Machine Learning Modelling of Catalytic Ammonia Decomposition Process

Bishan Kumar Deka

220107025

Final Project submission

Course Name: Applications of AI and ML for Chemical Engineering

Course Code: CL653

Submission Date: April 25, 2025

Contents

1 Executive Summary 3

2 Introduction 3

3 Methodology 4

4 Implementation Plan 7

5 Testing and Deployment 9

6 Results and Discussion 10

7 Conclusion and Future Work 12

8 References 12

9 Appendices 13

10 Auxiliaries 17

1 Executive Summary

This project focuses on applying machine learning techniques to model and optimize catalytic ammonia decomposition for hydrogen production. The traditional catalyst selection process is time-consuming and expensive, relying on extensive trial-and-error experiments. By leveraging machine learning models, we aim to establish relationships between catalyst properties, reaction parameters, and ammonia conversion efficiency, allowing for more efficient catalyst design and selection.

The project utilizes a dataset of 1492 samples containing catalyst properties and reaction parameters to predict ammonia conversion percentages. Through extensive data analysis, dimensionality reduction using Principal Component Analysis (PCA), and the implementation of various regression models including Random Forest, Decision Tree, and Support Vector Regression, we have developed predictive models that can estimate ammonia conversion rates based on catalyst characteristics.

Our findings demonstrate that machine learning approaches can effectively model the complex relationships in catalytic processes, providing a valuable tool for catalyst practitioners to accelerate the development of efficient hydrogen production methods via ammonia decomposition.

2 Introduction

Background:

Hydrogen, as a fuel, has tremendous importance in the modern world. It is used in oil refining, to power vehicles, as rocket fuel, as coolants, and as feedstock in the production of many industrial chemicals. Therefore, the production of hydrogen is a topic of utmost importance. Traditional methods of hydrogen production primarily use fossil fuels that emit CO2 during the manufacturing process. However, the pressing challenge of global warming brought by rapid urbanization, population expansion, and economic development in recent decades stresses the need for alternative methods for hydrogen production.

Ammonia decomposition emerges as a promising solution. Ammonia has a higher liquefaction temperature (-33.4°C) compared to hydrogen gas (-253°C) at atmospheric pressure, making its storage much easier. The existing infrastructure for synthesizing, transporting, and distributing

ammonia is readily available, making ammonia decomposition a frontrunner as a means of carbon-free hydrogen production.

Hydrogen is formed from ammonia through the endothermic reaction: $2NH_3(g) \rightarrow N_2(g) + 3H_2(g)$

Along with hydrogen, nitrogen is produced, which also has considerable industrial use. While the reaction can occur without catalysts, introducing them reduces the activation energy and results in energy savings by lowering the required temperatures for conversion.

Problem Statement:

The selection of catalysts for ammonia decomposition traditionally requires extensive trial-and-error experiments conducted by catalyst practitioners. This process is costly, time-consuming, and resource-intensive. There is a critical need for a more efficient approach to identify and optimize catalysts for ammonia decomposition.

This project addresses this challenge by applying machine learning techniques to model the relationship between catalyst properties, reaction conditions, and ammonia conversion efficiency. By developing accurate predictive models, we aim to provide a tool that can accelerate catalyst selection and optimization, reducing the experimental burden and facilitating the development of more efficient hydrogen production methods.

This study builds upon research presented in the article “Machine learning for predicting catalytic ammonia decomposition: An approach for catalyst design and performance prediction” (Source: https://www.sciencedirect.com/science/article/pii/S2352152X24012738).

Objectives:

- To analyze the relationships between catalyst properties, reaction conditions, and ammonia conversion efficiency using exploratory data analysis
- To develop and compare different machine learning models for predicting ammonia conversion rates
- To provide insights into the most important features affecting ammonia decomposition efficiency

3 Methodology

Data Source:

The dataset used in this project contains 1492 samples of catalyst properties and reaction parameters for ammonia decomposition. Each sample includes information about catalyst characteristics, reaction conditions, and the resulting ammonia conversion percentage. The data was sourced from published research in the field of catalytic ammonia decomposition, specifically from the study published at https://www.sciencedirect.com/science/article/pii/S2352152X24012738.

The dataset includes the following features:

- Catalyst total metal loading (%)
- Catalyst average crystallite size (nm)
- Catalyst crystallinity index (-)
- Catalyst specific surface area ($m^2.g^{-1}$)
- Catalyst pore volume ($cm^3.g^{-1}$)
- Catalyst average pore diameter (nm)
- Reaction temperature (°C)
- Gas hourly space velocity ($mL.h^{-1}.gcat^{-1}$)
- Ammonia conversion (%) [Target variable]

Data Preprocessing:

The data preprocessing phase included:

1. Removing non-numeric identifiers (DOI column)
2. Checking for missing values and handling them appropriately
3. Exploring the distribution of each feature using histograms
4. Analyzing correlations between features using Spearman correlation coefficients
5. Standardizing features to ensure all variables contribute equally to the analysis
6. Implementing Principal Component Analysis (PCA) to reduce dimensionality while preserving data variance.

Model Architecture:

Several regression models were implemented and compared to predict ammonia conversion percentages:

1. **Random Forest Regressor**: An ensemble learning method that constructs multiple decision trees during training and outputs the average prediction of the individual trees. Random forests are robust to overfitting and can handle non-linear relationships effectively.
2. **Decision Tree Regressor**: A non-parametric supervised learning method that creates a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.
3. **Support Vector Regressor (SVR)**: A type of Support Vector Machine that uses the same principles as SVM for classification, but adapted for regression problems. SVR works well for capturing complex patterns in data.

These models were chosen because of their ability to handle non-linear relationships, which are common in chemical processes like catalytic reactions, and their interpretability, which is important for providing insights into the factors affecting ammonia conversion.

Tools and Technologies:

The project was implemented using the following tools and technologies:

- **Python**: Primary programming language
- **NumPy**: For numerical computing and array operations
- **Pandas**: For data manipulation and analysis
- **Scikit-learn**: For implementing machine learning algorithms
- **Matplotlib and Seaborn**: For data visualization and plotting
- **Pickle**: For model serialization and storage

4 Implementation Plan

Development Phases:

The project was implemented in the following phases:

1. Data Collection and Exploration:
    - Import data and libraries
    - Perform exploratory data analysis (EDA)
    - Generate histograms for distribution of all variables
    - Create heatmap using Spearman correlation matrix
    - Develop scatter plots for visualization of different input features against ammonia conversion
2. Data Preprocessing and Dimensionality Reduction:
    - Apply StandardScaler for feature normalization
    - Perform PCA with 4 components
    - Create plots for explained variance by principal components and cumulative explained variance ratio
    - Generate scatter plots of principal components vs. ammonia conversion
3. Model Development and Evaluation:
    - Split data into training and testing sets
    - Implement Random Forest Regressor, Decision Tree Regressor, and SVR models
    - Evaluate models using Mean Absolute Error (MAE), R^2 score, and Root Mean Square Error (RMSE)
    - Compare model performance using line plots
4. Model Optimization:
    - Perform hyperparameter tuning using 5-fold cross-validation
    - Re-evaluate models with optimized parameters
    - Compare performance metrics of tuned models
    - Export optimized models using pickle

Model Training:

The models were trained on a portion of the dataset after preprocessing and PCA transformation. The training process included:

1. **Data Splitting**: The dataset was split into training (80%) and testing (20%) sets to ensure unbiased evaluation.
2. **Base Model Training**: Each model (Random Forest, Decision Tree, SVR) was trained with default parameters to establish baseline performance.
3. **Hyperparameter Tuning**: Grid search with 5-fold cross-validation was employed to find optimal hyperparameters for each model. Parameters tuned included:
    - Random Forest: number of estimators, maximum depth, minimum samples split
    - Decision Tree: maximum depth, minimum samples split, minimum samples leaf
    - SVR: kernel type, C parameter, gamma value
4. **Final Model Training**: Models were retrained with optimized hyperparameters.

Model Evaluation:

To ensure comprehensive assessment of model performance, multiple evaluation metrics were used:

1. **Mean Absolute Error (MAE)**: Measures the average magnitude of errors without considering their direction.
2. **Root Mean Square Error (RMSE)**: Emphasizes larger errors and provides a metric in the same units as the target variable.
3. **R^2 Score (Coefficient of Determination)**: Indicates the proportion of variance in the dependent variable predictable from the independent variables.

These metrics were calculated for both the base models and the hyperparameter-tuned models to quantify the improvement achieved through optimization.

5 Testing and Deployment

Testing Strategy:

The models were tested using a hold-out test set (20% of the original data) that was not used during the training phase. This approach ensures an unbiased evaluation of model performance on unseen data. The testing process involved:

1. Applying the same preprocessing steps and PCA transformation to the test data
2. Using the trained models to make predictions on the test set
3. Computing evaluation metrics (MAE, RMSE, R^2) on the test set predictions
4. Comparing the performance across different models

Additionally, k-fold cross-validation (k=5) was implemented during hyperparameter tuning to ensure robust model selection and to avoid overfitting to any particular subset of the data.

Ethical Considerations:

In the context of catalyst development for ammonia decomposition, ethical considerations include:

1. **Environmental Impact**: The models aim to optimize hydrogen production from ammonia, which represents a more sustainable alternative to fossil fuel-based methods. This aligns with global efforts to reduce carbon emissions.
2. **Data Quality and Transparency**: All data processing steps were documented to ensure transparency in how results were obtained. The source of the dataset has been properly attributed.
3. **Scientific Integrity**: Care was taken to avoid overfitting and to report model performance metrics honestly, avoiding exaggerated claims about predictive capabilities.
4. **Resource Efficiency**: By potentially reducing the need for extensive laboratory experiments, these models could lead to more efficient use of research resources, including time, materials, and energy

6 Results and Discussion

Findings:

The exploratory data analysis revealed several important patterns:

1. **Variable Distributions**: Histograms (see Appendix Figure 1) of the features showed varying distributions, with some features exhibiting skewness that justified the standardization step.
2. **Correlation Analysis**: The Spearman correlation matrix (see Appendix Figure 2) highlighted significant relationships between catalyst properties and ammonia conversion. Notably, reaction temperature showed a strong positive correlation (0.69) with ammonia conversion, while other parameters showed weaker correlations. The correlation matrix also revealed a strong relationship (0.71) between catalyst specific surface area and pore volume.
3. **PCA Results**: As shown in Appendix Figure 4, the first principal component captured approximately 25.5% of the variance, with subsequent components contributing progressively less (15.8%, 15.2%, and 12.2%). Together, the first four principal components captured approximately 68.7% of the total variance in the data.
4. **Model Performance Comparison**: Among the base models (see Appendix Figure 6), Random Forest Regressor demonstrated the best overall performance with the lowest RMSE (22.88), a MAE of 17.24, and the highest R^2 value (0.55). SVR showed moderate performance (RMSE: 27.76, MAE: 22.41, R^2: 0.33), while the Decision Tree performed least favorably (RMSE: 30.44, MAE: 21.91, R^2: 0.20).
5. **Hyperparameter Tuning Effects**: As illustrated in Appendix Figure 7, optimization through hyperparameter tuning led to improved performance across all models. The Random Forest RMSE improved from 22.88 to 22.12, the Decision Tree from 30.44 to 26.57, and the SVR from 27.76 to 23.57. These improvements demonstrate the importance of hyperparameter optimization in maximizing model performance.

Comparative Analysis:

When comparing the performance of different models:

Model	MAE (Base)	RMSE (Base)	R^2 (Base)	MAE (Tuned)	RMSE (Tuned)	R^2 (Tuned)
Random Forest	17.24	22.88	0.55	16.41	22.12	0.57
Decision Tree	21.91	30.44	0.20	19.35	26.57	0.39
SVR	22.41	27.76	0.33	18.73	23.57	0.49

The Random Forest model consistently outperformed other models, likely due to its ensemble nature which helps capture the complex non-linear relationships present in catalytic processes. The Decision Tree model, while simpler, provided reasonable performance after tuning and offers greater interpretability. SVR demonstrated substantial improvement after hyperparameter tuning, with its RMSE decreasing by approximately 15%.

Challenges and Limitations:

Several challenges and limitations were encountered during this project:

1. **Data Quality**: While the dataset was comprehensive (1492 samples), it represents published experimental results which may have inherent biases or limitations in the range of catalysts and conditions tested.
2. **Feature Engineering**: The correlation matrix (Appendix Figure 4) shows that many catalyst properties have weak correlations with ammonia conversion, suggesting that additional or transformed features might be needed to better capture the underlying relationships.
3. **Model Interpretability**: While Random Forest provided the best predictive performance, its complex ensemble nature makes it less interpretable than simpler models like Decision Trees.
4. **Extrapolation Limits**: The models may have limited predictive power for catalyst types or reaction conditions that fall outside the range of the training data.
5. **Experimental Validation**: Computational predictions ultimately require experimental validation, which was beyond the scope of this project.

7 Conclusion and Future Work

This project successfully demonstrated the application of machine learning techniques to model and predict ammonia conversion in catalytic decomposition processes. The developed models, particularly the optimized Random Forest Regressor with an R^2 score of 0.57 and RMSE of 22.12, show promising capability in predicting conversion rates based on catalyst properties and reaction conditions.

The analysis revealed that reaction temperature has the strongest correlation (0.69) with ammonia conversion among all variables studied, highlighting its critical role in the process. The PCA analysis showed that approximately 68.7% of the variance in the data could be explained by just four principal components, suggesting significant interdependence among the catalyst properties.

The insights gained from this study can guide catalyst practitioners in selecting and developing more efficient catalysts for ammonia decomposition, potentially accelerating the transition to more sustainable hydrogen production methods. By reducing the reliance on trial-and-error experimentation, these models contribute to more efficient use of research resources.

Future work could focus on:

1. **Expanding the Feature Set**: Incorporating additional catalyst characteristics, such as elemental composition, synthesis methods, or support materials, particularly given the relatively low correlation values observed between many current features and ammonia conversion.
2. **Advanced Model Architectures**: Exploring deep learning approaches such as neural networks that may capture more complex relationships in the data, potentially improving upon the R^2 score of 0.57 achieved by the Random Forest model.
3. **Transfer Learning**: Applying knowledge gained from ammonia decomposition modelling to related catalytic processes.
4. **Experimental Validation**: Testing model predictions with new catalyst formulations in laboratory settings, particularly focusing on conditions where reaction temperature is optimized based on our finding of its strong influence on conversion.
5. **Active Learning**: Implementing an iterative approach where the model suggests promising catalyst formulations for experimental testing, with results feeding back into model improvement.
6. **Feature Importance Analysis**: Conducting deeper analysis of the Random Forest model's feature importance to identify which catalyst properties have the greatest impact on performance, helping to guide future catalyst design.

By continuing to refine and expand this machine learning approach, we can further advance the field of catalyst design for sustainable hydrogen production, contributing to the global transition toward cleaner energy systems.

8 References

- Guo et al., "Machine learning for predicting catalytic ammonia decomposition: An approach for catalyst design and performance prediction" (https://www.sciencedirect.com/science/article/pii/S2352152X24012738)
- Official documentation: scikit-learn, matplotlib, seaborn, pandas, numpy

9 Appendices

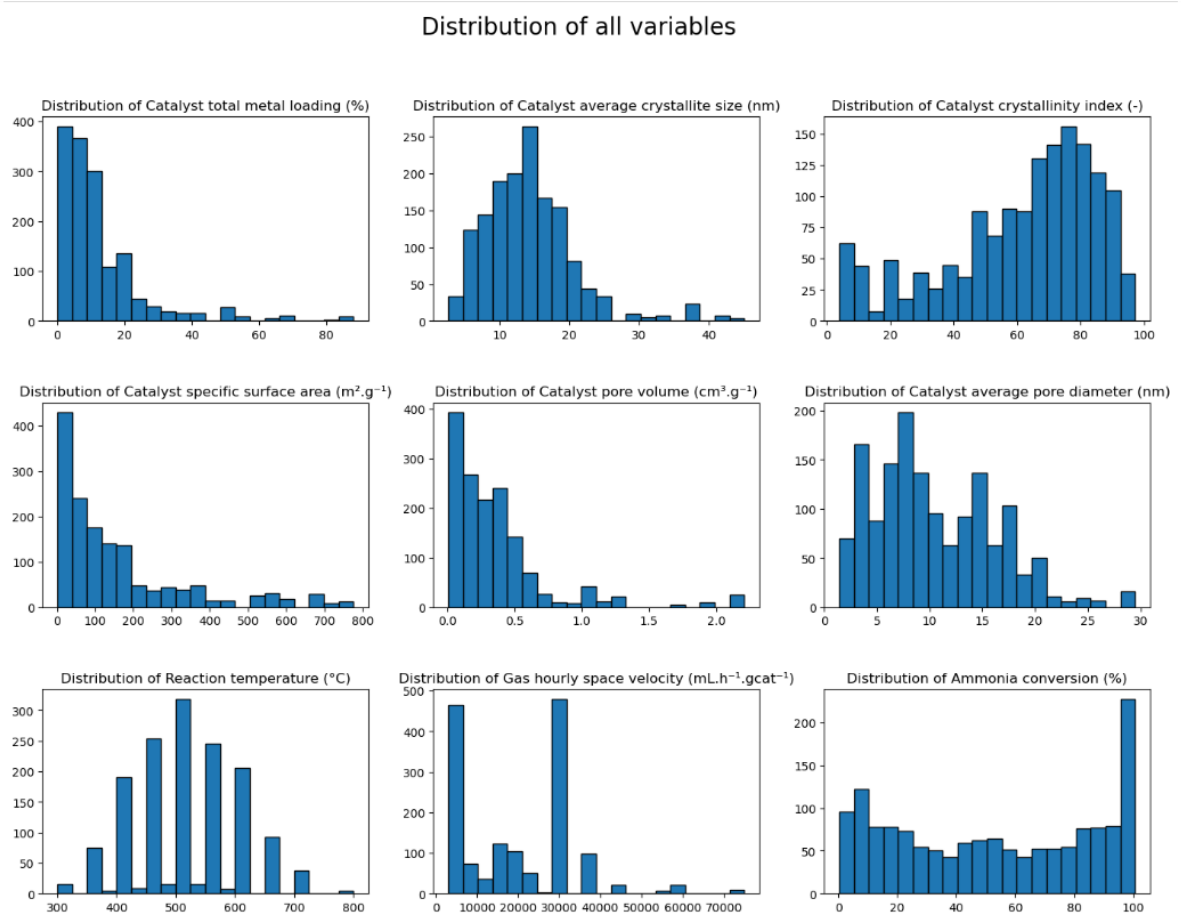

Fig1: Distribution of all variables using histogram plots

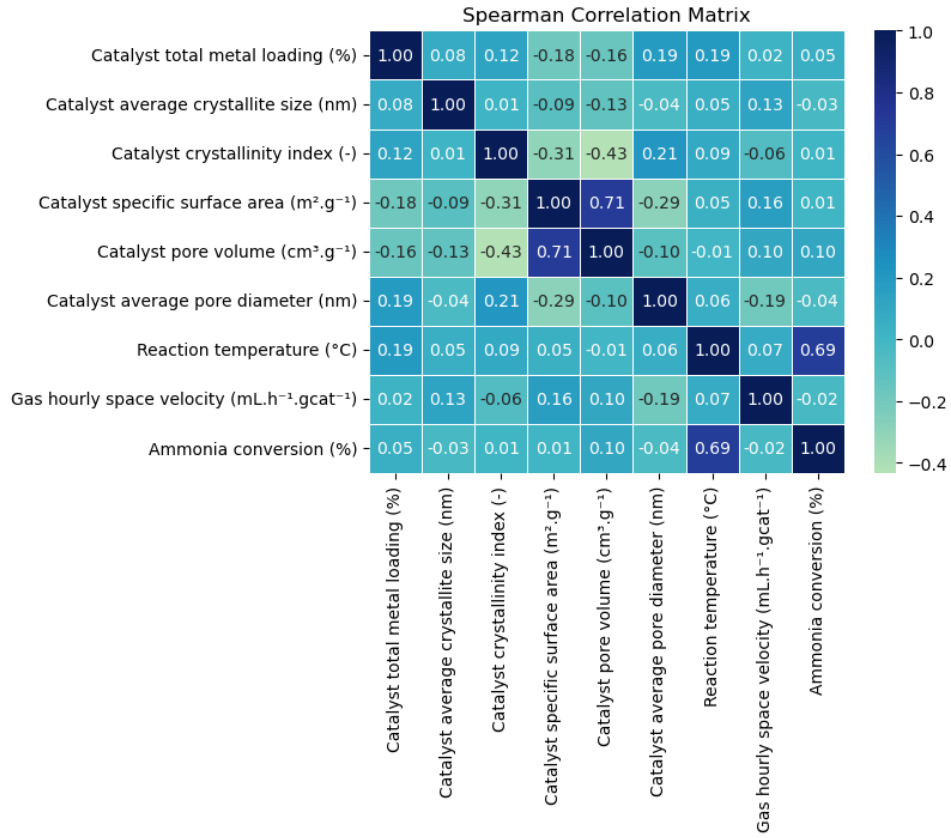

Fig2: Heatmap using Spearman Correlation Matrix

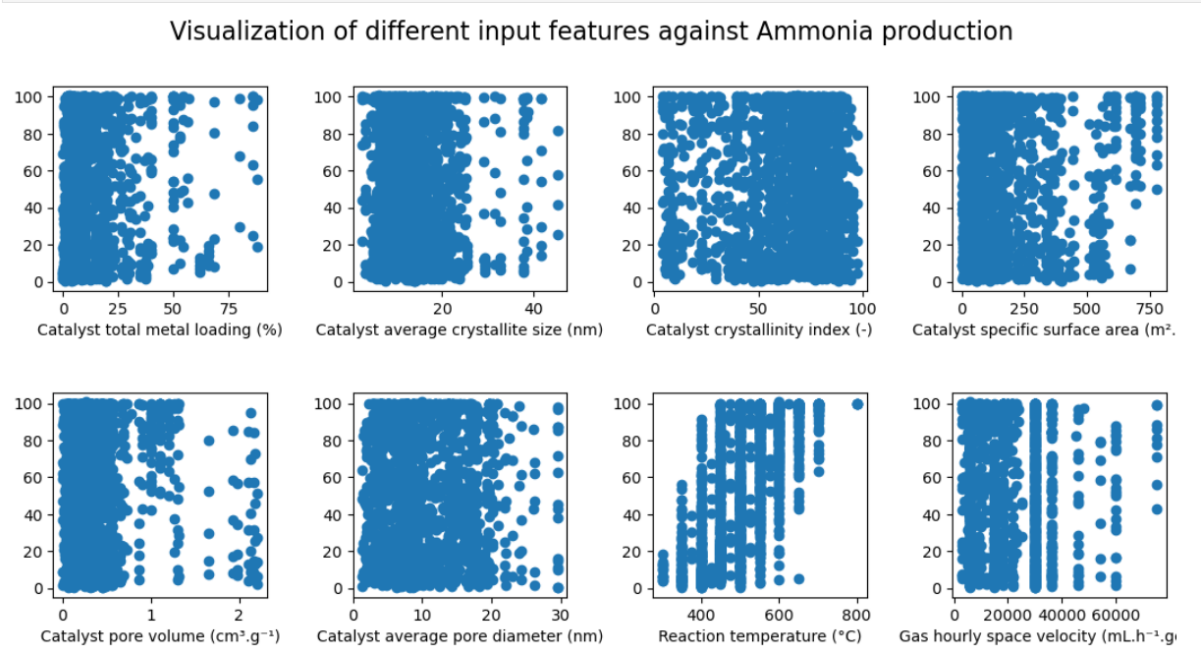

Fig3: Features vs Ammonia production scatter plots

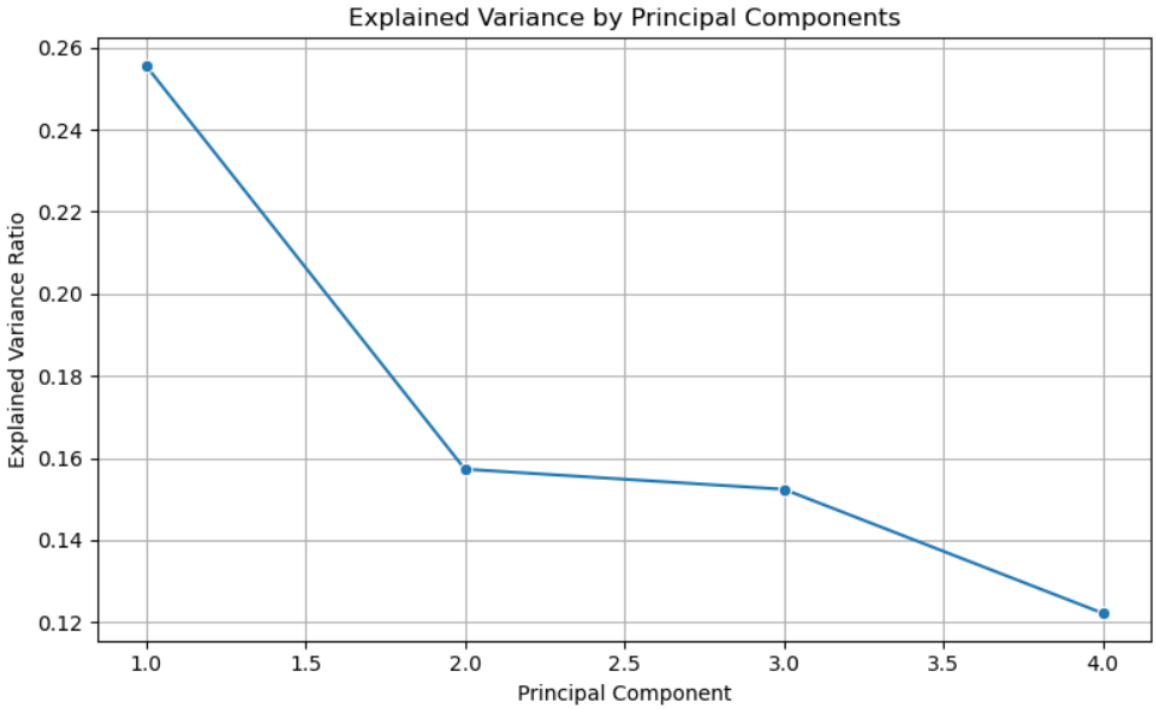

Fig4: Scree plot for PCA

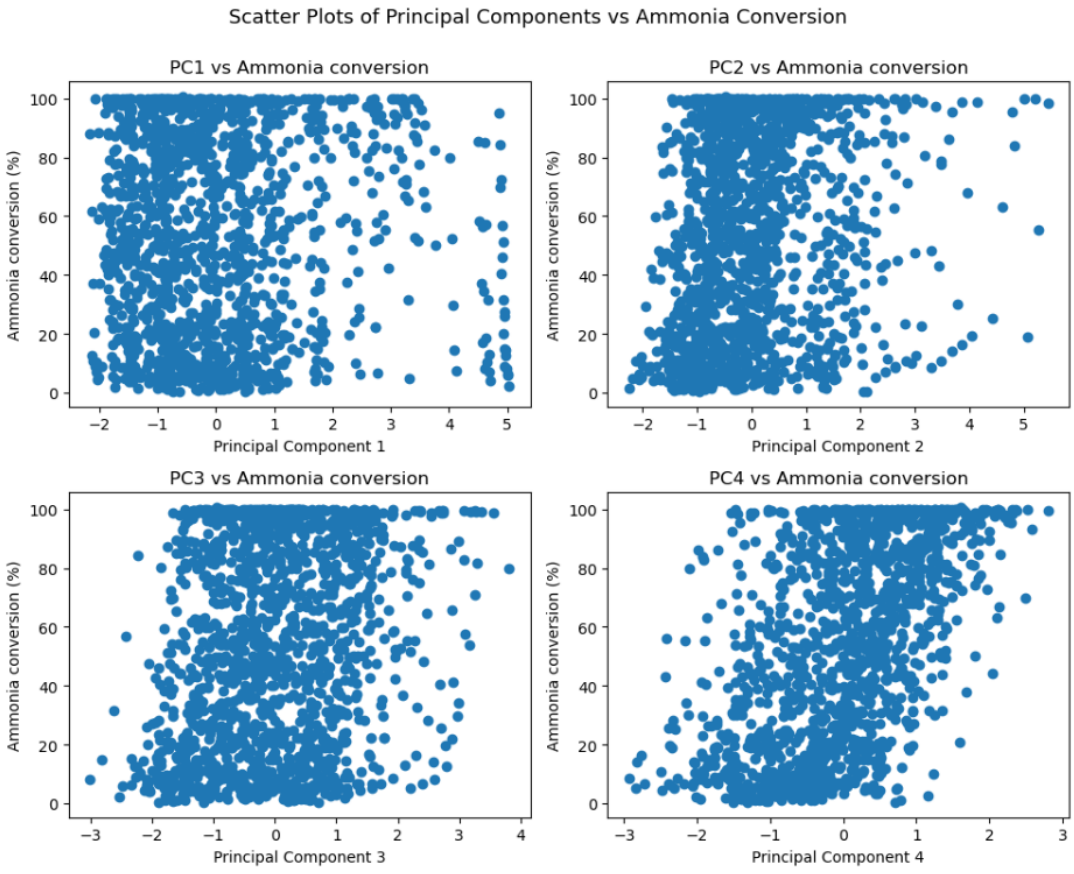

Fig5: Principal Components vs Ammonia Conversion scatter plots

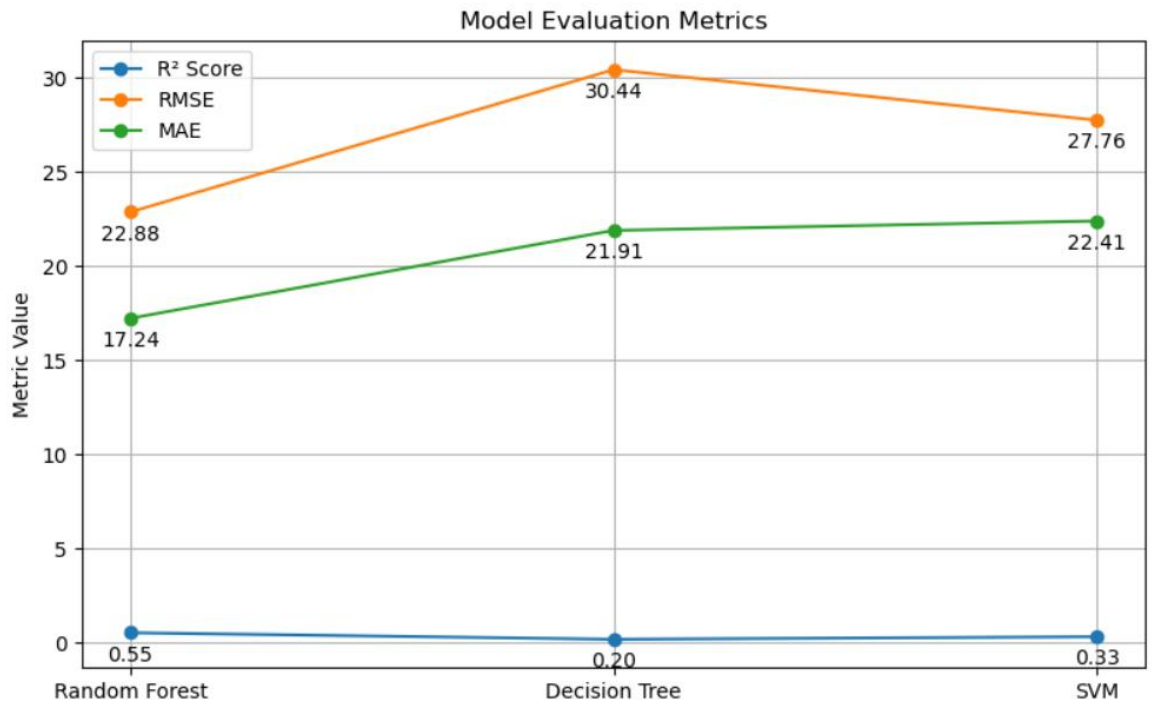

Fig6: Evaluation metrics for all three models

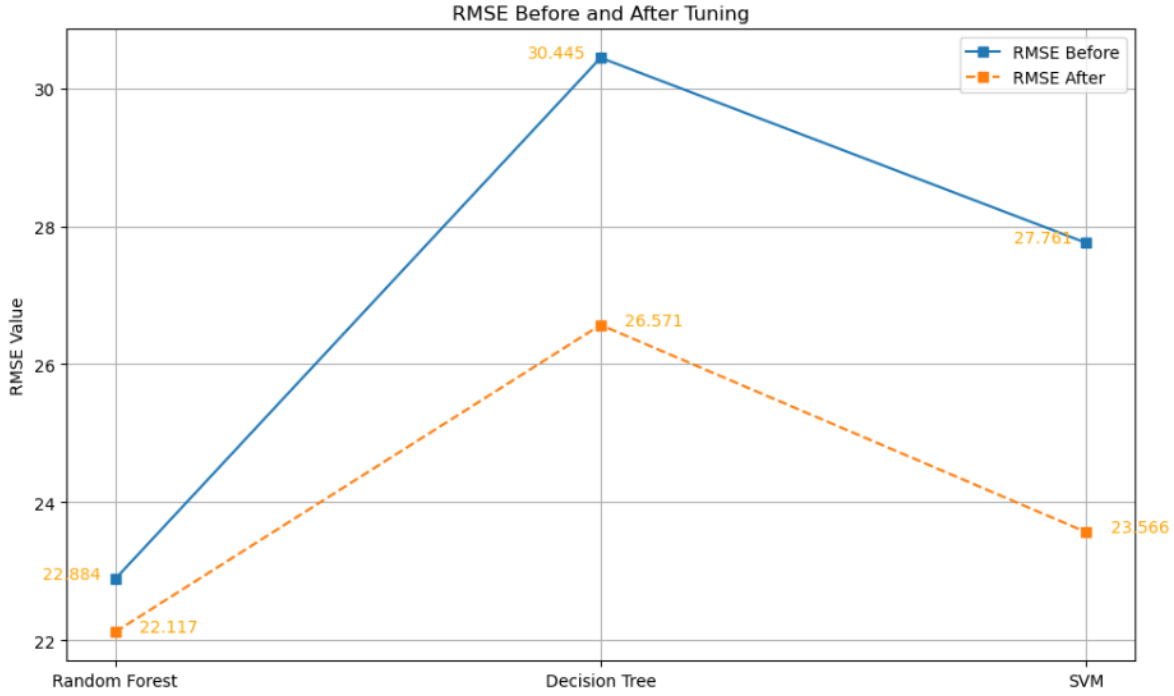

Fig7: Improvements in RMSE after Hyperparameter Tuning

10 Auxiliaries

Data Source: https://www.sciencedirect.com/science/article/pii/S2352152X24012738.

Python file:

https://drive.google.com/drive/folders/1gJaacOmFiuU6xJQuR9pSB2uNzp1opoqC?usp=drive_link