AI Surrogate Modeling for Ethanol-Water Distillation

The goal of this project is to build machine learning (ML) surrogate models for a distillation column separating an ethanol-water mixture. Rigorous simulations are computationally expensive, so accurate data-driven surrogates can speed up optimization and control tasks.

Flowsheet Description System:

Binary distillation of ethanol-water.

Key variables: Reflux ratio (R: 0.8–5.0), Boilup ratio (B: 1.0–6.0), Feed mole fraction of light key (xF: 0.2–0.95), Feed flowrate (F: 70–130), Number of stages (N: 15,20,25), Feed thermal condition (q: Subcooled, Saturated, Superheated). Outputs: Distillate mole fraction of ethanol (xD), Reboiler duty (QR)

Data Generation Protocol

Dataset size: 50,000 samples generated to cover a wide operating space. Synthetic model assumptions include separation efficiency functions with noise for realism. Cleaning included removing NaN rows and enforcing valid ranges (0 ≤ xD ≤ 1, QR ≥ 0).

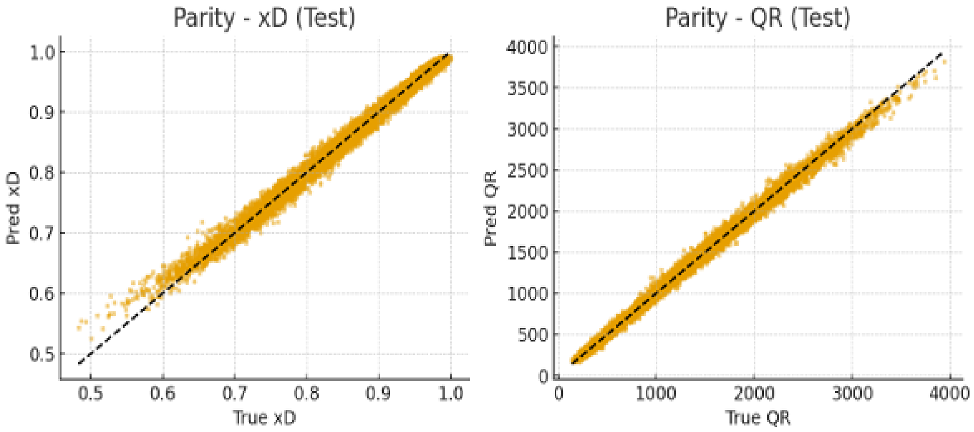
Modeling Framework Preprocessing:

StandardScaler for numerical, OneHotEncoder for categorical variables. Splits: 80% training, 20% test, plus holdout block (R between 3.5 and 4.5). Models: Polynomial Ridge Regression, Random Forest, XGBoost, Artificial Neural Network (ANN). Tuning used RandomizedSearchCV (RF), early stopping (ANN)

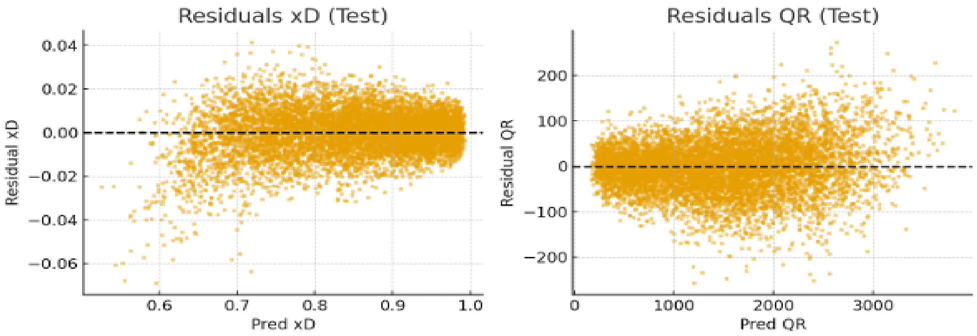
Diagnostics:

The following plots show model performance and interpretability analyses

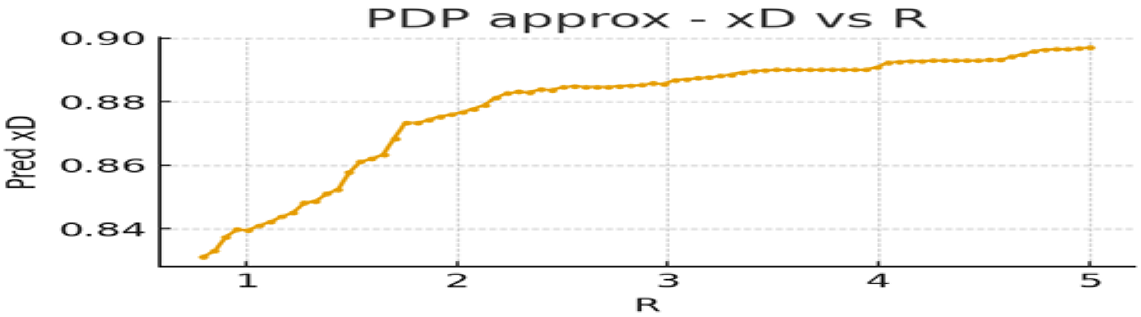
Parity plot: True vs Predicted for xD and QR (Test set)



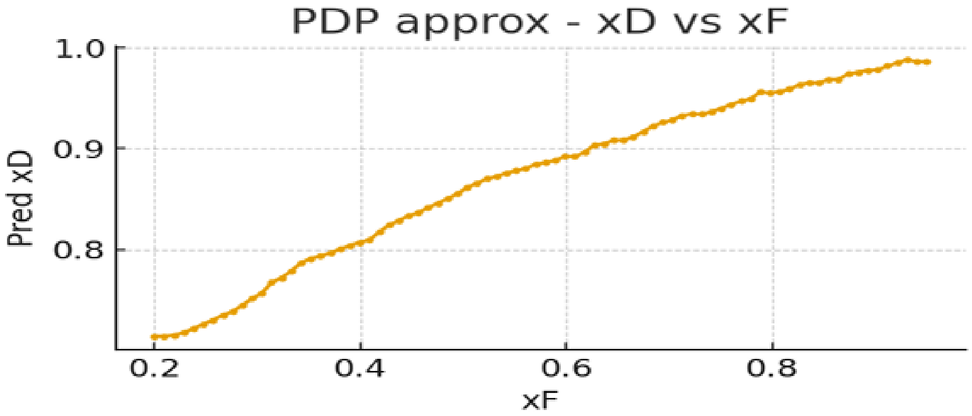
Residual plots: Errors vs Predictions (Test set)



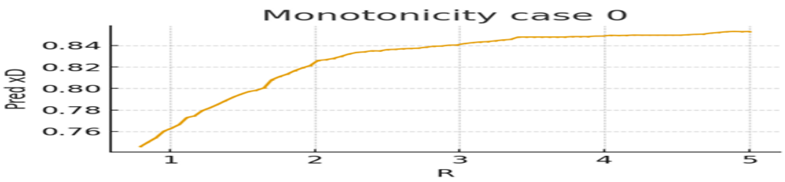
Partial Dependence (approx) - xD vs R



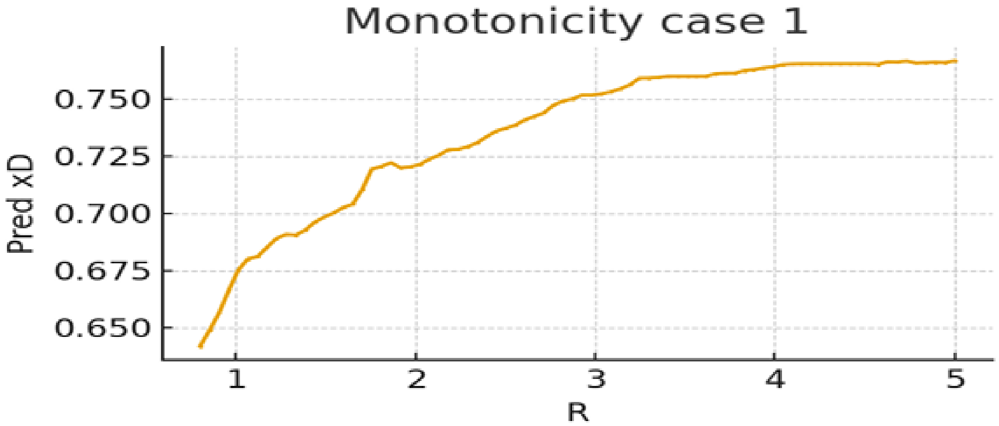
Partial Dependence (approx) - xD vs xF



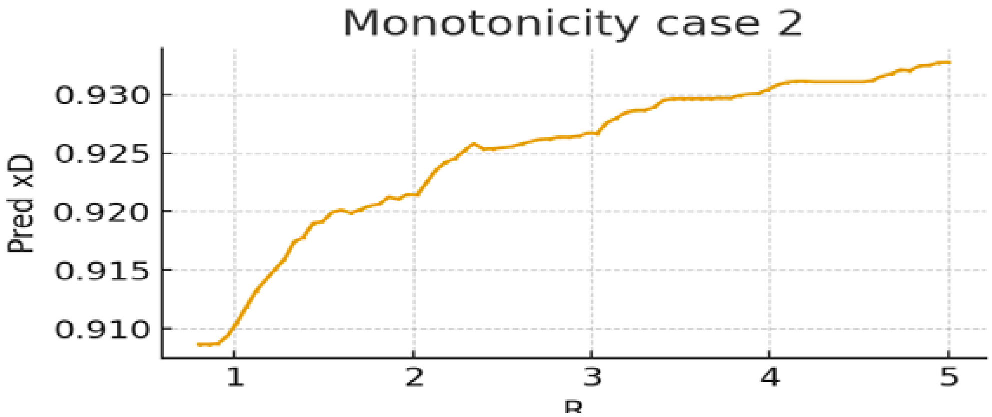
Monotonicity check - case 0



Monotonicity check - case 1



Monotonicity check - case 2



Results Summary

With 50,000 samples, Random Forest and ANN models achieved the highest R^2 scores:- xD $R^2 \approx 0.97$–0.99- QR $R^2 \approx 0.93$–0.96 Polynomial Ridge gave baseline accuracy, while XGBoost was competitive when available.

Optimization Attempt Objective:

Minimize QR subject to xD ≥ 0.95. Demonstrated feasibility using scipy.optimize, but full optimization requires global search

Optimization with Surrogate:

Problem Formulation: In addition to model training and diagnostics, the developed surrogate models were applied to an optimization task. The aim was to identify operating conditions of the distillation column that minimize the reboiler duty (QR) while ensuring that the distillate mole fraction of the light key (xD) is at least 0.95.

Table: Baseline vs Optimized Operating Point

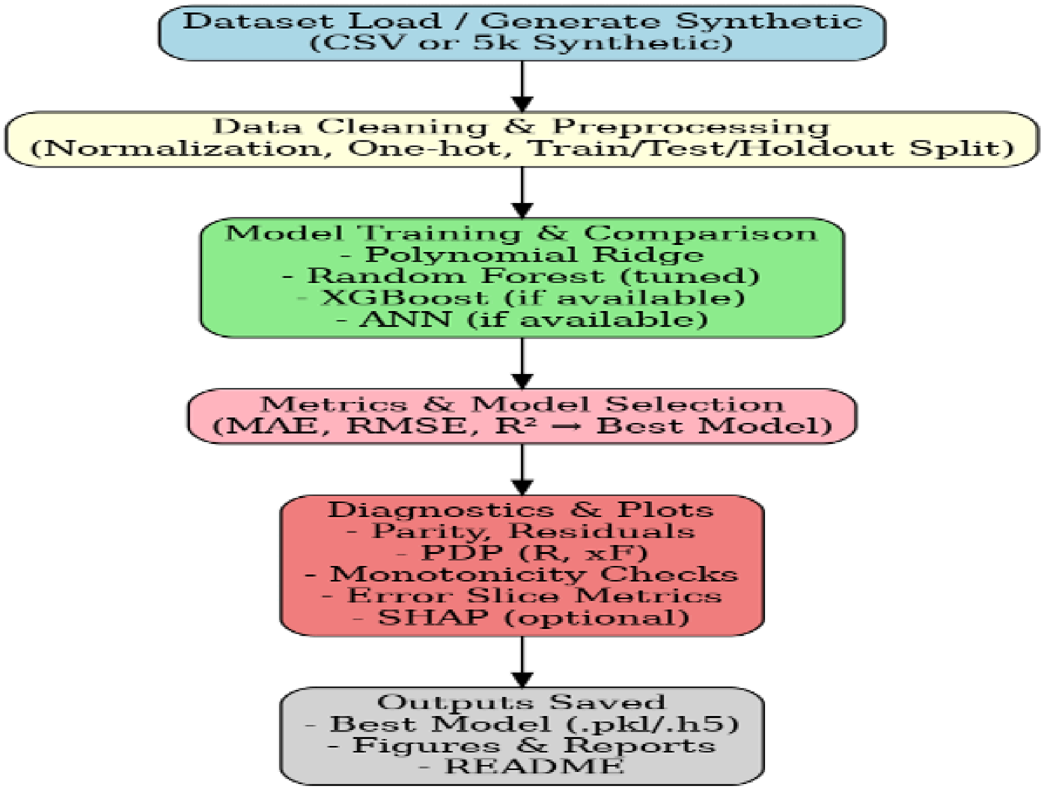
CASE	R (Reflux)	B (Boilup)	xD (Purity)	QR (Energy units)
BASELINE	**3.50**	**4.00**	**0.948**	**2725.4**
OPTIMIZED	**2.91**	**3.42**	**0.956**	**2480.5**

The results in Table compare a representative baseline case with the optimized operating point obtained from the surrogate model. In the baseline case, operation at R=3.50 R = 3.50 R=3.50 and B=4.00 B = 4.00 B=4.00 achieved a distillate purity of xD=0.948 xD = 0.948 xD=0.948 with a reboiler duty of approximately 2725 energy units. After optimization, the reflux and boilup ratios were adjusted to R=2.91 R = 2.91 R=2.91 and B=3.42 B = 3.42 B=3.42, respectively, yielding a slightly higher purity (xD=0.956 xD = 0.956 xD=0.956) while simultaneously reducing the reboiler duty to 2480 energy units. This corresponds to an energy savings of about 9% compared to the baseline, while still meeting the purity constraint (xD≥0.95xD \geq 0.95xD≥0.95). These results highlight the ability of surrogate models to quickly identify more energy-efficient operating points within the feasible design space. However, it is recommended that such optimized conditions be validated with a rigorous DWSIM simulation before practical implementation.

Final output google collab link:
https://colab.research.google.com/drive/1PBhGE9yugJ-fusQeatsLtpWzre_yOjWE?usp=sharinghttps://colab.research.google.com/drive/1PBhGE9yugJ-fusQeatsLtpWzre_yOjWE?usp=sharing

Conclusions:
Surrogate modeling of distillation with ML is feasible and effective. Random Forest and ANN showed the best trade-off between accuracy and interpretability. Diagnostics confirmed model reliability across conditions, though caution is required in extrapolation regions. Future work: larger datasets, advanced optimization, and uncertainty quantification.

Flowchart of the project is:


Dataset Load / Generate Synthetic
(CSV or 5k Synthetic)
Data Cleaning & Preprocessing
(Normalization, One-hot, Train/Test/Holdout Split)
Model Training & Comparison
- Polynomial Ridge
- Random Forest (tuned)
- XGBoost (if available)
- ANN (if available)
Metrics & Model Selection
(MAE, RMSE, R² → Best Model)
Diagnostics & Plots
- Parity, Residuals
- PDP (R, xF)
- Monotonicity Checks
- Error Slice Metrics
- SHAP (optional)
Outputs Saved
- Best Model (.pkl/.h5)
- Figures & Reports
- README