International Journal of Advances in Data and Information Systems
Vol. 5, No. 2, October 2024, pp. 216~228
ISSN: 2721-3056, DOI: 10.59395/ijadis.v5i2.1341

Predicting Methanol Space-Time Yield from CO_2 Hydrogenation Using Machine Learning: Statistical Evaluation of Penalized Regression Techniques

Harun Al Azies[1,2], Muhamad Akrom[1,2], Setyo Budi[3,2], Gustina Alfa Trisnapradika[1,2], Aprilyani Nur Safitri[1,2]
[1]Study Program in Informatics Engineering, Faculty of Computer Science, Universitas Dian Nuswantoro, Indonesia
[2]Research Center for Quantum Computing and Materials Informatics, Faculty of Computer Science, Universitas Dian Nuswantoro, Indonesia
[3]Study Program in Information Systems, Faculty of Computer Science, Universitas Dian Nuswantoro, Indonesia

Article Info

Article history:

Received Sep 09, 2024
Revised Sept 29, 2024
Accepted Oct 30, 2024

Keywords:

Penalized Regression
Ridge Regression
Methanol Production
CO_2 Hydrogenation
Lasso Regression
Elastic Net Regression

ABSTRACT

This study investigates the effectiveness of machine learning techniques, specifically penalized regression models Ridge Regression, Lasso Regression, and Elastic Net Regression in predicting methanol space-time yield (STY) from CO_2 hydrogenation data. Using a dataset derived from Cu-based catalyst research, the study implemented a comprehensive preprocessing approach, including data cleaning, imputation, outlier removal, and normalization. The models were rigorously evaluated through 10-fold cross-validation and tested on unseen data. Ridge Regression outperformed the other models, achieving the lowest Root Mean Squared Error (RMSE) of 0.7706, Mean Absolute Error (MAE) of 0.5627, and Mean Squared Error (MSE) of 0.5938. In comparison, Lasso and Elastic Net Regression models exhibited higher error metrics. Feature importance analysis revealed that Gas Hourly Space Velocity (GHSV) and Molar Masses of Support significantly influence catalytic activity. These findings suggest that Ridge Regression is a promising tool for accurately predicting methanol production, providing valuable insights for optimizing catalytic processes and advancing sustainable practices in chemical engineering.

Corresponding Author:

Harun Al Azies,
Study Program in Informatics Engineering,
Universitas Dian Nuswantoro,
127 Imam Bonjol Street, Semarang 50131, Indonesia.
Email: harun.alazies@dsn.dinus.ac.id

1. INTRODUCTION

The industrial production of carbon dioxide (CO_2) as a feedstock to manufacture value-added chemicals like methanol has gained interest in recent years due to its potential to ameliorate climate change and reduce dependency on fossil fuels [1]. CO_2, a major greenhouse gas, contributes heavily to global warming [2], and its conversion into useful chemicals, such as methanol[3], offers a sustainable approach to reducing its impact on the environment [4]. Methanol itself has a broad range of applications, including its use in the production of polymers [5], fuels, and various organic compounds, making it an essential component of modern industrial processes [6], [7]. As the global

demand for cleaner energy solutions grows, the need for efficient and sustainable methods of methanol production becomes even more critical [8].

Despite the potential benefits, the CO_2 hydrogenation process to methanol presents significant challenges [9]. The reaction involves a complex interplay of factors, such as catalyst type, temperature, pressure, and reactant ratios, all of which must be optimized to achieve efficient conversion. Copper (Cu)-based catalysts, known for their high activity and selectivity in producing methanol, are widely used for this purpose. However, achieving optimal conditions for maximum methanol yield is complicated due to the nonlinear nature of these interactions [10], [11]. Existing studies often rely on traditional optimization techniques, which fail to fully account for the complexities of these processes, leading to suboptimal performance and inefficiencies [12].

Recent advances in machine learning (ML) have opened new opportunities for improving complex chemical processes such as CO_2 hydrogenation [13]. Machine learning algorithms can model intricate relationships between multiple parameters and identify patterns that may be difficult to detect using conventional methods[14], [15]. In particular, penalized regression techniques, including Ridge Regression, Lasso Regression, and Elastic Net Regression, have proven effective in handling multicollinearity and preventing overfitting[16], [17], which are common issues in predictive modeling [17]. These algorithms introduce penalties to the regression model, encouraging simplicity and improving generalization by selecting the most relevant variables [18], [19]. Ridge Regression minimizes the size of the coefficients[20], while Lasso Regression automatically selects important features by shrinking the less important ones to zero[21]. Elastic Net Regression combines the strengths of both methods to offer a flexible and robust modeling approach [22].

This study addresses the gap in the existing literature by employing advanced machine learning techniques, specifically penalized regression models ridge regression, lasso regression, and elastic net regression to optimize methanol yield from CO_2 hydrogenation using Cu-based catalysts. Current models are often limited by their inability to fully account for the nonlinearities and interactions between process variables, which limits their applicability in real-world scenarios. This study proposes the use of machine learning algorithms, specifically penalized regression models, to improve the predictive accuracy and robustness of methanol yield predictions from CO_2 hydrogenation. By comparing the performance of Ridge, Lasso, and Elastic Net regression models, this research aims to identify the most effective approach for optimizing the CO_2 hydrogenation process.

The objectives of this study are to compare the performance of different penalized regression models and identify the most effective approach for maximizing methanol yield. The innovation of this study lies in its application of advanced machine learning techniques to address the complexities of CO_2 hydrogenation. While previous studies have focused on catalyst development and process optimization through experimental approaches, this research leverages data-driven methods to model and predict the behavior of the process more accurately. By utilizing data-driven techniques, this study not only improves methanol production efficiency but also contributes to the broader goal of developing sustainable energy solutions. The integration of penalized regression techniques offers a new perspective on process optimization that can lead to more informed decisions and better performance in industrial applications.

2. MATERIALS AND METHODS

This study follows a structured approach consisting of four stages: data collection and splitting, preprocessing, modeling, and evaluation. The overall methodology is depicted in Figure 1. These stages are designed to ensure a comprehensive analysis of the dataset and an effective application of machine learning models for predicting methanol space-time yield.

2.1. Data Collection

The dataset used in this study is derived from research by Suvarna, Araujo, and Pérez-Ramírez (2022) on CO_2 to methanol hydrogenation, covering the period from 1996 to 2021 [23]. The data, sourced from Web of Science and Scopus, focuses primarily on Cu-based catalysts, accounting for 55% of the dataset. Key variables include Metal Loading [wt.%], Molar Masses of Support 1 and 2 [g mol^{-1}], Total Molar Mass of Support [g mol^{-1}], Promoter 1 Loading [wt.%], Calcination Temperature [K], Calcination Duration [h], SBET [m^2 g^{-1}] (specific surface area), GHSV [cm^3 h^{-1}

g_cat^{-1}] (gas hourly space velocity), Catalyst Amount [g], Pressure [MPa], and Temperature [K]. The target variable is Methanol Space-Time Yield [gMeOH h^{-1} g_cat^{-1}], which quantifies methanol production per unit of catalyst per hour and serves as the key performance metric for catalyst efficiency. The dataset, consisting of 707 entries, was split into training and testing subsets using an 80:20 ratio [24]. This split allocated 80% of the data for training the models and 20% for testing, allowing for an unbiased evaluation of the model's performance on unseen data [25].

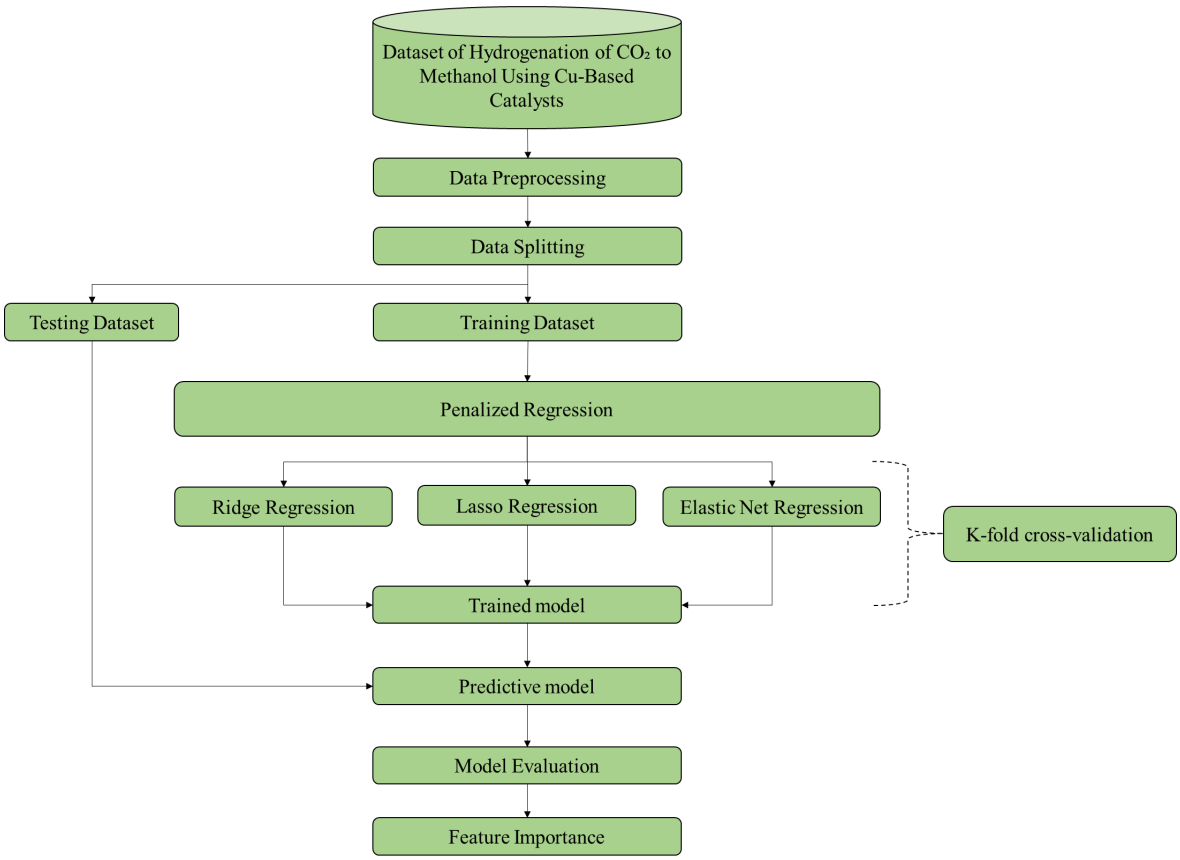


Figure 1. Research Framework for Predicting Methanol Space-Time Yield Using Machine Learning Techniques

2.2. Preprocessing Stage

In the preprocessing stage, the dataset was prepared for modeling through several critical steps [26]. First, data cleaning was performed to identify and resolve any missing values and inconsistencies [27]. Next, imputation techniques were applied to address any remaining missing data, ensuring the completeness of the dataset. Outlier removal was then conducted to eliminate data points that could adversely affect model performance[28], [29]. Finally, normalization and standardization were applied to the data to ensure consistency across all variables [30]. Additionally, Exploratory Data Analysis (EDA) was carried out to assess variable distributions and relationships, which informed the selection of relevant features for modeling [31]. These preprocessing steps collectively enhanced data quality, leading to improved model accuracy and reliability.

2.3. Model Development

In this study, three advanced penalized regression techniques Ridge Regression, Lasso Regression, and Elastic Net Regression were utilized to model methanol space-time yield in the CO_2 hydrogenation process. Each technique was selected for its ability to improve prediction accuracy while managing issues related to overfitting and multicollinearity. Ridge Regression, the first approach, incorporates an L2 regularization term, which applies a penalty to large regression coefficients [20], thereby reducing the risk of overfitting and addressing multicollinearity among

predictors. This regularization technique minimizes the influence of less significant features while preserving all variables in the model.

The second model, Lasso Regression, introduces an L1 regularization term, which not only reduces overfitting but also performs automatic feature selection by shrinking less important coefficients to exactly zero [32]. This makes Lasso particularly useful when dealing with high-dimensional data, where some predictors may be irrelevant or redundant. Lastly, Elastic Net Regression, a hybrid model that combines both L1 (Lasso) and L2 (Ridge) penalties, was employed to provide a balance between feature selection and regularization [33], [34]. By blending the strengths of both Ridge and Lasso, Elastic Net handles datasets with highly correlated predictors more effectively than either method alone.

All three models were trained using 10-fold cross-validation (CV) to ensure a robust evaluation [35]. This approach involved splitting the dataset into 10 subsets, training the model on nine subsets while testing it on the remaining one, and repeating this process iteratively[36]. Cross-validation helps minimize overfitting by ensuring that each data point is used for both training and validation, resulting in a more reliable assessment of model performance across different subsets of the data.

2.4. Model Evaluation

To assess the effectiveness of the penalized regression models developed in this study, a comprehensive evaluation framework was implemented. This evaluation aimed to quantify the accuracy of the models in predicting methanol space-time yield from the CO_2 hydrogenation process, ensuring that the models not only performed well on training data but also generalized effectively to unseen data. By using various error metrics, the evaluation provided insight into the models' ability to minimize prediction errors and capture the underlying relationships between the input features and the target variable. To assess the performance of the penalized regression models, a comprehensive evaluation was conducted using three key error metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Squared Error (MSE). These metrics provide different perspectives on the models' accuracy and help in understanding their predictive capabilities. Each metric was calculated as follows:

1. Root Mean Squared Error

RMSE measures the average magnitude of the prediction errors. It is computed by taking the square root of the mean of the squared differences between actual and predicted values [37]. This metric is sensitive to large errors, making it useful for identifying models that may have significant deviations. The formula for RMSE is:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}} \quad (1)$$

where y_i represents the actual values, y_i denotes the predicted values, and n is the number of observations. A lower RMSE indicates better model performance, with fewer errors.

2. Mean Absolute Error

MAE measures the average magnitude of the absolute differences between actual and predicted values [37]. Unlike RMSE, MAE does not square the errors, which makes it less sensitive to outliers. The formula for MAE is:

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}|}{n} \quad (2)$$

where $|y_i - \hat{y}|$ represents the absolute error for each observation. A lower MAE indicates that the model's predictions are closer to the actual values, providing a straightforward measure of prediction accuracy.

3. Mean Squared Error

MSE calculates the average of the squared differences between actual and predicted values. It provides a measure of the variance of the errors and is particularly useful for identifying models that may have high variance[37]. The formula for MSE is:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2 \quad (3)$$

where $(y_i - \hat{y})^2$ represents the squared error for each observation. Like RMSE, a lower MSE indicates better model performance, with fewer large errors. By employing these metrics, the study ensured a thorough evaluation of the penalized regression models' predictive accuracy, allowing for a well-rounded assessment of their performance in predicting methanol space-time yield in the CO_2 to methanol conversion process.

3. RESULTS AND DISCUSSION

This section outlines the research findings and discusses their implications. It includes an analysis of model performance, highlighting the accuracy and effectiveness of the employed regression techniques. The discussion also delves into the significance of key factors influencing methanol space-time yield.

3.1. Exploratory Data Analysis Results

The EDA provides an overview of the dataset, helping to identify key patterns and potential anomalies that may affect methanol space-time yield in CO_2 hydrogenation using Cu-based catalysts. Figure 2 presents violin plots illustrating the distribution of important variables, such as Metal Loading, Support Molar Weights, Promoter Loading, and Reaction Conditions. This analysis highlights data skewness and variability, offering insights into the factors that may influence methanol production and guiding the development of penalized regression models.

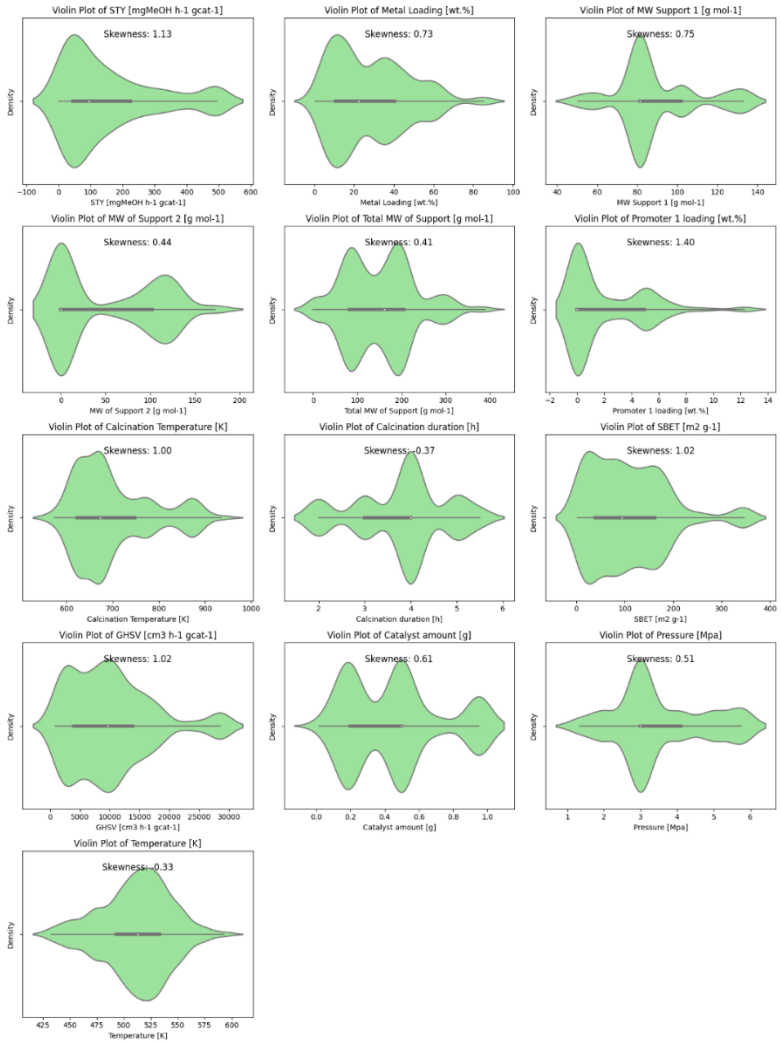


Figure 2. Violin Plots of Key Variables in CO_2 Hydrogenation to Methanol Using Cu-Based Catalysts

The violin plots in Figure 1 provide insights into the distribution of key variables relevant to CO_2 hydrogenation to methanol using Cu-based catalysts. Methanol Space-Time Yield shows a right-skewed distribution, indicating that while most experiments had lower yields, a few had significantly higher yields, potentially revealing optimal conditions or superior catalyst formulations. The skewness in Metal Loading and the molar weights of Support 1 and Support 2 suggests that most experiments used lower values, with higher values possibly linked to better catalytic performance. Promoter 1 Loading is notably skewed, with most data points at low levels and a few at higher levels, suggesting that higher promoter loadings may be crucial for optimizing catalyst performance. The moderate positive skewness in Calcination Temperature and SBET points to standard conditions being common, with some instances of higher values potentially enhancing catalytic properties. Variables like Gas Hourly Space Velocity show right skewness, reflecting typical gas flow conditions with a few high values possibly improving mass transfer rates. In contrast, Calcination Duration and Temperature exhibit mild negative skewness, indicating that longer durations and higher temperatures are more commonly applied, likely for better catalyst stability and reaction rates. Understanding these distribution patterns helps in developing accurate predictive models, where penalized regression techniques like Ridge, Lasso, and Elastic Net address non-normal distributions, minimize overfitting, and enhance model robustness.

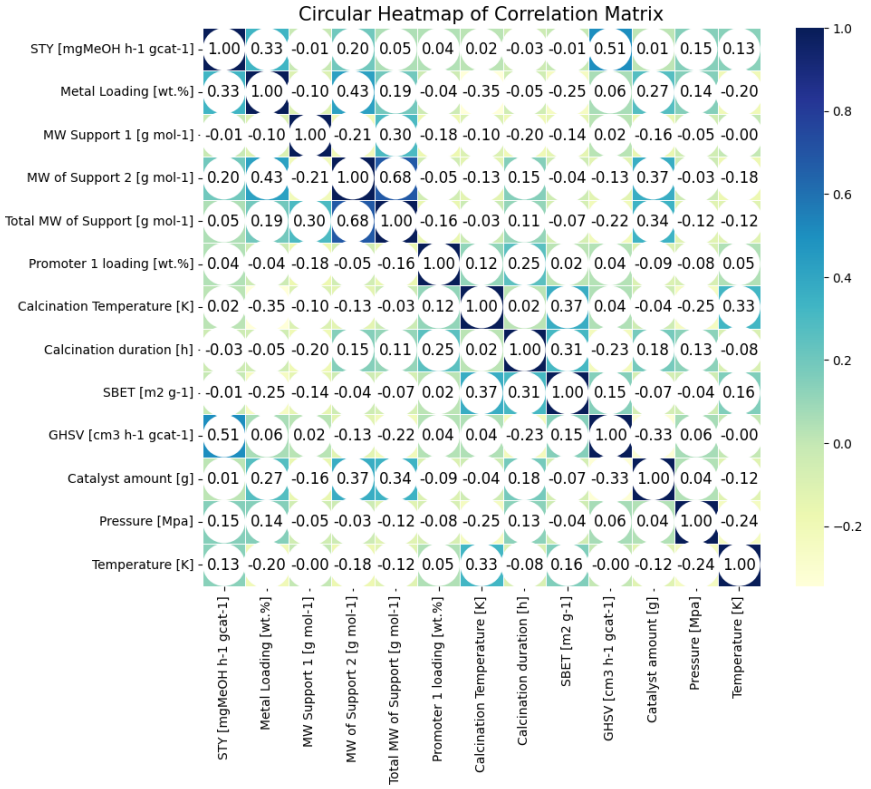


Figure 3. Heatmap of Correlation Coefficients Between Methanol Space-Time Yield and Experimental Variables

Following the exploratory data analysis, which provided insights into the distribution and characteristics of the dataset, a correlation analysis was conducted to further understand the relationships between Methanol Space-Time Yield [mgMeOH h^{-1} g_cat^{-1}] and various experimental variables. This analysis aimed to quantify how each variable influences methanol yield, providing a deeper understanding of the factors that impact the efficiency of CO_2 hydrogenation processes. Figure 3 displays the heatmap of these correlation coefficients, illustrating both the strength and direction of the relationships between STY and other variables. The analysis reveals that Metal Loading [wt.%] has a moderate positive correlation of 0.329 with STY, suggesting that higher metal loading is associated with increased methanol yields, which implies that optimizing metal content could enhance catalyst performance.Conversely, variables such as Molar Weight of Support 1 [g

mol⁻¹] and Total Molar Weight of Support [g mol⁻¹] exhibit negligible effects on methanol yield, with correlations of -0.011 and 0.053, respectively. This indicates that changes in these variables do not significantly impact STY. Molar Weight of Support 2 [g mol⁻¹] and Promoter 1 Loading [wt.%] show weak positive correlations of 0.199 and 0.039, respectively, suggesting minimal influence on methanol yield.

Calcination Temperature [K] and Calcination Duration [h] have very weak correlations (0.018 and -0.034, respectively), indicating their minimal impact on methanol production. The Specific Surface Area (SBET) [m^2 g^{-1}] also shows a minimal negative correlation of -0.007, reflecting that changes in surface area have little effect on methanol yield. Gas Hourly Space Velocity [cm^3 h^{-1} g_cat^{-1}] demonstrates a strong positive correlation of 0.506, highlighting its significant influence on methanol production efficiency. Catalyst Amount [g], Pressure [MPa], and Temperature [K] show very weak positive correlations (0.012, 0.145, and 0.135, respectively), indicating a modest influence on methanol yield.

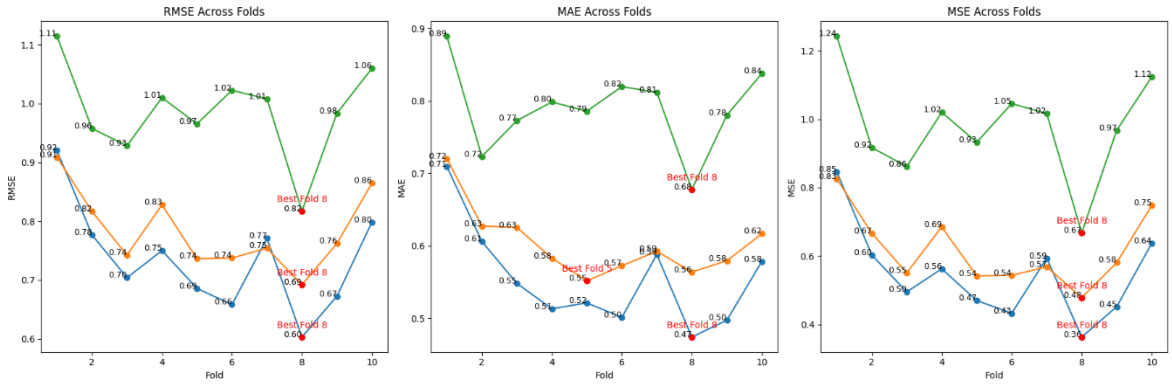
3.2. Model Development Phase

In the model development phase, penalized regression techniques—Ridge Regression, Lasso Regression, and Elastic Net Regression—were employed to predict STY from CO_2 hydrogenation data. Each model underwent a rigorous evaluation using 10-fold cross-validation to ensure robustness and generalizability. The results of the cross-validation are summarized in Table 1.

Table 1. Performance Metrics of Penalized Regression Models

Variable	Statistics	RMSE	MAE	MSE
Ridge Regression	Mean	0.7339	0.5538	0.5459
	Stdev	0.0849	0.0661	0.1295
Lasso Regression	Mean	0.7841	0.6034	0.6190
	Stdev	0.0641	0.0461	0.1029
Elastic Net Regression	Mean	0.9867	0.7897	0.9792
	Stdev	0.0757	0.0557	0.1463

Ridge Regression emerged as the most effective model, achieving a RMSE of 0.7339, a MAE of 0.5538, and a MSE of 0.5459. These metrics indicate that Ridge Regression provided a balanced approach to model complexity and accuracy, yielding reliable predictions for methanol production. Lasso Regression showed slightly less accuracy with a mean RMSE of 0.7841, mean MAE of 0.6034, and mean MSE of 0.6190. Although Lasso Regression is known for its feature selection capabilities, its more aggressive regularization resulted in marginally higher error metrics compared to Ridge Regression. Elastic Net Regression had the highest error metrics, with a mean RMSE of 0.9867, mean MAE of 0.7897, and mean MSE of 0.9792. Despite combining elements of both Ridge and Lasso, its complex regularization framework did not perform as well in this context.



(a) (b) (c)

Figure 4. Performance Metrics of Penalized Regression Models Across 10-Fold Cross-Validation: (a) RMSE, (b) MAE, (c) MSE

Figure 4 presents these results, displaying the performance metrics across each fold of the cross-validation process. Fold 8 emerged as the most consistently effective across all models. Consequently, the Ridge Regression model trained with data from Fold 8 was selected for predicting methanol yield on the testing data. This choice ensures the most reliable and accurate predictions, contributing to the optimization of methanol production processes from CO_2 hydrogenation and supporting advancements in sustainable chemical manufacturing.

3.3. Model Evaluation

Based on the model development phase, the performance of the penalized regression models—Ridge Regression, Lasso Regression, and Elastic Net Regression—was further assessed on the test set, using the best-performing fold from the cross-validation. The evaluation results are summarized in Table 2. Ridge Regression, with an RMSE of 0.7706, MAE of 0.5627, and MSE of 0.5938, emerged as the most reliable model for predicting Methanol STY. This performance indicates that Ridge Regression provides a robust and consistent prediction of methanol production efficiency. Its lower error metrics suggest that it effectively balances the complexity of the model with predictive accuracy, making it well-suited for practical applications in optimizing catalyst performance for CO_2 hydrogenation.

Table 2. Evaluation Results on the Test Set

Model	RMSE	MAE	MSE
Ridge Regression	0.7706	0.5627	0.5938
Lasso Regression	0.8419	0.6416	0.7087
Elastic Net Regression	1.0107	0.7907	1.0214

In contrast, Lasso Regression showed an RMSE of 0.8419, MAE of 0.6416, and MSE of 0.7087. Although Lasso Regression excels in feature selection by shrinking less relevant coefficients to zero, its performance on the test set was slightly less accurate than Ridge Regression. This outcome suggests that while Lasso's feature selection is valuable, its more aggressive regularization resulted in higher prediction errors. This could imply that in this context, reducing model complexity did not translate into improved performance. Elastic Net Regression, which combines elements of both Ridge and Lasso regularization, had the highest error metrics, with an RMSE of 1.0107, MAE of 0.7907, and MSE of 1.0214. Despite its ability to address multicollinearity and perform feature selection, the complex regularization framework of Elastic Net did not perform as effectively as Ridge Regression. This suggests that the model's added complexity may not have been beneficial for this particular dataset and prediction task. Ridge Regression proved to be the most effective model for predicting methanol space-time yield from CO_2 hydrogenation data. Its performance underscores its suitability for scenarios where both accuracy and model stability are crucial. The higher error metrics observed in Lasso and Elastic Net Regression highlight the challenges of feature selection and regularization in this context, suggesting that simpler models like Ridge Regression may offer better practical performance for optimizing catalyst efficiency.

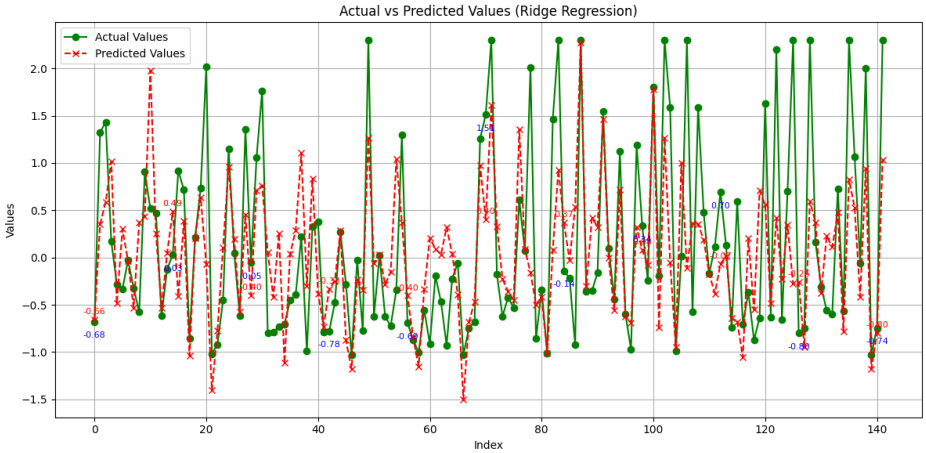


Figure 5. Comparison of Actual and Predicted Methanol Space-Time Yield for Ridge Regression Model

The performance of the Ridge Regression model was further evaluated through a visual comparison of actual versus predicted Methanol Space-Time Yield values. Figure 4 illustrates this comparison, presenting a scatter plot where each point represents an actual and predicted STY value pair. In Figure 5, the diagonal line serves as a reference for perfect prediction, with actual values plotted against their predicted counterparts. A strong alignment of points along this line would indicate high predictive accuracy. However, the plot reveals a mix of scenarios: some points are closely aligned with the diagonal line, suggesting that the model performs well in those regions, while other points show noticeable deviations. Certain areas of the plot exhibit clusters of points that are tightly packed along the diagonal, reflecting accurate predictions and indicating that the Ridge Regression model effectively captures the trends in methanol yield for those data points. In contrast, there are regions where points are dispersed further from the diagonal, suggesting discrepancies between actual and predicted values. These deviations highlight areas where the model's performance may be less reliable, potentially indicating regions with complex underlying patterns or outliers that the model struggles to capture.

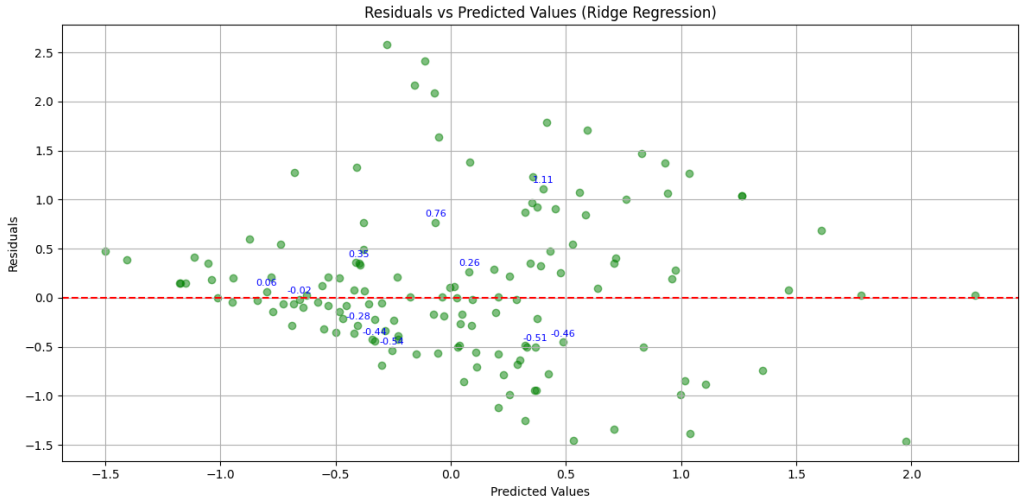


Figure 6. Residual Plot for Ridge Regression Model: Distribution of Residuals versus Predicted Values

Figure 6 presents the residual plot for the Ridge Regression model, offering a detailed view of the relationship between predicted values and residuals. In this plot, residuals—representing the differences between actual values and model predictions—are displayed against the predicted values. The plot shows that most residuals are clustered around zero, indicating that the Ridge Regression model captures the underlying patterns in the data effectively. This distribution suggests that the model's predictions are generally accurate and that it successfully minimizes prediction errors without any discernible systematic bias. However, the plot also reveals some deviations from this trend. A few points display larger residuals, suggesting instances where the model's predictions differ significantly from the actual values. These deviations might highlight areas where the model could be overfitting or underfitting or where additional features might enhance the model's accuracy.

3.4. Determinants of Catalytic Efficiency in Methanol Production

Following the evaluation of model performance, an analysis was performed to identify the features that most significantly influence catalytic activity in methanol production. This feature importance analysis, derived from the Ridge Regression model, offers valuable insights into the key variables impacting the efficiency of methanol synthesis. Figure 7 highlights the importance of various factors in the catalytic process.

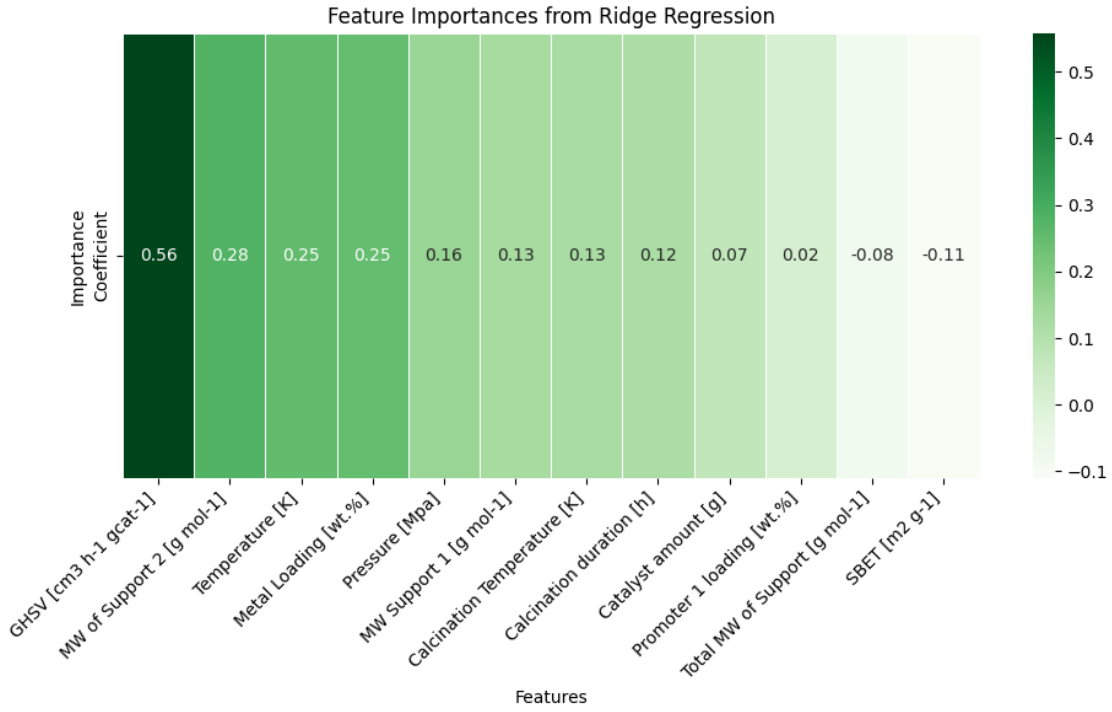


Figure 7. Feature Importance Analysis for Catalytic Activity in Methanol Production

Gas Hourly Space Velocity (GHSV) [cm³ h^{-1} g_cat^{-1}] emerged as the most influential variable, with a coefficient of 0.5571. This significant positive impact suggests that higher GHSV improves the methanol production rate, making it a crucial parameter for optimizing catalyst performance. Molar Weight of Support 2 [g mol^{-1}] and Temperature [K] also play substantial roles, with coefficients of 0.2809 and 0.2515, respectively. Their considerable influence indicates that these factors are vital in enhancing methanol production efficiency, potentially by affecting the catalyst's stability and activity under varying operational conditions.

Metal Loading [wt.%] follows with a coefficient of 0.2481, reinforcing its importance in the catalytic process. A higher metal loading generally enhances the catalyst's ability to facilitate the methanol synthesis reaction. Additional factors such as Pressure [MPa] and Molar Weight of Support 1 [g mol^{-1}] contribute to the model's predictions with coefficients of 0.1561 and 0.1280. These variables, while important, have a somewhat lesser impact compared to GHSV and Temperature. Calcination Temperature [K] and Calcination Duration [h] also influence the catalytic activity, though to a moderate extent, with coefficients of 0.1254 and 0.1174. Catalyst Amount [g] shows a minor but positive effect with a coefficient of 0.0729, indicating a smaller but still relevant contribution to methanol production.

Conversely, Promoter 1 Loading [wt.%], Total Molar Weight of Support [g mol^{-1}], and Specific Surface Area (SBET) [m^2 g^{-1}] have relatively minor or negative impacts. This suggests that these factors are less critical for optimizing catalytic performance in methanol production within the context of this study.The analysis underscores that GHSV, Molar Weight of Support 2, and Temperature are pivotal factors for enhancing catalytic activity. These insights are crucial for refining methanol production processes and improving catalyst design and operation in industrial applications.

4. CONCLUSION

This study aimed to enhance the understanding of factors affecting Methanol Space-Time Yield and to demonstrate the effectiveness of penalized regression models in predicting methanol production from CO_2 hydrogenation. The findings highlight that the Ridge Regression model outperformed other models, effectively capturing the intricate relationships between experimental variables and methanol yield.

Key factors influencing catalytic activity were identified, particularly Gas Hourly Space Velocity (GHSV) and Molar Weight of Support 2, which play crucial roles in optimizing methanol production processes. Additionally, the research emphasizes the importance of understanding these variables to refine catalytic strategies and enhance production efficiency.

Looking forward, future research should focus on integrating additional variables and exploring alternative machine learning approaches to further refine predictive accuracy. Moreover, applying these insights to real-world catalytic systems could lead to significant improvements in industrial methanol production processes. Expanding research to include diverse catalytic environments or operational conditions will enhance the applicability and drive further advancements in the field.

ACKNOWLEDGEMENTS

This article is the research result funded by Universitas Dian Nuswantoro under Hibah Penelitian Pemula Perguruan Tinggi (Grant:076/A.38-04/UDN-09/VII/2024). Furthermore, the authors would also like to thank Universitas Dian Nuswantoro for the continuous support in completing this study.

REFERENCES

[1] A. AlNouss, G. Mckay, and T. Al-Ansari, "Utilisation of Carbon Dioxide and Gasified Biomass for the Generation of Value Added Products," *Computer Aided Chemical Engineering*, vol. 50, pp. 1567–1572, Jan. 2021, doi: 10.1016/B978-0-323-88506-5.50242-4.

[2] M. N. Anwar *et al.*, "CO2 utilization: Turning greenhouse gas into fuels and valuable products," *J Environ Manage*, vol. 260, p. 110059, Apr. 2020, doi: 10.1016/J.JENVMAN.2019.110059.

[3] T. Patil, A. Naji, U. Mondal, I. Pandey, A. Unnarkat, and S. Dharaskar, "Sustainable methanol production from carbon dioxide: advances, challenges, and future prospects," *Environmental Science and Pollution Research 2024 31:32*, vol. 31, no. 32, pp. 44608–44648, Jul. 2024, doi: 10.1007/S11356-024-34139-3.

[4] A. Saravanan *et al.*, "A comprehensive review on different approaches for CO2 utilization and conversion pathways," *Chem Eng Sci*, vol. 236, p. 116515, Jun. 2021, doi: 10.1016/J.CES.2021.116515.

[5] S. S. Tabibian and M. Sharifzadeh, "Statistical and analytical investigation of methanol applications, production technologies, value-chain and economy with a special focus on renewable methanol," *Renewable and Sustainable Energy Reviews*, vol. 179, p. 113281, Jun. 2023, doi: 10.1016/J.RSER.2023.113281.

[6] A. Sonthalia, N. Kumar, M. Tomar, V. Edwin Geo, S. Thiyagarajan, and A. Pugazhendhi, "Moving ahead from hydrogen to methanol economy: scope and challenges," *Clean Technol Environ Policy*, vol. 25, no. 2, pp. 551–575, Mar. 2023, doi: 10.1007/S10098-021-02193-X/METRICS.

[7] Z. Tian, Y. Wang, X. Zhen, and Z. Liu, "The effect of methanol production and application in internal combustion engines on emissions in the context of carbon neutrality: A review," *Fuel*, vol. 320, p. 123902, Jul. 2022, doi: 10.1016/J.FUEL.2022.123902.

[8] A. Ullah, N. A. Hashim, M. F. Rabuni, and M. U. Mohd Junaidi, "A Review on Methanol as a Clean Energy Carrier: Roles of Zeolite in Improving Production Efficiency," *Energies 2023, Vol. 16, Page 1482*, vol. 16, no. 3, p. 1482, Feb. 2023, doi: 10.3390/EN16031482.

[9] F. Sha, Z. Han, S. Tang, J. Wang, and C. Li, "Hydrogenation of Carbon Dioxide to Methanol over Non−Cu-based Heterogeneous Catalysts," *ChemSusChem*, vol. 13, no. 23, pp. 6160–6181, Dec. 2020, doi: 10.1002/CSSC.202002054.

[10] M. Ren, Y. Zhang, X. Wang, and H. Qiu, "Catalytic Hydrogenation of CO2 to Methanol: A Review," *Catalysts 2022, Vol. 12, Page 403*, vol. 12, no. 4, p. 403, Apr. 2022, doi: 10.3390/CATAL12040403.

[11] C. Wu *et al.*, "Inverse ZrO2/Cu as a highly efficient methanol synthesis catalyst from CO2 hydrogenation," *Nature Communications 2020 11:1*, vol. 11, no. 1, pp. 1–10, Nov. 2020, doi: 10.1038/s41467-020-19634-8.

[12] M. B. Gawande *et al.*, "Cu and Cu-Based Nanoparticles: Synthesis and Applications in Catalysis," *Chem Rev*, vol. 116, no. 6, pp. 3722–3811, Mar. 2016, doi: 10.1021/ACS.CHEMREV.5B00482/ASSET/IMAGES/LARGE/CR-2015-004823_0039.JPEG.

[13] E. G. Aklilu and T. Bounahmidi, "Machine learning applications in catalytic hydrogenation of carbon dioxide to methanol: A comprehensive review," *Int J Hydrogen Energy*, vol. 61, pp. 578–602, Apr. 2024, doi: 10.1016/J.IJHYDENE.2024.02.309.

[14] M. Shehab *et al.*, "Machine learning in medical applications: A review of state-of-the-art methods," *Comput Biol Med*, vol. 145, p. 105458, Jun. 2022, doi: 10.1016/J.COMPBIOMED.2022.105458.
[15] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput Sci*, vol. 2, no. 3, pp. 1–21, May 2021, doi: 10.1007/S42979-021-00592-X/FIGURES/11.
[16] C. J. Greenwood *et al.*, "A comparison of penalised regression methods for informing the selection of predictive markers," *PLoS One*, vol. 15, no. 11, p. e0242730, Nov. 2020, doi: 10.1371/JOURNAL.PONE.0242730.
[17] U. Sharma, N. Gupta, and M. Verma, "Prediction of compressive strength of GGBFS and Flyash-based geopolymer composite by linear regression, lasso regression, and ridge regression," *Asian Journal of Civil Engineering*, vol. 24, no. 8, pp. 3399–3411, Dec. 2023, doi: 10.1007/S42107-023-00721-2/METRICS.
[18] Q. Chen, B. Xue, and M. Zhang, "Rademacher Complexity for Enhancing the Generalization of Genetic Programming for Symbolic Regression," *IEEE Trans Cybern*, vol. 52, no. 4, pp. 2382–2395, Apr. 2022, doi: 10.1109/TCYB.2020.3004361.
[19] M. Nicolau and A. Agapitos, "Choosing function sets with better generalisation performance for symbolic regression models," *Genet Program Evolvable Mach*, vol. 22, no. 1, pp. 73–100, Mar. 2021, doi: 10.1007/S10710-020-09391-4/METRICS.
[20] M. Arashi, M. Roozbeh, N. A. Hamzah, and M. Gasparini, "Ridge regression and its applications in genetic studies," *PLoS One*, vol. 16, no. 4, p. e0245376, Apr. 2021, doi: 10.1371/JOURNAL.PONE.0245376.
[21] M. Hamada, J. J. Tanimu, M. Hassan, H. A. Kakudi, and P. Robert, "Evaluation of Recursive Feature Elimination and LASSO Regularization-based optimized feature selection approaches for cervical cancer prediction," *Proceedings - 2021 IEEE 14th International Symposium on Embedded Multicore/Many-Core Systems-on-Chip, MCSoC 2021*, pp. 333–339, 2021, doi: 10.1109/MCSOC51149.2021.00056.
[22] J. K. Tay, B. Narasimhan, and T. Hastie, "Elastic Net Regularization Paths for All Generalized Linear Models," *J Stat Softw*, vol. 106, 2023, doi: 10.18637/JSS.V106.I01.
[23] M. Suvarna, T. P. Araújo, and J. Pérez-Ramírez, "A generalized machine learning framework to predict the space-time yield of methanol from thermocatalytic CO2 hydrogenation," *Appl Catal B*, vol. 315, p. 121530, Oct. 2022, doi: 10.1016/J.APCATB.2022.121530.
[24] V. R. Joseph and A. Vakayil, "SPlit: An Optimal Method for Data Splitting," *Technometrics*, vol. 64, no. 2, pp. 166–176, 2022, doi: 10.1080/00401706.2021.1921037/SUPPL_FILE/UTCH_A_1921037_SM8231.PDF.
[25] V. Roshan, J. H. M. Stewart, R. Joseph, and H. M. Stewart, "Optimal ratio for data splitting," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 15, no. 4, pp. 531–538, Aug. 2022, doi: 10.1002/SAM.11583.
[26] J. Luengo, D. García-Gil, S. Ramírez-Gallego, S. García, and F. Herrera, "Big Data Preprocessing: Enabling Smart Data," *Big Data Preprocessing: Enabling Smart Data*, pp. 1–186, Jan. 2020, doi: 10.1007/978-3-030-39105-8/COVER.
[27] A. A. Dharmasaputro, N. M. Fauzan, M. Kallista, I. P. D. Wibawa, and P. D. Kusuma, "Handling Missing and Imbalanced Data to Improve Generalization Performance of Machine Learning Classifier," *2021 International Seminar on Machine Learning, Optimization, and Data Science, ISMODE 2021*, pp. 140–145, 2022, doi: 10.1109/ISMODE53584.2022.9743022.
[28] R. Dawson, "How Significant is a Boxplot Outlier?," *Journal of Statistics Education*, vol. 19, no. 2, 2011, doi: 10.1080/10691898.2011.11889610.
[29] R. C. Pfaffenberger and T. E. Dielman, "A Comparison of Regression Estimators When Both Multicollinearity and Outliers Are Present," *Robust Regression*, pp. 243–270, May 2019, doi: 10.1201/9780203740538-13.
[30] L. Huang, J. Qin, Y. Zhou, F. Zhu, L. Liu, and L. Shao, "Normalization Techniques in Training DNNs: Methodology, Analysis and Application," *IEEE Trans Pattern Anal Mach Intell*, vol. 45, no. 8, pp. 10173–10196, Aug. 2023, doi: 10.1109/TPAMI.2023.3250241.
[31] R. Indrakumari, T. Poongodi, and S. R. Jena, "Heart Disease Prediction using Exploratory Data Analysis," *Procedia Comput Sci*, vol. 173, pp. 130–139, Jan. 2020, doi: 10.1016/J.PROCS.2020.06.017.
[32] Y. Wang *et al.*, "Regression with adaptive lasso and correlation based penalty," *Appl Math Model*, vol. 105, pp. 179–196, May 2022, doi: 10.1016/J.APM.2021.12.016.
[33] S. Srivatsaan, A. Sankar, and M. Karthikeyan, "Impact Of Elastic Net and Lasso Regularization Techniques on the NHANES Dataset," *AIP Conf Proc*, vol. 3075, no. 1, Jul. 2024, doi: 10.1063/5.0217034/3305152.

[34] M. Hajihosseinlou, A. Maghsoudi, and R. Ghezelbash, "Regularization in machine learning models for MVT Pb-Zn prospectivity mapping: applying lasso and elastic-net algorithms," *Earth Sci Inform*, pp. 1–15, Aug. 2024, doi: 10.1007/S12145-024-01404-5/METRICS.

[35] S. Bates, T. Hastie, and R. Tibshirani, "Cross-Validation: What Does It Estimate and How Well Does It Do It?," *https://doi.org/10.1080/01621459.2023.2197686*, 2023, doi: 10.1080/01621459.2023.2197686.

[36] H. Al Azies, N. Ariyanto, and I. B. Dikaputra, "Data-Driven Analytical Model Using Machine Learning Algorithms," *International Journal of Advances in Data and Information Systems*, vol. 5, no. 1, pp. 1–11, Mar. 2024, doi: 10.59395/IJADIS.V5I1.1309.

[37] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Comput Sci*, vol. 7, pp. 1–24, Jul. 2021, doi: 10.7717/PEERJ-CS.623/SUPP-1.