# The Impact of Data Preprocessing Order on LASSO and Elastic Net Capabilities

**Geneveive Yii Ven Tang**
Department of Mathematics and Statistics, Faculty of Applied Science and Technology, University Tun Hussein Onn Malaysia, Campus Pagoh, Muar, Johor, Malaysia
geneveivetang0806@gmail.com

**Khuneswari Gopal Pillay**
Department of Mathematics and Statistics, Faculty of Applied Science and Technology, University Tun Hussein Onn Malaysia, Campus Pagoh, Muar, Johor, Malaysia
khuneswari@uthm.edu.my (corresponding author)

**Aida Mustapha**
Department of Mathematics and Statistics, Faculty of Applied Science and Technology, University Tun Hussein Onn Malaysia, Campus Pagoh, Muar, Johor, Malaysia
aidam@uthm.edu.my

*Received: 14 November 2024 | Revised: 15 December 2024 and 29 December 2024 | Accepted: 1 January 2025*

*Licensed under a CC-BY 4.0 license | Copyright (c) by the authors | DOI: https://doi.org/10.48084/etasr.9611*

## ABSTRACT

**The Food Security Index (FSI) evaluates affordability, accessibility, utilization, and food availability. However, previous research on food security in Malaysia has primarily focused on production, neglecting a detailed analysis of economic factors. The Overnight Policy Rate (OPR), set by Bank Negara Malaysia (BNM), regulates economic activity by controlling the interest rate at which commercial banks borrow and lend overnight. This study explores the impact of data preprocessing sequences on the performance of LASSO and Elastic Net regression models in predicting Malaysia's FSI. Using macroeconomic data from 2010 to 2023, this study evaluates the effects of different sequences of outlier detection and missing data imputation. The findings reveal that the LASSO model achieves the highest accuracy and the lowest error rates with outlier detection performed after imputation. This study underscores the importance of preprocessing order in enhancing model reliability and provides insight into the economic factors that influence food security in Malaysia. The results show that OPR reduces Malaysia's FSI by 0.151 units, while inflation increases it by 0.022. The LASSO regression model offers a novel perspective on the economic factors influencing food security, providing a more comprehensive understanding of food security in Malaysia.**

***Keywords-preprocessing sequences; missing data; outliers; LASSO; Elastic Net***

## I. INTRODUCTION

The United Nations Committee on World Food Security (WFS) defines food security as the adequate physical, social, and economic access to safe and nutritious food required for a healthy and active life [1]. Ensuring food security has become increasingly complex due to rapid population growth, fluctuating energy prices, and volatile exchange rates. These challenges are directly related to Sustainable Development Goal (SDG) 2: Zero Hunger, which aims to end hunger, achieve food security, and improve nutrition by 2030 [2]. Addressing these issues is critical for building resilient and sustainable food systems, particularly in regions such as Malaysia.

In Malaysia, macroeconomic variables directly and indirectly influence food security indices, including the Overnight Policy Rate (OPR), inflation, production costs, exchange rate, and energy price fluctuations. There is a significant relationship between the exchange rate and global oil prices. High energy prices exerted greater upward pressure on local food prices due to increases in processing costs, transportation, and food distribution [3]. As these costs rise, companies typically pass on a portion of the increased expenses to consumers by raising selling prices. Additionally, high local demand further drives up commodity prices. When demand remains high while supply is constrained, this imbalance causes further price increases [4]. The combined effect of these factors increases the cost of related goods, exacerbating the overall financial pressure on consumers. These dynamics contribute to a broader inflationary environment, with food security indices directly reflecting these changes in prices.

In response to such inflationary pressures, many countries, including Malaysia, can attempt to stimulate the purchasing power of citizens by increasing the money supply [5]. However, excessive money printing can decrease the value of money, fueling even higher inflation. One of the policy responses involves the adjustment of the OPR, which serves as a tool to control the money supply and stabilize economic conditions [6]. Changes in OPR affect consumer spending behavior and food price inflation. A higher OPR typically encourages saving and overspend, reducing aggregate demand and alleviating price pressures on food items [6]. However, the effects of adjustments in the money supply and interest rates on food prices are not immediate. Studies have shown that these impacts may take between 8 and 16 months to manifest fully, suggesting that policy interventions aimed at stabilizing food prices require a longer-term perspective [5].

Given these complexities, a comprehensive analysis of macroeconomic factors, including OPR and inflation, is crucial to formulate effective policy recommendations. Understanding the intricate relationships between these variables can guide decision-makers in balancing economic growth with food security. Theoretical expectations, based on economic models of inflation and monetary policy, suggest that targeted adjustments on OPR could help mitigate food price inflation in the medium to long term, benefiting both consumers and producers in Malaysia's evolving economic environment.

To effectively analyze these macroeconomic factors, data from multiple sources are often utilized. However, real-world datasets typically present various challenges, such as missing observations, outliers, noise, inconsistencies, and multicollinearity [7]. These imperfections can significantly compromise the accuracy and reliability of predictive models if not addressed properly [8]. Preprocessing techniques such as Multiple Imputation Chained Equations (MICE) and robust outlier detection are widely used to mitigate these issues and enhance model performance. Despite their importance, existing research often focuses on individual preprocessing methods while overlooking the impact of the sequence in which these steps are applied [9].

Ensuring data quality is a crucial aspect of preprocessing and involves steps such as handling missing values, addressing outliers, smoothing noisy data, and resolving inconsistencies [8]. Addressing missing values is particularly critical, as unhandled gaps can introduce bias and skew analytical results [10]. Early resolution of missing data is frequently required to preserve the integrity of the dataset because statistical analysis relies on complete datasets for accuracy and representativeness. Addressing missing data before outlier detection can result in imputations that are skewed by the presence of outliers, complicating subsequent analyses. In contrast, detecting and removing outliers before imputation can improve the accuracy of imputed values by mitigating bias. However, the optimal sequence of these preprocessing steps depends on the dataset's characteristics and analytical goals. For example, some studies suggest that imputation-first approaches better preserve data structure in datasets with a high prevalence of outliers [11]. On the other hand, outliers can significantly influence variable selection processes, particularly in regression models such as LASSO, which may compromise prediction reliability [12].

Outliers disrupt the accuracy of imputations, causing the imputed values to deviate substantially from their true values and leading to biased estimates [11]. Identifying and removing outliers before imputation can reduce these distortions, enhancing the reliability of the dataset. However, this approach carries risks, such as potentially losing valuable information due to masking and swamping effects. These effects can obscure the detection of additional outliers or reveal new ones only after removing an initial outlier [13]. These challenges highlight the need for a systematic investigation into how the order of preprocessing steps affects the performance of regression models, particularly in scenarios involving high-dimensional data and complex relationships.

Another prevalent issue in macroeconomic datasets is multicollinearity, where independent variables are highly correlated. This can compromise the precision of coefficient estimates in traditional regression models. Regularization methods, such as LASSO and Elastic Net, address this issue by applying penalties that shrink correlated variables' coefficients to zero [14]. Elastic Net, which combines L1 and L2 penalties, offers a flexible approach to variable selection, particularly in datasets with high multicollinearity. However, the influence of preprocessing sequences, such as the timing of outlier detection and missing data imputation, on the performance of these regularization methods remains underexplored.

This study aims to bridge this gap by systematically evaluating how different preprocessing sequences affect the performance of LASSO and Elastic Net models. Using macroeconomic data from Malaysia, including variables such as OPR, inflation rates, production costs, and FSI, this study examines the effects of outlier detection and missing data imputation on variable selection and model accuracy. The findings contribute to understanding regularized regression techniques and offer practical insights into designing effective preprocessing pipelines for predictive modeling in economic and agricultural studies.

## II. MATERIALS AND METHODOLOGY

### A. *Data Description*

The dataset used in this study consists of one dependent variable, the Malaysia FSI, and 16 independent variables. The data were obtained from the OpenDOSM portal [15] and span monthly observations from 2010 to 2023, resulting in a total of 2,856 data points. Both dependent and independent variables are organized into 168 rows. Table I provides a summary of the variables included in the dataset.

### B. *Data Preprocessing*

#### 1) *Missing Data*

The MICE method was employed to address the issue of missing data [16]. MICE was selected due to its ability to handle different types of variables and its effectiveness in producing multiple plausible datasets, which account for the uncertainty introduced by missing data.

TABLE I. SUMMARY OF DATA VARIABLES

| Variables | Data type | Unit | Dependent or independent variable |
|---|---|---|---|
| Malaysia Food Security Index (*FSI*) | Index | - | Dependent, $Y$ |
| Gross Domestics Product (*GDP*) | Index | - | Independent, $X_1$ |
| Consumer Price Index (*CPI*) | Index | - | Independent, $X_2$ |
| Agricultural Production Cost (*agri*) | Numerical | RM | Independent, $X_3$ |
| Mining Production Cost (*mining*) | Numerical | RM | Independent, $X_4$ |
| Manufacturing Production Cost (*manufac*) | Numerical | RM | Independent, $X_5$ |
| Electricity Production Cost (*electric*) | Numerical | RM | Independent, $X_6$ |
| Water Production Cost (*water*) | Numerical | RM | Independent, $X_7$ |
| Overnight Policy Rate (*OPR*) | Rate | - | Independent, $X_8$ |
| Palm Oil Price (*palm.oil*) | Numerical | MYR per metric ton | Independent, $X_9$ |
| Crude Oil Price (*crude.oil*) | Numerical | MYR per Barrel | Independent, $X_{10}$ |
| Wholesale Price Index (*wpi*) | Index | - | Independent, $X_{11}$ |
| Population Growth (*ppln*) | Rate | - | Independent, $X_{12}$ |
| Unemployment Rate (*unemployment*) | Rate | - | Independent, $X_{13}$ |
| Inflation Rate (*inflation*) | Rate | - | Independent, $X_{14}$ |
| Producer Price Index (*ppi*) | Index | - | Independent, $X_{15}$ |
| Exchange Rate of Ringgit Malaysia against US Dollar (*exchange.rate*) | Rate | - | Independent, $X_{16}$ |

In R, packages known as mice are used to address the problem of missing data [17]. A unique prediction model that is adapted to the distribution of the variable and its interactions with other variables is used to impute each variable with missing data. Predictive Mean Matching (PMM) is applied as an imputation method to ensure that the imputed values are consistent with the observed data. Five datasets were generated and pooled for subsequent analyses to account for the variability in imputed values, improving the robustness of the results.

### 2) *Outliers Detection*

The detection of outliers is performed using the boxplot method [18]. In [18], it was emphasized that using a box plot is especially effective for detecting outliers within multivariate datasets. Any rows containing outlier data points were removed from the dataset [19].

### 3) *Multicollinearity*

The multicollinearity method assesses whether independent variables in the datasets are highly correlated. The Variance Inflation Factors (VIF) is used to quantify the degree of correlation. The interpretation benchmark for VIF values can be described as follows [20]:

- $1 \leq$ VIF $\leq 5$ suggests a modest correlation among independent variables,
- $6 \leq$ VIF $< 10$ indicates a significant correlation among independent variables.
- VIF $\geq 10$ suggests the presence of multicollinearity in the dataset.

## C. *Analysis Flowchart*

Figure 1 illustrates the study's overall analysis flowchart. The entire dataset underwent several preprocessing steps at the beginning of the procedure. After the preprocessing stage, three different datasets were produced, followed by data splitting, achieved by k-fold cross-validation. A second round of k-fold cross-validation was then applied to obtain the optimal regularization parameters $\lambda$, which were subsequently used to calculate the regression coefficients. The regression models were evaluated to assess their suitability in capturing all information in the data. Finally, the selected best-fitted model explains the relationship between Malaysia's FSI and macroeconomic variables.

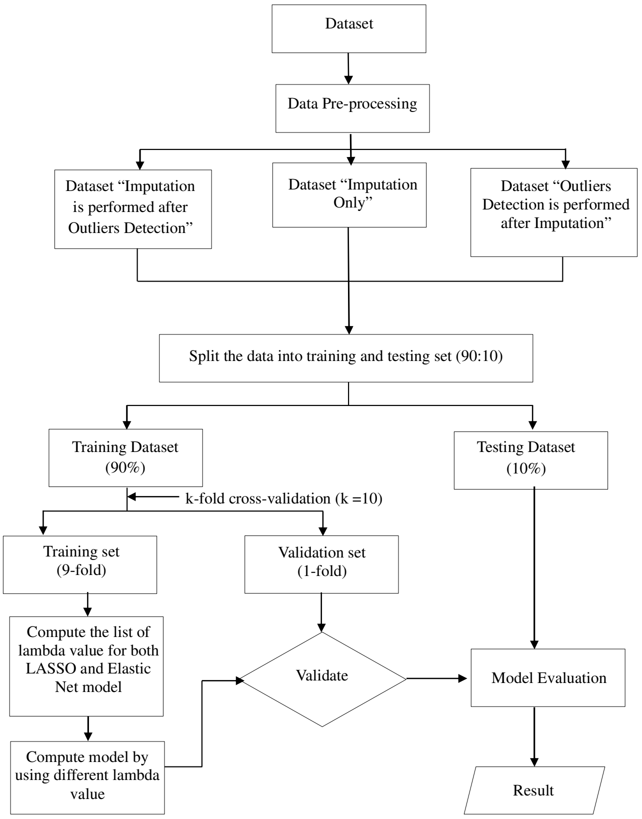

Dataset
Data Pre-processing
Dataset "Imputation is performed after Outliers Detection"
Dataset "Imputation Only"
Dataset "Outliers Detection is performed after Imputation"
Split the data into training and testing set (90:10)
Training Dataset (90%)
Testing Dataset (10%)
k-fold cross-validation (k =10)
Training set (9-fold)
Validation set (1-fold)
Compute the list of lambda value for both LASSO and Elastic Net model
Validate
Model Evaluation
Compute model by using different lambda value
Result

Fig. 1. Analysis flowchart.

## D. *Regularized Regression Model*

### 1) *Least Absolute Shrinkage and Selection Operator (LASSO)*

The LASSO regression model utilizes the L1 regularization technique [21]. The regularization equation of the LASSO regression model is defined as:

$\mathcal{L}(\beta;\lambda) =$

$$\sum_{i=1}^{n}(y_t-\beta_0-\sum_{j=1}^{t}\beta_j x_{ij})^2+\lambda\sum_{j=1}^{p}|\beta_j| \qquad (1)$$

where $L(\beta;\lambda)$ is the loss function of the LASSO regression model, $\beta_j$ is the coefficient of independent variables, and $\lambda$ is a tuning parameter that controls the amount of shrinkage and adjusts the penalty terms.

### 2) *Elastic Net Regression Model*

Elastic Net regression combines both L1 and L2 regularization techniques [22]. The Elastic Net regression model can be defined as:

$$\mathcal{L}(\beta;\lambda)=\sum_{i=1}^{n}(y_t-\beta_0-\sum_{j=1}^{t}\beta_j x_{ij})^2+$$

$$\lambda\left[\frac{1}{2}(1-\alpha)\sum_{j=1}^{p}\beta_j^2+\alpha\sum_{j=1}^{p}|\beta_j|\right] \qquad (2)$$

The glmnet package in R was used to implement both regularized regression models [23]. Using the package's cv.glmnet function, k-fold cross-validation was used to obtain the optimum value of $\lambda$. This approach ensures that the $\lambda$ value minimizes prediction error and optimizes model performance.

The performance of both regularized regression models was evaluated using datasets prepared by different preprocessing sequences. Metrics such as prediction accuracy and information criteria were used to assess and compare their effectiveness in capturing underlying data patterns.

## E. *Model Selection and Evaluation*

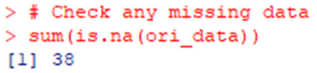
The Bayesian Information Criterion (BIC) and the Akaike Information Criterion (AIC) are used to select the best-fitting regression model. AIC estimates prediction error and evaluates the quality of the model, while BIC favors the simplest model. The model with the lowest AIC and BIC values is considered the most effective at describing variable relationships [24]. These metrics align with the study's goal of balancing model fit and complexity by assessing both goodness-of-fit and parsimony. AIC prioritizes prediction accuracy and better-fit models, even at the cost of slightly increased complexity. In contrast, BIC imposes a stricter penalty for additional parameters, reducing the risk of overfitting and ensuring model generalizability. This approach helps identify preprocessing sequences that enhance model reliability while maintaining simplicity, making it applicable to economic and agricultural research. Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) were used to evaluate prediction accuracy, while the Jarque-Bera test (JB) was used to confirm that the selected regression model follows a normal distribution.

# III. RESULTS AND DISCUSSION

## A. *Data Preprocessing*

### 1) *Missing Data*

Figure 2 shows that 38 data points (0.23%) in the original dataset of 2,856 data points were missing. Therefore, imputation using the MICE approach was required. This produced a dataset called "Dataset done Imputation Only" which was then analyzed further.



```
> # Check any missing data
> sum(is.na(ori_data))
[1] 38
```

Fig. 2. Missing data check.

Another subset of the original data was detected for outliers before imputation. This subset, comprising 1,292 data points, had 12 missing data points (0.93%). Again, MICE was used for imputation to handle the missing data. This dataset is termed "Dataset where Imputation was Performed After Outlier Detection."

### 2) *Outliers Detection*

The dataset with only imputation underwent outlier detection and is now referred to as "Dataset where Outlier Detection is Performed After Imputation." Its 2,125 data points (74%) were identified as outliers and removed from further modeling analysis.

Conversely, in the "Dataset where Imputation is Performed After Outlier Detection," 1,564 data points (constituting 54.76% of the total) were flagged as outliers. After removing these outliers and imputing missing data using the MICE technique, the resulting datasets were utilized for multicollinearity checking. As a result, there were three different datasets:

- Dataset done Imputation Only,
- Dataset Where Outlier Detection is Performed After Imputation,
- Dataset Where Imputation is Performed After Outlier Detection.

Table II shows the training and testing dataset overview for these various preprocessing procedures. The training sets for the dataset where outlier detection was performed after imputation and the dataset where imputation was performed after outlier detection were relatively small, with 35 and 61 observations, respectively. This reduction is attributed to removing outlier observations during the preprocessing stages. Conversely, the dataset with imputation only maintained a larger training set of 135 observations due to no outlier exclusion. These variations in training set sizes highlight the trade-off between handling data irregularities, such as outliers, and preserving sample size for model development.

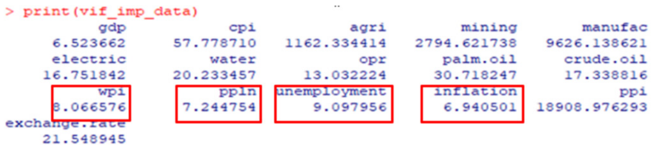
TABLE II. SAMPLE SIZE OF TRAINING AND TESTING SET FOR DIFFERENT PRE-PROCESSING SEQUENCES

| Sample size | Dataset with Imputation Only | Imputation is Performed After Outlier Detection | Outlier Detection is Performed After Imputation |
|---|---|---|---|
| Training set | 135 | 61 | 35 |
| Testing set | 33 | 15 | 8 |

### 3) *Multicollinearity Checking*

Figures 3, 4, and 5 show the results of multicollinearity evaluation of the three different datasets. Figure 3 displays the multicollinearity analysis results for the dataset where only imputation was performed. All variables in this dataset have a VIF value of more than 10, which suggests multicollinearity.

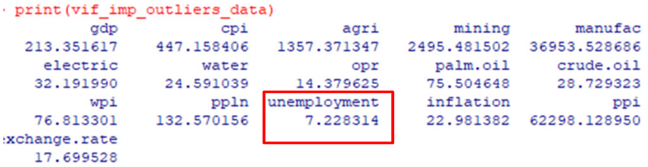
However, it is noteworthy that variables such as *GDP*, *wpi*, *ppln*, *unemployment*, and *inflation* have VIF values less than 10, suggesting no significant multicollinearity among each other.



```
> print(vif_imp_data)
          gdp           cpi          agri        mining       manufac 
     6.523662     57.778710   1162.334414   2794.621738   9626.138621 
     electric         water           opr      palm.oil     crude.oil 
    16.751842     20.233457     13.032224     30.718247     17.338816 
          wpi          ppln  unemployment     inflation           ppi 
     8.066576      7.244754      9.097956      6.940501  18908.976293 
exchange.rate 
    21.548945 
```

Fig. 3. VIF result of the dataset with imputation only.

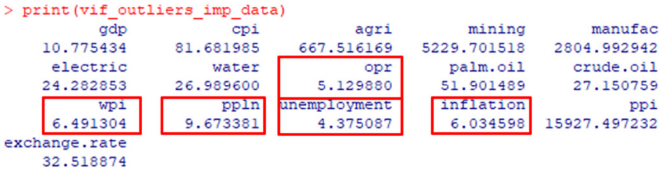
Figure 4 displays the multicollinearity analysis results for the dataset where outlier detection was performed after imputation. In general, the VIF value of this dataset is higher than the dataset with only imputation. The VIF values for all variables in this dataset exceed 10 except the *unemployment* variable, indicating multicollinearity within and among the variables.



```
· print(vif_imp_outliers_data)
          gdp           cpi          agri        mining       manufac 
   213.351617    447.158406   1357.371347   2495.481502  36953.528686 
     electric         water           opr      palm.oil     crude.oil 
    32.191990     24.591039     14.379625     75.504648     28.729323 
          wpi          ppln  unemployment     inflation           ppi 
    76.813301    132.570156      7.228314     22.981382  62298.128950 
:xchange.rate 
    17.699528 
```

Fig. 4. VIF result of dataset where outlier detection is performed after imputation.

Figure 5 shows the multicollinearity assessment result of the dataset where imputation followed outlier detection. It is worth noting that the variables *opr*, *wpi*, *ppln*, *unemployment*, and *inflation* exhibit no significant correlation with each other, as their VIF values do not exceed 10. Furthermore, the VIF value for each variable is relatively smaller than the dataset where outlier detection was performed after imputation. Although these variables have VIF values of more than 10, they are retained in the dataset as L1 and L2 regularization techniques are applied during the modeling step.



```
> print(vif_outliers_imp_data)
          gdp           cpi          agri        mining       manufac 
    10.775434     81.681985    667.516169   5229.701518   2804.992942 
     electric         water           opr      palm.oil     crude.oil 
    24.282853     26.989600      5.129880     51.901489     27.150759 
          wpi          ppln  unemployment     inflation           ppi 
     6.491304      9.673381      4.375087      6.034598  15927.497232 
exchange.rate 
    32.518874 
```

Fig. 5. VIF result of dataset where Imputation is performed after outlier detection.

### *B. Model Evaluation of LASSO and Elastic Net Regression*

The results in Table III reveal that the LASSO regression model computed using the dataset with imputation only (i.e., no outlier detection) produced the highest error rates among the evaluated approaches. Elevated RMSE indicates insufficient predictive power and a failure to preserve information fidelity relative to the actual data, raising concerns about the model's reliability [25]. Similarly, the high MAPE underscores substantial discrepancies between the predicted and observed values, further weakening the model's credibility. High AIC (248.885) and BIC (294.144) values further suggest that the absence of outlier detection undermines model optimization, as undetected outliers distort variable relationships and magnify prediction errors. Consequently, this configuration demonstrates the weakest explanatory power among the three preprocessing approaches.

The LASSO regression model applied to the dataset where outliers were handled before missing data imputation exhibits moderate performance in terms of model evaluation and selection. This is supported by lowered AIC (-69.185) and BIC (-42.550) values, which show an improvement in model fit compared to the model utilizing imputation alone. Reducing evaluation metrics such as MAE, RMSE, and MAPE confirms the significance of addressing outliers in the preprocessing stage. However, although this approach improved model performance relative to the dataset with imputation only, it did not produce the best-fitting LASSO regression model.

The LASSO regression model calculated using the dataset where outlier detection was performed after imputation had the lowest error rates. With the most negative AIC (-154.093) and BIC (-139.355) values, this configuration showed the best model structure with the least information loss. These findings demonstrate that addressing missing data first preserves data integrity, enhances the accuracy of outlier identification, and reduces prediction errors. This result underscores the importance of preprocessing sequence in improving model performance, particularly for complex datasets where data quality significantly affects predictive accuracy.

TABLE III. MODEL EVALUATION RESULTS OF THE LASSO REGRESSION MODEL

| | **Dataset with Imputation Only** | **Imputation is Performed After Outlier Detection** | **Outlier Detection is Performed After Imputation** |
|---|---|---|---|
| MAE | 1.223 | 0.327 | 0.066 |
| RMSE | 1.489 | 0.356 | 0.069 |
| MAPE | 1.776 | 0.490 | 0.096 |
| AIC | 248.885 | -69.185 | -154.093 |
| BIC | 294.144 | -42.550 | -139.355 |

Table IV shows the evaluation results for the Elastic Net regression model across different data preprocessing sequences. Like the findings for the LASSO regression model, the Elastic Net model using the dataset with imputation only yielded high prediction error rates. This configuration recorded elevated values for MAE, RMSE, and MAPE, indicating higher prediction error and model unreliability when no outlier detection is applied. These results suggest that imputation alone, without addressing outliers, fails to sufficiently correct data distortions, which may lead to misleading variable associations and inflated residual errors.

In contrast, the Elastic Net model with outlier detection performed after imputation showed the best performance, while the model with imputation after outlier detection exhibited moderate performance. At the same time, both the LASSO and Elastic Net models performed similarly when comparing the datasets where imputation was performed after outlier detection and outlier detection was performed after imputation. Hence, it can be concluded that the regularized regression model computed using the dataset where outlier detection was performed after imputation obtained outstanding capability in capturing the information present in the dataset.

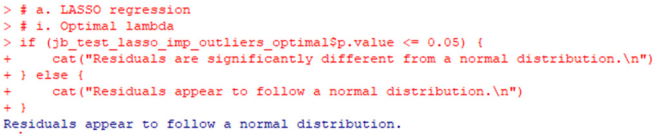
However, comparing the explanatory among the LASSO and Elastic Net regression models, LASSO regression outperformed Elastic Net regarding model evaluation and selection. LASSO achieved an AIC of -154.093 and BIC of -139.355, compared to Elastic Net's AIC of -148.034 and BIC of -128.383. LASSO's lower AIC and BIC values indicate a more efficient and better-fitting model, with reduced information loss and a simpler structure. This is likely due to LASSO's robust variable selection ability, which reduces model complexity while maintaining predictive power. In contrast, Elastic Net, which combines both LASSO and ridge penalties, tends to retain more variables, leading to slightly higher AIC and BIC values. Therefore, the LASSO regression model is recommended when interpretability and simplicity are key considerations.

TABLE IV. MODEL EVALUATION RESULT OF ELASTIC NET REGRESSION MODEL

| Dataset Type / Model Evaluation | Dataset with Imputation Only | Imputation is Performed After Outlier Detection | Outlier Detection is Performed After Imputation |
|---|---|---|---|
| MAE | 1.220 | 0.341 | 0.066 |
| RMSE | 1.484 | 0.369 | 0.069 |
| MAPE | 1.772 | 0.511 | 0.097 |
| AIC | 249.859 | -64.107 | -148.034 |
| BIC | 298.135 | -37.473 | -128.383 |

### C. *Goodness of Fit Test*

Figure 6 shows the best model's Jarque-Bera test result. The test's *p*-value was contrasted with a significance level of 0.05. Since the *p*-value is higher than this significance level, it can be concluded that the residuals follow a normal distribution with no significant deviation from normality. In conclusion, the Jarque-Bera test result validates the LASSO regression model's residuals' assumption of normality, showing that the model is suitably described and capable of producing accurate predictions.



```
> # a. LASSO regression
> # i. Optimal lambda
> if (jb_test_lasso_imp_outliers_optimal$p.value <= 0.05) {
+     cat("Residuals are significantly different from a normal distribution.\n")
+ } else {
+     cat("Residuals appear to follow a normal distribution.\n")
+ }
Residuals appear to follow a normal distribution.
```

Fig. 6. Jarque-Bera test result of the best model.

### D. *Model Explanation*

The LASSO regression model with the preprocessing sequence where outlier detection is applied after imputation was chosen:

$$FSI = 67.851 - 0.0104X_1 - 0.005X_2 + 0.004X_3 - 0.007X_4 - 0.151X_8 + 1.465X_{12} + 0.022X_{14} + 0.148X_{16} \quad (3)$$

OPR demonstrates the most substantial negative relationship, with a coefficient of -0.151. This indicates that for each unit increase in OPR, the FSI is expected to decrease by 0.151 units, assuming all other variables remain constant. In contrast, the *ppln* variable positively impacts FSI, contributing 1.465 units to FSI for each unit increase in the Malaysia population growth, with other variables held constant.

## IV. FINDINGS

The study's primary finding is that the sequence of preprocessing steps significantly affects the performance of the LASSO and Elastic Net regression models. Specifically, outlier detection after imputation resulted in the most reliable models with lower prediction error rates and improved variable selection. This result is consistent with earlier research indicating that early imputation maintains data integrity and improves later analysis [9]. However, unlike previous research that focused on individual preprocessing methods, this study provides a systematic comparison of preprocessing sequences, filling an important gap in the literature. The results have significant implications for food security analysis in Malaysia. By identifying the optimal preprocessing sequence, the study enhances the reliability of the models used to examine the relationships between macroeconomic variables and food security indices. The findings highlight that population growth positively affects Malaysia's FSI, while OPR negatively affects it. These observations can help policymakers develop plans to strike a compromise between the goals of food security and economic growth.

## V. CONCLUSION

This study emphasizes the importance of preprocessing sequences in the preliminary stages of analysis, especially when it comes to improving the performance of predictive models in economic and agricultural settings. While previous research often overlooks how preprocessing steps interact with model accuracy, this work highlights, for the first time, the significant impact that the order of these steps can have. The study explores complex economic interactions using the LASSO and Elastic Net regularized regression approaches by creating a novel framework that maximizes predictive performance. Furthermore, it offers valuable insights for shaping evidence-based food security policies in Malaysia, pushing the boundaries of existing methods, and providing unique contributions to the field.

## ACKNOWLEDGEMENT

The author would like to thank the funding from the research grant number GPPS-Q540 provided by the Universiti Tun Hussein Onn Malaysia (UTHM). The authors would also like to thank the Faculty of Applied Sciences and Technology (FAST) Universiti Tun Hussein Onn Malaysia for its support.

## REFERENCES

[1] "Global Strategic Framework for Food Security & Nutrition (GSF)." https://www.fao.org/cfs/policy-products/onlinegsf/en/.

[2] Martin, "Goal 2: Zero Hunger," *United Nations Sustainable Development*. https://www.un.org/sustainabledevelopment/hunger/.

[3] A. K. Tiwari, S. Nasreen, M. Shahbaz, and S. Hammoudeh, "Time-frequency causality and connectedness between international prices of

energy, food, industry, agriculture and metals," *Energy Economics*, vol. 85, Jan. 2020, Art. no. 104529, https://doi.org/10.1016/j.eneco.2019.104529.

[4] R. Kollmann, "Effects of Covid-19 on Euro area GDP and inflation: demand vs. supply disturbances," *International Economics and Economic Policy*, vol. 18, no. 3, pp. 475–492, Jul. 2021, https://doi.org/10.1007/s10368-021-00516-3.

[5] A. O. El Alaoui, H. B. Jusoh, S. A. Yussof, and M. H. Hanifa, "Evaluation of monetary policy: Evidence of the role of money from Malaysia," *The Quarterly Review of Economics and Finance*, vol. 74, pp. 119–128, Nov. 2019, https://doi.org/10.1016/j.qref.2019.04.005.

[6] B. Kuma and G. Gata, "Factors affecting food price inflation in Ethiopia: An autoregressive distributed lag approach," *Journal of Agriculture and Food Research*, vol. 12, Jun. 2023, Art. no. 100548, https://doi.org/10.1016/j.jafr.2023.100548.

[7] S. Batra and S. Sachdeva, "Organizing standardized electronic healthcare records data for mining," *Health Policy and Technology*, vol. 5, no. 3, pp. 226–242, Sep. 2016, https://doi.org/10.1016/j.hlpt.2016.03.006.

[8] J. V. den Broeck, S. A. Cunningham, R. Eeckels, and K. Herbst, "Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities," *PLOS Medicine*, vol. 2, no. 10, 2005, Art. no. e267, https://doi.org/10.1371/journal.pmed.0020267.

[9] P. Misra and A. S. Yadav, "Impact of Preprocessing Methods on Healthcare Predictions," in *Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE)*, 2019, https://doi.org/10.2139/ssrn.3349586.

[10] E. de Jonge and M. van der Loo, *An Introduction to Data Cleaning with R*. Statistics Netherlands, 2013.

[11] C. Quintano, R. Castellano, and A. Rocca, "Influence of outliers on some multiple imputation methods," *Advances in Methodology and Statistics*, vol. 7, no. 1, Jan. 2010, https://doi.org/10.51936/tuki4538.

[12] J. Jeong and C. Kim, "Effect of outliers on the variable selection by the regularized regression," *Communications for Statistical Applications and Methods*, vol. 25, no. 2, pp. 235–243, Mar. 2018, https://doi.org/10.29220/CSAM.2018.25.2.235.

[13] F. E. Grubbs, "Procedures for Detecting Outlying Observations in Samples," *Technometrics*, vol. 11, no. 1, pp. 1–21, Feb. 1969, https://doi.org/10.1080/00401706.1969.10490657.

[14] E. M. Raouhi, M. Lachgar, and A. Kartit, "Comparative Study of Regression and Regularization Methods: Application to Weather and Climate Data," in *WITS 2020*, vol. 745, S. Bennani, Y. Lakhrissi, G. Khaissidi, A. Mansouri, and Y. Khamlichi, Eds. Springer Singapore, 2022, pp. 233–240.

[15] Department of Statistics Malaysia, "Data Catalogue | OpenDOSM." https://open.dosm.gov.my.

[16] S. Van Buuren, "Multiple imputation of discrete and continuous data by fully conditional specification," *Statistical Methods in Medical Research*, vol. 16, no. 3, pp. 219–242, Jun. 2007, https://doi.org/10.1177/0962280206074463.

[17] S. V. Buuren and K. Groothuis-Oudshoorn, "mice : Multivariate Imputation by Chained Equations in *R*," *Journal of Statistical Software*, vol. 45, no. 3, 2011, https://doi.org/10.18637/jss.v045.i03.

[18] J. Laurikkala, M. Juhola, and E. Kentala, "Informal identification of outliers in medical data," in *Fifth International Workshop on Intelligent Data Analysis in Medicine and Pharmacology*, vol. 1, no. 1, pp. 20–24.

[19] C. Chen and L. M. Liu, "Joint Estimation of Model Parameters and Outlier Effects in Time Series," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 284–297, Mar. 1993, https://doi.org/10.1080/01621459.1993.10594321.

[20] N. Shrestha, "Detecting Multicollinearity in Regression Analysis," *American Journal of Applied Mathematics and Statistics*, vol. 8, no. 2, pp. 39–42, Jun. 2020, https://doi.org/10.12691/ajams-8-2-1.

[21] R. Tibshirani, "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, Jan. 1996, https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[22] H. Zou and T. Hastie, "Regularization and Variable Selection Via the Elastic Net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, Apr. 2005, https://doi.org/10.1111/j.1467-9868.2005.00503.x.

[23] J. Friedman *et al.*, "glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models." Jun. 02, 2008, https://doi.org/10.32614/CRAN.package.glmnet.

[24] E. A. Mohammed, C. Naugler, and B. H. Far, "Emerging Business Intelligence Framework for a Clinical Laboratory Through Big Data Analytics," in *Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology*, Elsevier, 2015, pp. 577–602.

[25] M. A. A. Abdullah, L. Jesintha, G. P. Khuneswari, S. A. M. Jamil, and O. R. Olaniran, "Comparison of Multiple Regression and Model Averaging Model-Building Approach for Missing Data with Multiple Imputation," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 18502–18508, Dec. 2024, https://doi.org/10.48084/etasr.8909.