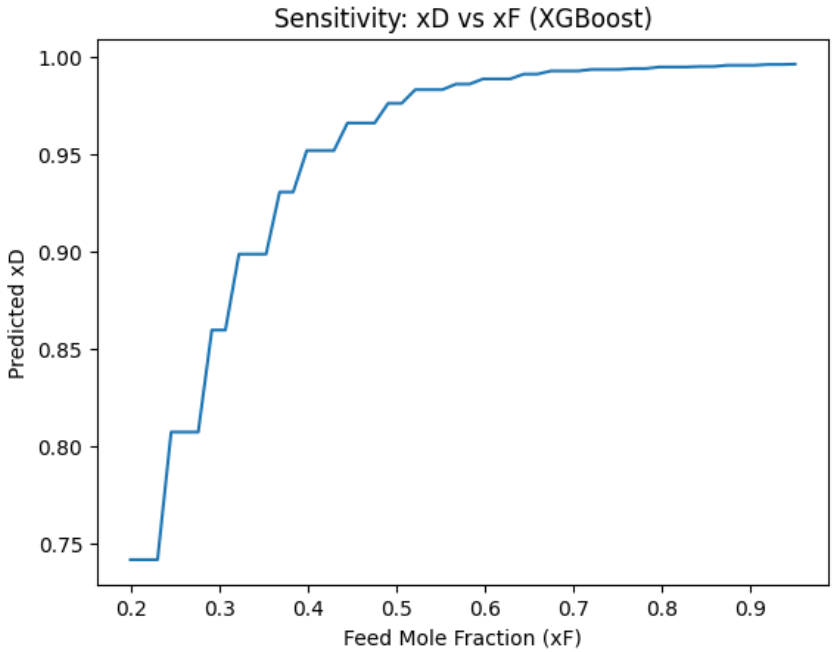


Figure 9

8. Conclusions

The evaluation of six different surrogate models revealed that multiple algorithms can accurately predict the performance of the Methanol-Ethanol distillation column.

In terms of pure predictive accuracy, Support Vector Regression (SVR) was the top-performing model. It achieved the lowest error metrics and near-perfect R2 scores.

However, for practical engineering applications, XGBoost is selected as the final recommended surrogate model. While its accuracy is marginally lower than SVR, it offers an outstanding balance of high performance, robustness, and superior interpretability (e.g., feature importance analysis). This transparency makes it a more reliable and trustworthy tool for process optimization and control. The model's predictions are physically consistent and it performs reliably in the critical high-purity region, making it a powerful substitute for the rigorous DWSIM simulator.