MJFAS
Malaysian Journal of Fundamental and Applied Sciences

Abu Afouna and Majahar Ali | Malaysian Journal of Fundamental and Applied Sciences, Vol. 20 (2024) 1344-1362

RESEARCH ARTICLE

The Impact of Heterogeneity in High-Ranking Variables Using Precision Farming

Nour Abu Afouna, Majid Majahar Ali*

School of Mathematical Sciences, Universiti Sains Malaysia, 11800 USM, Penang, Malaysia

Abstract Smart precision farming combines IoT, cloud computing, and big data to optimize agricultural productivity, reduce costs, and advance sustainability through digitalization and intelligent approaches. However, precision farming grapples with challenges like managing complex variables, addressing multicollinearity, handling outliers, ensuring model robustness, and improving accuracy, particularly with smaller or medium-sized datasets. Reducing retraining time and solving the calamity of complexity are necessary to overcome these obstacles and improve machine learning algorithms' performance, scalability, and efficiency—especially when working with big or high-dimensional datasets. In a recent study with 435 drying parameters and 1914 observations. In this study, we employed Ridge, Lasso, and Elastic Net regression techniques to address the challenges of multicollinearity and heterogeneity within our dataset. Traditional regression models, such as ordinary least squares (OLS), often struggle with multicollinearity, leading to unstable and unreliable coefficient estimates. Ridge regression mitigates this issue by adding an L_2 penalty, stabilizing the coefficients. Lasso regression introduces an L_1 penalty, which further enhances the model by performing variable selection. Elastic Net, a combination of L_1 and L_2 penalties, effectively handles both multicollinearity and heterogeneity by selecting relevant variables and capturing varying patterns across different subgroups. Our study's use of Ridge, Lasso, and Elastic Net regression techniques has broad practical applications across various fields. In economics, they help identify key indicators for economic forecasting; in healthcare, they improve predictions of patient outcomes for personalized treatment; in finance, they create more stable models for market behavior; and in social sciences, they reveal influential factors in behavioral studies. These methods effectively manage multicollinearity and heterogeneity, making them valuable tools for decision-making and policy development across these domains. The objective was to identify significant drying parameters both before and after heterogeneity, while selecting varying numbers of variables (50, 100, 150, 200, 250, 300) based on validation metrics such as MAPE, MSE, SSE, and R^2. The results revealed that the Ridge model demonstrated the highest efficiency, exhibiting the smallest values for MAPE, MSE, SSE, and the largest value for R^2, both before and after heterogeneity.

Keywords: Machine learning, lasso, ridge, elastic net, validation metrics, smart farming.

***For correspondence:**
majidkhanmajaharali@usm.my

Received: 10 May 2024
Accepted: 20 Oct. 2024

Introduction

Smart farming, also referred to as smart agriculture, entails the utilization of cutting-edge technologies and data-driven strategies to enhance efficiency and sustainability in agricultural practices. This approach integrates artificial intelligence (AI), automation, and the Internet of Things (IoT) to optimize various aspects of farming operations [1]. In Smart Farming technology, IoT involves linking smart machines and sensors throughout agricultural operations, enabling data-driven and data-enabled farming processes [2]. Solar dryers in precision farming offer energy efficiency, environmental benefits, precise control over drying conditions, and economic advantages. These benefits collectively contribute to the sustainability and profitability of agricultural practices, making solar dryers a valuable tool in precision farming [3, 4, 5, 6, 7].

In agriculture and aquaculture, producing high-value end products requires several phases, including cultivation, pre and post-harvest. Drying is synonym for post-harvest. Drying and dehydration play pivotal roles in agriculture. These processes involve the removal of moisture from food materials, enhancing their longevity, preventing spoilage, and elevating their overall quality [8]. Drying is a widely recognized technique for food preservation, achieved by lowering the moisture content to an ideal level. Different techniques are utilized to dry food, such as direct exposure to sunlight, oven drying, warmth pump drying, and solar energy-based drying processes [9].

The conventional method open-air drying in the sun remains a popular alternative because of its low cost. However, this approach is very sensitive to ambient circumstances and is at risk of pollution by dust, storms, moisture, diseases, and rodents in particular which leads to lower product quality. In contrast, the hybrid solar drier provides a quicker, better performing, and cleaner alternative, causing lower losses to crop than typical in the open drying in the sunlight. In the sunlight's air-drying process, moisture is taken from raw agricultural materials using all three of the modes of heat transmission [10].

A solar dryer is an efficient method for preserving agricultural products by utilizing solar radiation for drying. Various studies have focused on designing and evaluating solar dryers for different agricultural commodities. For instance, studies have developed solar dryers with components like solar collectors, drying chambers, and fans to enhance drying efficiency [11, 12, 13]. solar dryers offer a sustainable and cost-effective solution for post-harvest preservation of agricultural products.

A solar drier has been studied in various kinds of fields, including seaweed. Ibidoja study has demonstrated that solar drying technologies are useful in retaining high-quality seaweed biomass for food and feed markets [14, 15, 16, 17]. Solar-assisted warmth pump drying systems have been created to minimize time to drying, consumption of energy, and drier performance, making them a sustainable and energy-saving alternative to traditional drying techniques [18].

The existing problems in dryer/IoT systems issues related to traditional drying, such as limited resources leading to the leading to not efficient drying systems. Additionally, challenges like unpredictable weather affecting the drying process have been highlighted, prompting the need for cost-effective solutions. Furthermore, the impact of high temperatures on the efficiency of solar panels used in solar dryers, leading to utilize excess heat for drying purposes while maintaining panel efficiency. Moreover, the importance of monitoring and maintaining specific parameters like temperature, humidity, solar radiation, and air velocity in dryers to ensure food quality and safety has been emphasized, necessitating real-time alerts and control systems in IoT-based dryers [19].

The Internet of Things (IoT) entails managing numerous parameters and their interplay, resulting in substantial complexity in handling big data when recording in a cloud database. In the dryer, 29 sensors were strategically placed to gather data on drying parameters. Due to the sheer quantity of sensors, it becomes challenging to pinpoint significant variables and construct an effective predictive model [20]. Drying processes can benefit significantly from IoT technology, as seen in various studies. IoT-based systems allow for real-time monitoring and control of drying parameters like temperature, relative humidity, and air velocity, ensuring food quality and safety [21, 22]. Overall, integrating IoT into drying processes enhances efficiency, quality control, and automation capabilities.

Dealing with numerous sensor variables is a significant challenge in IoT systems for agriculture [23, 24, 25]. These systems aim to enhance farm productivity and quality by monitoring various parameters like soil pH, humidity, temperature, water salinity, and environmental conditions [26]. Implementing precision farming using IoT and IIoT infrastructure involves fusion of multiple sensors to collect critical data for decision-making, including monitoring weather variability, automating irrigation, and extracting soil properties [27]. Smart farming methods based on IoT offer high precision crop control and automated farming techniques, monitoring soil humidity, temperature, and automating irrigation system without constant manual intervention. The data obtained from sensors aids in improving cultivation precision, optimizing watering, pesticide application, and adjusting environmental variables for optimal plant growth. In land or water as mentioned in SDG 13. So, we need to focus more on how to dealt with complex variables.

Heterogeneity refers to the quality or state of being diverse or varied. In the context of data, it describes a dataset that contains differing elements or characteristics. This could involve variations in data types (numerical, categorical, textual), sources (different sensors, surveys, experiments), scales (different units of measurement), or distributions (skewness, kurtosis). In complex systems like socio-ecological systems, understanding the diverse interactions and feedbacks across time and space is key. Analyzing this heterogeneity is essential for sustainability analysis and governance, as it aids in effectively modeling non-linearities and system dynamics [28]. The data in this study is continuous data and does not contain any missing values.

Literature Review

Anam Javaid *et al*. (2019a) [29] conducted a study to analyze the impacts of interactions among key factors within a drier integrated with an IoT system. The research focused on five variables, including one dependent variable—moisture ratio (Y)—and four independent variables: chamber temperature (X1), chamber humidity (X2), solar radiation (X3), and collector efficiency (X4). Using multiple regression analysis, this research studied interactions until to the third degree, yielding a total of 32 possible models. This comprehensive investigation provided valuable insights into the drying process facilitated by solar driers, emphasizing the significance of considering intricate interactions among variables. The same author in (2019b) [30] examined the influencing factors on collector efficiency in solar driers. Their study focused on five variables, including one response variable—collector efficiency (Y)—and four predictor variables: time (X1), inlet temperature (X2), collector average temperature (X3), and solar radiation (X4). They explored interactions until to the third level across 32 models, comparing the outcomes of two regression analyses. In ordinary least squares (OLS) regression, the final model had three individual variables and three association variables, whereas with LASSO regression, it included of three individual variables and five association factors that contributed to predicting solar drier collector performance. However, identified the outlier.

Anam Javaid *et al*. (2020) [31] undertook a study aimed at selecting the most effective model for predicting collector efficiency in solar dryers. They employed a hybrid approach combining LASSO (Least Absolute Shrinkage and Selection Operator) and robust regression to address outliers, as robust methods are essential for outlier detection and removal. Additionally, Ridge regression was utilized to mitigate multicollinearity, although it can be influenced by outliers. The study focused on five variables, including collector efficiency (Y) as the response variable, and time (X1), inlet temperature (X2), collector average temperature (X3), and solar radiation (X4) as predictor variables. After investigating interactions until to the third level among 32 various models, they determined that using LASSO with the Huber M estimator generated the most efficiency model compared to other approaches. Consequently, they identified significant variables such as time, collector average temperature, solar radiation, the association between time and solar radiation, and the association between time, collector average temperature, and solar radiation as key determinants for forecasting collector efficiency in solar dryers. This model is well-prepared for predicting the collector efficiency of solar dryers and decreasing outlier.

Next, Hui Yin Lim *et al*. (2020) [32] examined the accurate model selection and prediction of fish drying using Ridge regression against OLS. The study focused on six variables, including one response variable—fish moisture content (Y)—and five predictor variables: inlet temperature chamber (X1), outlet temperature chamber (X2), outlet humidity chamber (X3), inlet humidity chamber (X4), and solar radiation (X5). Their findings revealed that Ridge regression was the best way for creating the most effective prediction model for fish drying with the V-Groove Hybrid Solar Drier.

Again, Anam Javaid *et al*. (2021) [33] focused on the efficient selection of models for removing moisture ratio from seaweed, utilizing a hybrid approach combining sparse and robust regression analysis. The dataset utilized was sourced from a solar drier, with moisture ratio removal (%) serving as the response variable. Predictor variables included ambient temperature, chamber temperature, collector temperature, chamber relative humidity, ambient relative humidity, and solar radiation. The analysis considered a total of 192 models, aiming to identify the significant factors affecting solar drier efficiency and moisture ratio removal, including their interaction effects. The hybrid of LASSO and robust regression emerged as an effective method for pinpointing these significant factors and interaction terms within the dataset.

In their paper, Mukhtar *et al*. (2022a) [34] analyze the effect of three distinct machine learning algorithms on variable selection: random forest, support vector machine, and boosting. Furthermore, outliers were dealt with using M robust regression approaches. Use a set of 1924 data to examine the effects of 29 different predictor variables on one response variable. The data follows a secondary interaction procedure. This dataset includes the effects in 435 different interaction predictor factors on the response variable. They showed that the random forest-Hampel's model is the best one to use for food safety and sustainable farming.

Mukhtar *et al*. (2022b) [35] three variable selection strategies were used with regularization methods of regression to examine the effect on seaweed drying data. They focused simply on the data's second-order interaction, which included 435 different predictor interaction factors. Following this analysis, they compared the effectiveness of these techniques. Subsequently, they employed robust regression methods, including Tukey Bi-Square, Hampel, and Huber, to further evaluate their findings. Their study concluded that the Lasso-Hampel method demonstrated reliability in accurately assessing large datasets derived from both regularization and robust regression approaches.

In 2023 study, Ibidoja *et al*. [36] using a dataset comprising 435 parameters with 1914 observations each. Initially, they used four ML algorithms—random forest, support vector machine, bagging, and boosting—to identify relevant parameters, choosing 15, 25, 35, and 45 parameters, respectively. Subsequently, they crafted a hybrid model integrating robust methodologies like M. Bi-Square, M. Hampel, and M. Huber. Their findings showed that using the hybrid approach to handle impacted seaweed large data resulted in a significant reduction in outliers and improved predictive skills. Additionally, as the most important variable, which included 45 crucial seaweed drying factors, the bagging M. Bi-Square hybrid model outperformed the others.

In their 2023(b) study, Ibidoja *et al*. [20] initially worked with 29 drying factors, each boasting 1914 data. By considering interactions until to the second order, they expanded the variables to 435 from the original 29. This expansion posed a challenge due to having fewer observations than variables in high-dimensional data. To solve this, the researchers proposed an approach that uses the variance inflation factor to detect heterogeneous parameters. They used seven forecasting models—ridge, random forest, support vector machine, bagging, boosting, LASSO, and elastic net—to identify 15, 25, 35, and 45 relevant drying factors for removing seaweed moisture content. Subsequently, they crafted hybrid models incorporating robust statistical techniques. Their findings indicated that for pre-heterogeneity analysis, the hybrid model, which combined random forest and M. Hampel, outperformed the other models. In post-heterogeneity analysis, the hybrid model mixing boosting and M. Hampel outperformed their alternatives. As shown in Table 1.

Table 1. Summary for literature review

#	Authors	Variables	Problems	Solving
1.	Anam Javaid *et al*., (2019a)	1 response variable 4 predictor (single) variables 28 interaction Total 32 with interaction -temperature chamber (X_1), humidity chamber (X_2) , solar radiation (X_3) collector efficiency (X_4)	Analysed the impact of interactions between primary components. multicollinearity outliers check assumption normality independent, randomness, homogeneous	Using multiple regression models
2.	Anam Javaid *et al*., (2019b)	1 response variable 4 predictor (single) variables 28 interaction Total 32 with Interaction - time (X_1) ,inlet temperature (X_2) ,collector average temperature(X_3), solar radiation (X_4)	Examined the primary components and their interactions impacting collector efficiency. multicollinearity	Using 8SC for LASSO and multiple regression models.
3.	Anam Javaid *et al*., (2020)	1 response variable 4 predictor (single) variables 28 interaction Total 32 with Interaction - time (X_1) - inlet temperature (X_2) -collector average temperature (X_3) - solar radiation (X_4)	Comparison of ordinary least squares (OLS) after multicollinearity and coefficient tests vs ridge regression analysis.	Using a mix of LASSO and robust regression (Huber M estimator), and comparing to OLS and ridge regression analysis
4.	Hui Yin Lim *et al*., (2020)	1 response variable moisture content of fish: Y) 5 predictor (single) variables 75 interaction Total 80 models with Interaction -inlet temperature chamber (X_1) -outlet temperature chamber (X_2) -outlet humidity chamber (X_3) -inlet humidity chamber (X_4) - solar radiation (X_5)	Determine the most appropriate model for predicting the moisture content of dried fish. Multicollinearity coefficient test outliers check assumption normality, randomness, independent, homogeneous	- Using ordinary least squares (OLS) regression and ridge regression with 8SC for model construction

#	Authors	Variables	Problems	Solving
5.	Anam Javaid *et al*., (2021)	1 response variable 6 predictor (single) variables 186 interaction Total 192 with Interaction - ambient temperature(X_1) - temperature chamber(X_2) - collector temperature(X_3) - chamber relative humidity (X_4) - ambient relative humidity(X_5) - solar radiation(X_6)	examined main factors with their interaction terms on the moisture ratio removal by considering a large data multicollinearity coefficient test outliers	-Using LASSO with robust regression (Huber M, Hampel M, Bi square M) - Following the multicollinearity and coefficient tests, the approaches are compared using ridge regression and OLS (ordinary least squares).
6.	Mukhtar *et al*.,(2022a)	1 response variable (moisture: *Y*)) 29 predictor variables Total 435 models with Interaction	Examining the effectiveness of three variable selection strategies with machine learning techniques Examine the impact of 435 independent factors on a single dependent variable, selecting 30 important variables. determine the irrelevant variables for big data. determine the outliers for big data	- Hybrid models incorporate machine learning methodologies like random forest (RF), support vector machines (SVM) and boosting techniques. -M-bi square, M-Hampel, and M-Huber are used for robust regression.
7.	Mukhtar *et al*.,(2022b)	1 response variable (moisture: *Y*) 29 predictor variables Total 435 models with Interaction	Determine the impact of three variable selection techniques and regularization regression algorithms on seaweed drying performance. Selecting importance variable (30 variables) Outliers)	-Using ML such as (LASSO, Elastic net, Ridge) -Using robust regression with Tukey- Bi Square, Hampel, and Huber
8.	Ibidojaa et. al., (2023a)	1 response variable (moisture: *Y*) 29 predictor variables Total 435 models with Interaction	Identify significant parameters by choosing 15, 25, 35, and 45 Outliers for big data	- Using four machine learning algorithms: random forest, support vector machine, bagging, and boosting - hybrid model was created utilizing strong approaches like M. Bi-Square, M. Hampel, and M. Huber.
9.	Ibidojaa et. al., (2023b)	1 response variable (moisture content: *Y*) 29 predictor variables Total 435 models with Interaction	identify the heterogeneity parameters Determine the relevant drying parameters (15, 25, 35, and 45) for before and after heterogeneity. outliers for big data before and after heterogeneity	-Using the variance inflation factor - Using seven models for prediction, including ridge, random forest, support vector machine, bagging, boosting, LASSO, and elastic net - The hybrid model was constructed using robust methodologies like M. Bi-Square, M. Hampel, and M. Huber (MM).

Most prior studies, as outlined in Table 1, focus on the implementation of a few variables using statistical and machine learning methods. Current literature underscores that multicollinearity and outliers continue to be significant challenges in big data analysis. Additionally, there is a scarcity of studies investigating the effects of interaction variables, particularly within ultra-dimensional contexts. Addressing these issues in greater detail is essential for developing accurate prediction models for the drying process. Moreover, there is a noticeable research gap concerning the impact of significant variables in high-ranking datasets, such as those involving 435 factors, especially in the context of seaweed. Furthermore, there is a limited body of research dedicated to identifying the most effective prediction models using existing frameworks for high-ranking variables. In this work, we attempted to examine the effect of 435 predictor factors on a one response variable using a dataset of 1924 observations. There are various challenges with big data, including irrelevant factors. the research specifically delved into the analysis

and comparison of outcomes using three distinct variable selection methods grounded in machine learning techniques. Three machine learning algorithms—Ridge, LASSO and Elastic Net regression methods are utilized due to their ability to address issues like multicollinearity, overfitting, and model complexity [37, 38]. Ridge regression reduces overfitting by applying a penalty equal to the square of the size of the coefficients, allowing all predictors to remain in the model [39]. Lasso regression, on the other hand, introduces sparsity by shrinking some coefficients to zero, aiding in feature selection and model simplification. Elastic Net combines the sparsity of Lasso with the grouping effect of Ridge, offering a balance between variable selection and model interpretability [40]. These methods collectively enhance prediction accuracy, handle multicollinearity issues, and provide more interpretable models, making them valuable tools in various research domains. were applied to identify significant parameters considering selections of 50, 100, 150, 200, 250, and 300 variables, both before and after accounting for heterogeneity and comparing the effect of three different machine learning techniques for predicting the efficient model: mean absolute percentage error (MAPE), mean square error (MSE), sum squares of error (SSE), and R-squared (R^2).

Data Description

Data were collected during the drying process of seaweed using a v-Groove Hybrid Solar Drier (v-GHSD). Key factors examined included temperature, ambient relative humidity, chamber relative humidity, and solar radiation. Table 2 presents the 29 primary parameters, each with 1,914 data points. Due to time and complexity constraints, the system needs to be simplified by treating each observation area as a single parameter. The addition of second-order interactions among the initial 29 parameters increased the total number to 435. The optimization process involved identifying the top 50, 100, 150, 200, 250, and 300 most significant factors. Building on the work of Ibidoja *et al*. (2023a), who selected up to 45 variables, this study will extend the analysis to 50 variables. Additionally, the number of variables will be doubled to evaluate the impact of adding another 50 variables to the models, thus selecting 100, 150, 200, 250, and 300 variables.

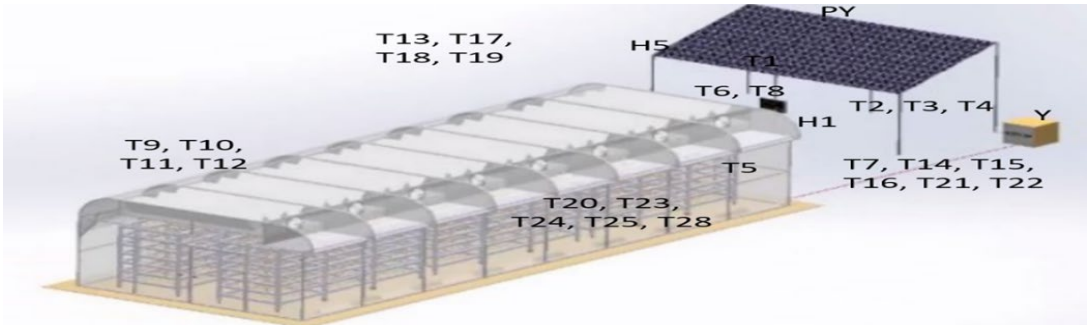

Figure 1. v-GHSD simulation diagram [41]

Figure 1 provides additional details on the drying parameters. Some parameters are missing and are not in sequential order due to measurement errors in the sensors used for data collection. The dataset in this study comprises 1914 data points, featuring 29 predictor variables and one response variable. Interaction effects among various factors are explored, such as T2*T5 representing the interaction between T2 and T5. The dataset encompasses main effects of 29 variables along with interaction effects of 406 variables, all contributing to the determination of the moisture content represented by the dependent variable *Y*. In total, there are 435 predictor variable models that influence the moisture content *Y*.

Materials and Methods

Three machine learning algorithms—Ridge, LASSO, and Elastic Net— were used to identify important features. The study involved selecting subsets of 50, 100, 150, 200, 250, and 300 variables, both before and after adjusting for heterogeneity. The primary goal was to compare the effectiveness of these algorithms in forecasting the most efficient model. The procedure and methodology followed in this investigation are outlined in the flowchart depicted in Figure 2.

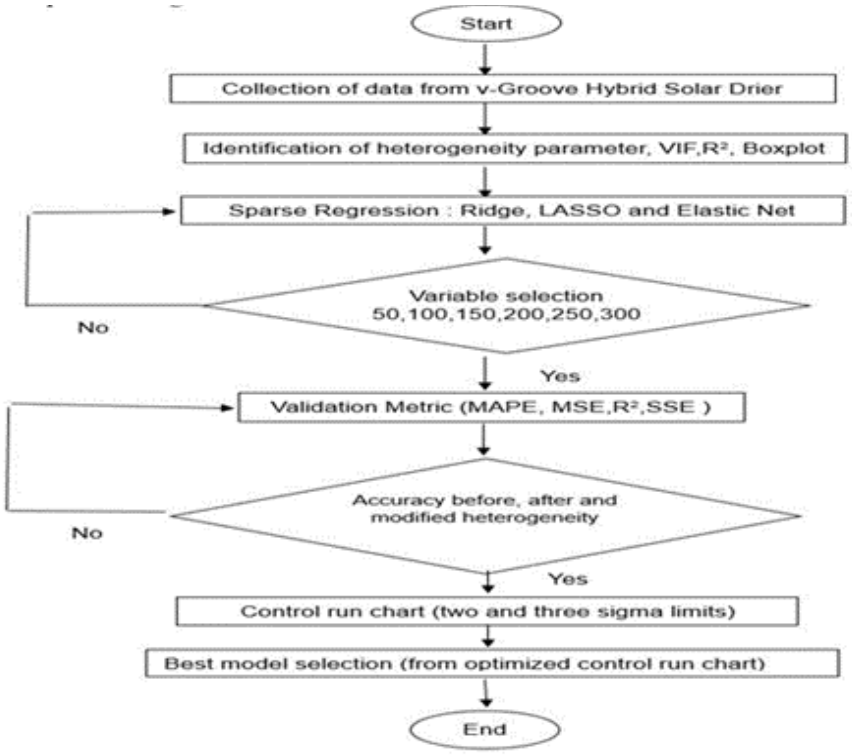

Figure 2. Methodology of the flowchart

Figure 2 presents the flowchart outlining the study's progression. The dataset, consisting of 1914 observations and 29 features, underwent first-order interactions, resulting in 435 distinct features for the dependent variable. The methodology begins by collecting data from the v-Groove Hybrid Solar Drier, incorporating various critical parameters for analysis. The next step involves assessing the heterogeneity of the data using several techniques: The Variance Inflation Factor (VIF) to identify multicollinearity, the coefficient of determination (R^2) to assess model fit, and boxplots to visualize data distribution and detect potential outliers. To address multicollinearity and perform variable selection, sparse regression techniques such as Ridge, LASSO, and Elastic Net are employed. Ridge regression is particularly useful when multicollinearity is present, as it adds a penalty equal to the square of the coefficients' magnitude to the loss function. LASSO (Least Absolute Shrinkage and Selection Operator) introduces a penalty based on the absolute value of the coefficients' magnitude, which helps in variable selection. Elastic Net combines the penalties from both Ridge and LASSO to manage multicollinearity and variable selection effectively with dealing high ranking variable. Following this, the process involves selecting variables based on predefined thresholds (50, 100, 150, 200, 250, 300), ensuring that only the most significant variables are included in the model. The model's performance is then evaluated using validation metrics such as Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), R^2, and Sum of Squared Errors (SSE). These metrics are crucial for assessing the model's accuracy before and after adjustments for heterogeneity. Once refined, the model's performance is monitored over time using a control run chart with two and three sigma limits, ensuring stability and consistency within acceptable control boundaries. Finally, the best-performing model is selected based on its performance in the optimized control run chart, resulting in a model that is both accurate and reliable for predicting outcomes in the v-Groove Hybrid Solar Drier system. The study's impact could be substantial, offering significant improvements in the reliability, efficiency and accuracy of precision farming in high-ranking variables. By tackling critical issues like environmental dependency, analytical complexity, and errors in prediction models, the research has the potential to deliver valuable insights and practical solutions. These advancements could drive the evolution of precision farming, leading to more sustainable and productive agricultural practices that benefit both farmers and the wider agricultural sector. In the context of Ridge, Lasso and elastic net regression, the focus is on predicting continuous values rather than classifying instances into discrete categories. Thus, classification metrics like ROC, AUC, and F1 scores are not suitable for evaluating regression models. Instead, metrics tailored for regression are used to assess the accuracy and performance of the model in predicting continuous outcomes.

Multiple Linear Regression

Consider a multiple regression model:

$$y = X\beta + \varepsilon \tag{1}$$

where y is a $\mathrm{n} \times 1$ vector of response variables, X is known as the design matrix of order $\mathrm{n} \times \mathrm{p}$, β is a $p \times 1$ vector of unknown parameters and ε is a $\mathrm{n} \times 1$ vector of identically and independent distributed errors. The Ordinary Least Squares (OLS) is popularly used to estimate the unknown parameters in a regression model. According to [42, 43] the ordinary least squares (OLS) estimator of β is obtained as follows:

By minimizing $\varepsilon\varepsilon' = (y - X\beta)'(y - X\beta) = y'y - 2\beta'X'y + \beta'X'X\beta$

$$\frac{\delta(\varepsilon\varepsilon')}{\delta\beta} = -2X'y + 2X'X\beta = 0$$

$$X'X\beta = X'y$$

$$\hat{\beta} = (X'X)^{-1}X'y \tag{2}$$

Heterogeneity Identification and Variance Inflation Factor (VIF)

Heterogeneity refers to the variation of observations. The variability lead to incompatible forecasts and affects results [44]. Consider multiple linear regression (MLR):

$$Y_i = \beta_0 + \beta_1 T_{i,1} + \beta_2 T_{i,2} + \cdots + a_j + \varepsilon_i \tag{3}$$

where Y_i, $i = 1, 2, \ldots, n$ is the response value for the i^{th} case moisture content, estimates $\beta's$ are the regression coefficients for the predictor variables (drying parameter) $T's$, a_j denote heterogeneity, for $j = 1, 2, \ldots, f$. That is, the parameters that exhibit heterogeneity and ε is the random error. In the equation above, if the estimates of the regression equation are computed and a crucial variable is omitted, then the estimate β will be biased and inconsistent. It is also possible that some variables are correlated with the error term, which violates the assumption of regression. According to [45] the variance inflation factor in multiple regression is used to quantify the level of severity. The coefficient of determination can be written as $R^2 = 1 - \frac{1}{VIF}$ .If the R^2 satisfied certain conditions, then the parameter is said to exhibit heterogeneity. [45] stated that variance inflation factor in multiple regression is used to quantify the level of severity. It can be computed with R_l^2 where R_l^2 for $l = 1,2, \ldots, p$ denote the quantity of determination between the l^{th} variable x_l in the predictors matrix and the variables not related to it.

Let $\boldsymbol{X}^* = \begin{bmatrix} 1 & X_{11} & \ldots & X_{1,p-1} \\ 1 & X_{21} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \cdots & X_{n,p-1} \end{bmatrix}$. We can define $\boldsymbol{X}^{*\prime}\boldsymbol{X}^* = \begin{bmatrix} n & 0' \\ 0 & r_{XX} \end{bmatrix}$,So that r_{xx} will be the correlation matrix representing the $\boldsymbol{X}$ variables. Since

$\sigma^2\{\hat{\beta}\} = \sigma^2(\boldsymbol{X}^{*\prime}\boldsymbol{X}^*)^{-1} = \sigma^2 \begin{bmatrix} \frac{1}{n} & 0' \\ 0 & r_{XX}^{-1} \end{bmatrix}$. The VIF_l for $l = 1,2,3, \ldots, p-1$ stands for the l^{th} diagonal element of $r_{\boldsymbol{XX}}^{-1}$. If we can show the proof for $l = 1$, then the rows and columns $r_{\boldsymbol{XX}}$ can be permutated to obtain the result for the remaining l.

Let $\boldsymbol{X}_{(-1)} = \begin{bmatrix} X_{12} & \cdots & X_{1,p-1} \\ X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots \\ X_{n2} & \cdots & X_{n,h-1} \end{bmatrix}$, $X_1 = \begin{bmatrix} X_{11} \\ X_{21} \\ \vdots \\ X_{n1} \end{bmatrix}$. By applying Schur's complement, $r_{XX}^{-1}(1,1) = \left(r_{11} - r_{1X_{(-1)}} r_{X_{(-1)}X_{(-1)}}^{-1} r_{X_{(-1)}} 1\right)^{-1} = \left(r_{11} - \left[r_{1X_{(-1)}} r_{X_{(-1)}X_{(-1)}}^{-1}\right] r_{X_{(-1)}X_{(-1)}} \left[r_{X_{(-1)}X_{(-1)}}^{-1} r_{X_{(-1)}} 1\right]\right)^{-1} = \left(1 - \beta'_{1X_{(-1)}} X'_{(-1)} X_{(-1)} \beta_{1X(-1)}\right)^{-1}$,where $\beta_{1X_{(-1)}}$ means the regression coefficient of X_1 on $X_2, \ldots, X_{p-1}$ except the intercept. For clarity, R_1^2 and VIF_1 can be written as $R_1^2 = \frac{SSR}{SSTO} = \frac{\beta'_{1X_{(-1)}} X'_{(-1)} X_{(-1)} \beta_{1X_{(-1)}}}{1} = \beta'_{1X_{(-1)}} X_{(-1)} \beta_{1X_{(-1)}}$ and $VIF_1 = r_{XX}^{-1}(1,1) = \frac{1}{1-R_1^2}$.

Ridge Regression

In Equation 2, if the explanatory variables are nearly dependent, the matrix $X'X$ becomes ill conditioned. The ridge parameter k is important to manage the bias towards the mean of the dependent variable[46]. The standard errors are reduced and variance of the estimated parameters are reduced [47]. According to [47], suppose Y be the response vector and X predictor matrix, the ridge regression coefficient can be given by :

$$\hat{\beta}(K) = (X'XkI)^{-1}X'Y \tag{4}$$

where k denote the ridge parametr and I is the identity matrix. If $k = 0$, the estimate is $(X'X)^{-1}X'Y$ and if $k = 1$, $\hat{\beta}(k) = 0$. If we chose little positive values for k , it will improve the problem of conditioning and the variance of the estimates are reduced. The quantity of the shrinkage depends on k, it is the ridge penalty. If chose large values for k, it means more shrinkage. This means we are going to have different coefficient estimates for different values of k. One of the challenges in using the ridge regression is how to choose the value of k. Many authors have proposed how to select value for k in the literature. See for instance [48, 49, 50, 51, 52, 53, 54]; stated that a graphic can be used and called it the ridge trace. The plot may show the ridge coefficients to be a function of k. During the inspection of the ridge trace, k is chosen and the regression coefficients have a satisfactory magnitude, stability and sign, also the mean squared error (MSE) is not clearly inflated. Furthermore,

$$\widehat{K} = \frac{\hat{\sigma}^2}{\hat{\beta}^2_{max}} \tag{5}$$

The value of k selected is small enough, where the mean squared error of ridge estimator, is smaller than the mean squared error of OLS estimator.

$$\hat{\sigma}^2 = \frac{(Y-X\hat{\beta})'(Y-X\hat{\beta})}{n-p-1} \quad \text{(Hoerl \& Kennard, 1970)} \tag{6}$$

LASSO Regression

LASSO combines shrinkage and variable selection[55] and [56]. It can shrink the coefficients towards 0 when λ increases. Similarly, many coefficients shrink to exact value 0 when λ is sufficiently big. However, shrinkage improves the forecast accuracy because of bias variance tarde off. Though LASSO is a good method for achieving optimal prediction and consistent variable selection [57]. Let assume that we have data $X^i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$, $i = 1,2,3, \dots, n$ are the explanatory variables and y_i are the dependent variables. The assumption is that the observations are independent or the conditionality independent of $y_i s$ given the $x_{ij} s$. Also, x_{ij} is assumed to be standardised so that

$$\frac{\sum_i x_{ij}}{n} = 0, \sum_i x_{ij}^2 = 1 \text{ , i} = 1,2,3, \dots, \text{n} \tag{7}$$

$$\hat{\beta}(\text{LASSO}) = \text{argmin} \left\| y - \sum_{j=1}^{p} x_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{8}$$

where λ is a regularization parameter that is positive,the l_1 is the second part which is important for the LASSO [58].

Elastic Net Regression

Let p be the covariates with n observations. If $y = (x_{1j}, x_{2j}, \dots, x_{nj})^T$ be dependent variables and the matrix of the model $X = \left(X_1 \mid \dots \mid X_p\right)$, $x_j = (x_{1j}, x_{2j}, \dots, x_{nj})^T$, $j = 1,2, \dots, f$ are the independent variables. The assumption is that the dependent varaiable is centred. According to [59], the Elastic Net is explained as follows:

$\sum_{i=1}^{n} x_{ij} = 0$, $\sum_{i=1}^{n} y_i = 0$ and $\sum_{i=1}^{n} x_{ij}^2 = 1$, for $j = 1,2,3, \dots, p$ .If λ_1 and λ_2 are fixed and positive, the Elastic Net criterion is given as

$$L(\lambda_1, \lambda_2, \beta) = \|y - X\beta\|^2 + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|^2 \tag{9}$$

Where $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ and $\|\beta\|^2 = \sum_{j=1}^{p} \beta_j^2$.

The estimator $\hat{\beta}$ for the Elastic Net minimized the equation

$$\hat{\beta} = arg\min_{\beta}\{L(\lambda_1, \lambda_2, \beta)\} \quad (10)$$

The method used here is the penalised least square method. If $\frac{\lambda_2}{\lambda_1+\lambda_2} = \alpha$, $\hat{\beta} = arg\min_{\beta}\|y - X\beta\|^2$. This is the same with the problem of optimization $\hat{\beta} = arg\min_{\beta}\|y - X\beta\|^2$, subject to $\alpha\|\beta\|^2 + (1-\alpha)\|\beta\|_1 \leq t$ for some t. $\alpha\|\beta\|^2 + (1-\alpha)\|\beta\|_1$ is the Elastic Net penalty and it combines LASSO and Ridge.

Control Limit

Control limits are crucial components of statistical process control (SPC) charts, serving as akin to traffic lanes that indicate whether a process is stable and predictable. Control charts apply control limits to assist determine whether a process has dramatically altered or to isolate an unexpected occurrence. Because control limits are calculated from data, you won't know what they are until you've collected a representative set of data. A control chart always has the following types of lines, which are determined from previous data.

1) A central line (CL) is a horizontal graphical line that represents the mean or median of process measurements.
2) The upper control limit (UCL) is shown by a horizontal red line above the process average. Generally thought to be three times the standard variation of process measurements.
3) Lower line (LCL) is the lower control limit. This is shown as a horizontal red line below the process average. Generally thought to be three times the standard variation of process measurements. 'These control limits, determined from historical data, aid in identifying significant process changes or unusual events, enabling effective monitoring and improvement efforts [60].

Evaluation Metric

Evaluation metrics are critical in measuring the performance of machine learning models in regression analysis. These metrics, including Mean Absolute Percentage Error (MAPE), Mean Squared Error (MSE), Sum Squares of Error (SSE), and Coefficient of Determination (R^2), have been specifically developed for this purpose. Leveraging the R package 'metric' provides a comprehensive toolkit for evaluating prediction performance in regression models, which in turn streamlines model validation and assessment processes. These metrics are indispensable for ensuring the reproducibility and predictability of regression algorithms. Further elaboration on the benefits of each error metric is provided in their respective sections below.

Mean Absolute Percentage Error

The Mean Absolute Percentage Error (MAPE) is a significant metric of accurate predictions in a variety of industries, including agricultural forecasting. MAPE is calculated by considering the absolute percentage difference between predicted and actual values, making it a useful tool for evaluating the precision of forecasting methods [61]. The formula for MAPE is:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right| * 100$$

Where,
Y_i is the actual value for observation i
$\hat{Y}_i$ is the predicted (estimated) value for observation i
n is the total number of observations

Mean Squared Error

The MSK evaluation metric, while not a widely recognized acronym, might refer to a specific context or could be related to the well-known Mean Squared Error (MSE), a common metric used to assess the performance of predictive models. In the realm of statistical analysis, the Dynamic Linear Model (DLM) method plays a crucial role, especially in time series analysis. DLMs are valuable because they allow model parameters to change over time, making them particularly useful for analysing data in dynamic environments where underlying processes evolve, such as in financial modelling or precision farming. By continuously updating predictions with new data, DLMs provide adaptive and accurate insights, helping to refine decision-making processes.in this study, using Mean Squared Error (MSE) is a crucial metric in statistical analysis, reflecting both the bias (accuracy) and variance (precision) of an estimator

[62]. It calculates the average squared difference between the estimated values and the actual values of a parameter, providing a measure of how close the estimator is to the actual value [63, 64]. MSE is particularly valuable in regression analysis, where it helps compare the accuracy of different estimators. Furthermore, in image quality assessment, MSE serves as a reliable indicator when comparing images with similar bias/variance ratios. The formula for MSE is:

$$MSE = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}$$

where,
Y_i is the actual value for observation i
$\hat{Y}_i$ is the predicted (estimated) value for observation i
n is the total number of observations

Sum Squares of Error

The Sum Squares of Error (SSE) is a crucial metric utilized in various fields like optimization and clustering to evaluate the accuracy of models or algorithms [65, 66]. Additionally, SSE plays a pivotal role in assessing errors, optimizing models, and enhancing the accuracy of various analytical processes. SSE simply sums up the squares of the errors between the predicted and actual values. The formula for SSE is:

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Where,
Y_i is the actual value for observation i
$\hat{Y}_i$ is the predicted (estimated) value for observation i
n is the total number of observations

Coefficient of Determination

The Coefficient of Determination, often known as R^2, is a statistical metric that quantifies the proportion of the variation in a response variable that is foreseeable from the predictor variables in a regression model. Simply put, it shows how well the predictor factors explain the variability of the response variable [67]. The formula for R^2 is:

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}$$

Where,
Y_i is the actual value for observation i
$\bar{Y}$ is the mean value for observation i
$\hat{Y}_i$ is the predicted (estimated) value for observation i
n is the total number of observations

R^2 varies from 0 to 1. A score closer to 1 implies that the predictor variables account for a greater share of the variation in the response variable, implying that the model fits the data better. In contrast, a score closer to 0 indicates that the predictor factors do not explain much of the variance in the response variable.

Percentage Change

Percentage change is a metric that expresses the difference between two variables as a percentage of one of them. It's calculated using the formula:

$$\text{Percentage Change (\%)} = \frac{after\ value - before\ value}{before\ value} * 100\%$$

This formula gives you the percentage increase or decrease from the old value to the new value. If the result is positive, it indicates an increase, while a negative result indicates a decrease [68].

Results and Discussion

Based on the findings presented in Table 3, the variables T6, T7,T8, T11, H1, H5 and PY are heterogeneity. The variation inflation factor (VIF) values are increased, with the largest value at 75,337.29. It demonstrates the high amount of multicollinearity [30].

Figure 3 represent a box-and-whisker plot, A box-and-whisker plot, also known as a five-number summary, including the Minimum, First quartile (Q1), Second quartile (Q2), Third quartile (Q3) and Maximum values and whisker line, a visual representation used in statistical analysis to display key characteristics of a dataset [69, 70]. It provides insights into central tendency, dispersion, asymmetry, and extremes of the data, making it valuable for exploring distributions without assuming normality [71]. Box-and-whisker plots are particularly useful in identifying outliers, which can significantly impact forecasting accuracy [72]. They offer a practical way to introduce to statistical concepts, encouraging exploration and understanding of data properties like medians and means [73]. In educational settings, box-and-whisker plots are essential for teaching how to organize and interpret data effectively, connecting concepts of centre, spread, outliers, skewness, and measures of dispersion.

Figure 3 displays the distribution of 29 variables before the removal of heterogeneity parameters, enabling an assessment of the widest range among variables and the identification of those with significant outliers. Notably, variables H1, H5, T11, and T7 demonstrate the highest number of outliers. Furthermore, in variable PY, the whiskers extending from the box signify variability beyond the upper and lower quartiles, indicating heterogeneity. This suggests that H1, H5, T11, T7, and PY are all affected by heterogeneity parameters, thereby impacting forecasting accuracy.
In Figure 4, the distribution of 22 variables is presented after the removal of heterogeneity parameters. It appears that all variables demonstrate statistical significance after this adjustment. Additionally, the distribution in the box-and-whisker plots is observed to be positively skewed due to the shorter whisker towards the lower end (skewed right).

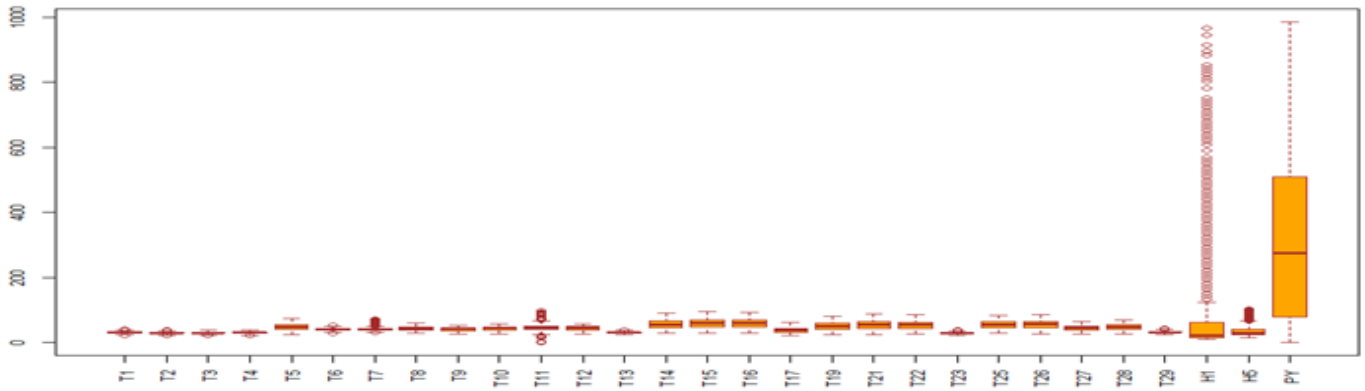

Figure 3. Box-and-whisker plot for 29 predictor variables before remove heterogeneity variables

Table 2. Heterogeneity variables

Smallest VIF	Largest VIF	Smallest R^2	Largest R^2	90% CI	Heterogeneity Variables
3.067297	75337.29	0.67398	0.999987	[0.786375,0.8875918]	T6, T7, T8, T11, H1, H5 , PY

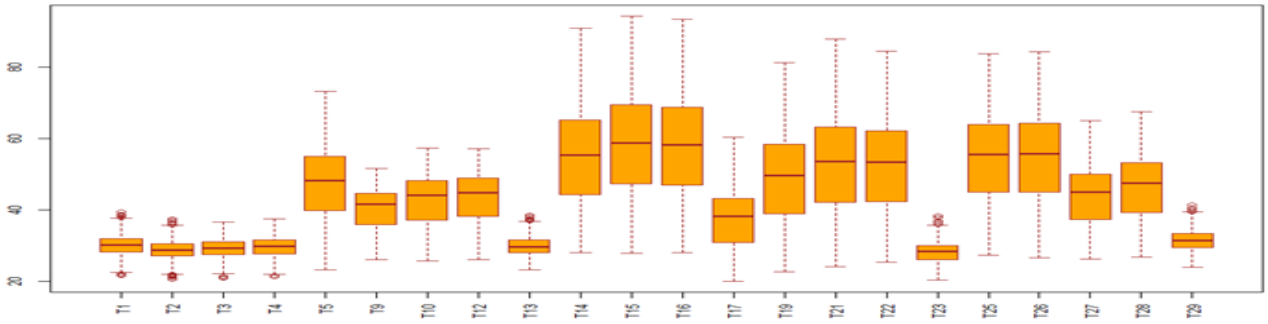

Figure 4. Box-and-whisker plot for 22 predictor variables after remove heterogeneity variables

Table 3. Evaluation metrics for the 50, 100, 150, 200, 250 and 300 high-ranking variables for before and after heterogeneity model

ML	High Ranking variables	Before heterogeneity				After heterogeneity				Percentage change %
		MAPE	MSE	SSE	R^2	MAPE	MSE	SSE	R^2	
Ridge	50	9.45909	41.5978	79618.2	0.847945	10.0197	45.1986	86510.2	0.834783	-14%
	100	8.30465	33.3634	63857.6	0.878044	8.99888	37.8396	72425.0	0.861682	-10%
	150	7.89390	30.6079	58583.6	0.888117	8.51171	34.4437	65925.2	0.874096	-10%
	200	7.67288	29.1636	55819.2	0.893396	8.19289	31.9777	61205.4	0.883110	-10%
	250	7.62552	28.7248	54979.3	0.895000	8.16304	31.6841	60643.4	0.884183	-9%
	300	**7.06351**	**25.8776**	**49529.7**	**0.905408**	**7.01913**	**25.7288**	**49245.0**	**0.905952**	**-10%**
LASSO	50	8.95830	38.8604	74378.9	0.857951	8.93358	38.5744	73831.5	0.858996	-14%
	100	8.82383	37.7268	72209.0	0.862095	8.81149	37.6604	72082.1	0.862337	-14%
	150	8.34749	34.2659	65584.9	0.874746	8.37318	34.2922	65635.4	0.874649	-13%
	200	8.33922	33.7814	64657.7	0.876517	8.32850	33.7935	64680.9	0.876472	-12%
	250	8.30830	33.5502	64215.2	0.877362	8.30903	33.6824	64468.2	0.876878	-12%
	300	**8.27855**	**33.1606**	**63469.4**	**0.878786**	**8.23513**	**33.0172**	**63195.0**	**0.879310**	**-12%**
Elastic Net	50	9.21914	41.1365	78735.3	0.849631	9.39125	42.8362	81988.6	0.843418	-14%
	100	8.90806	38.2638	73237.0	0.860132	8.75290	37.3240	71438.2	0.863567	-14%
	150	8.40840	34.4483	65934.1	0.874079	8.38454	34.3626	65770.0	0.874392	-13%
	200	8.37494	34.3496	65745.1	0.874440	8.34541	34.2235	65503.7	0.874901	-13%
	250	8.28678	33.6327	64373.0	0.877060	8.28987	33.7577	64612.2	0.876603	-12%
	300	**8.23751**	**33.1948**	**63534.8**	**0.878661**	**8.25215**	**33.4058**	**63938.7**	**0.877890**	**-12%**

Table 3 provides a summary of the predictive performance of regression algorithms, evaluated quantitatively before and after considering heterogeneity. The evaluation metrics include Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE), Sum Squares of Error (SSE), and Coefficient of Determination (R^2). The results are presented for different numbers of high-ranking variables: 50, 100, 150, 200, 250, and 300.

Before heterogeneity, the Ridge model exhibited decreasing values of MAPE, MSE, and SSE as the number of high-ranking variables increased. For instance, for 50 highest ranking variables, the MAPE was 9.459094 and the MSE was 41.59782, while for 300 highest ranking variables, the MAPE reduced to 7.063511 and the MSE decreased to 25.8776. Similarly, the LASSO and Elastic Net models also displayed decreasing trends in MAPE, MSE, and SSE with increasing numbers of high-ranking variables before heterogeneity.

After heterogeneity, similar decreasing trends were observed in MAPE, MSE, and SSE for all models across different numbers of high-ranking variables. For example, in the Ridge model, after addressing heterogeneity, the MAPE decreased from 10.0197 for 50 highest ranking variables to 7.01913 for 300 highest ranking variables. The same decreasing trends in MAPE, MSE, and SSE were observed in the LASSO and Elastic Net models after heterogeneity across varying numbers of high-ranking variables.

Before heterogeneity, the 300 high-ranking variables showed that the Ridge model had an R^2 value of 0.905408, Displaying 90.54% of the variance in the response variable was explained by the predictor variables. The LASSO model had an R^2 value of 0.878786, suggesting that 87.88% of the variance could be explained, while the Elastic Net model had an R^2 of 0.878661, explaining 87.87% of the variance. After heterogeneity, the Ridge model exhibited an improved R^2 value of 0.905952, explaining 90.60% of the variance. The LASSO model also saw a slight increase to an R^2 of 0.879310, explaining 87.93% of the variance, whereas the Elastic Net model showed a decrease to an R^2 of 0.877890, explaining 87.79% of the variance. Comparing these models, the Ridge model, post-heterogeneity, emerges as the best-performing one due to its highest R^2 value. This indicates that the Ridge model provides more significant results compared to the others.

The 300 high-ranking variables show a percentage change of -10% for Ridge, indicating a decrease of 10% before and after heterogeneity. Conversely, both LASSO and Elastic Net exhibit a percentage change of -12%, signifying a 12% decrease before and after heterogeneity. In comparison, Ridge stands out as the most significant, showcasing its superiority in performance. In statistical modelling, a smaller percentage change suggests that the model's performance is more stable and less affected by changes in conditions or variables. Therefore, Ridge appears to be more robust and reliable in maintaining its effectiveness across different conditions or scenarios. This interpretation suggests that Ridge may offer more consistent and dependable results compared to LASSO and Elastic Net in the context of this analysis.

In summary, by comparing the measure validation for after and before heterogeneity, the ridge is a good comparison among between LASSO and elastic net. The findings of various researchers [25, 74,75,76,77,78,79, 80] support the notion that Ridge regression stands out when compared to Lasso and Elastic Net. It demonstrates efficiency, particularly with larger sample sizes, although it differs from Lasso in its inability to shrink coefficients to zero. Ridge is favoured for its adeptness in handling multicollinearity while preserving model simplicity. By introducing a penalty term that reduces coefficients without eliminating them entirely, Ridge effectively manages multicollinearity. Moreover, it exhibits stability in variable selection, especially in high-dimensional datasets marked by collinearity. Notably, Ridge outperforms Lasso and Elastic Net in forecasting the moisture content of fish and analysing factors influencing fiscal revenue, attributed to its superior goodness of fit, smaller error rates, and enhanced model performance. Its reliance on L_2 norm penalization proves beneficial in managing multicollinearity in complex, high-dimensional settings. Overall, Ridge's ability to handle multicollinearity without reducing coefficients to zero makes it advantageous over Lasso and Elastic Net in various analytical contexts.

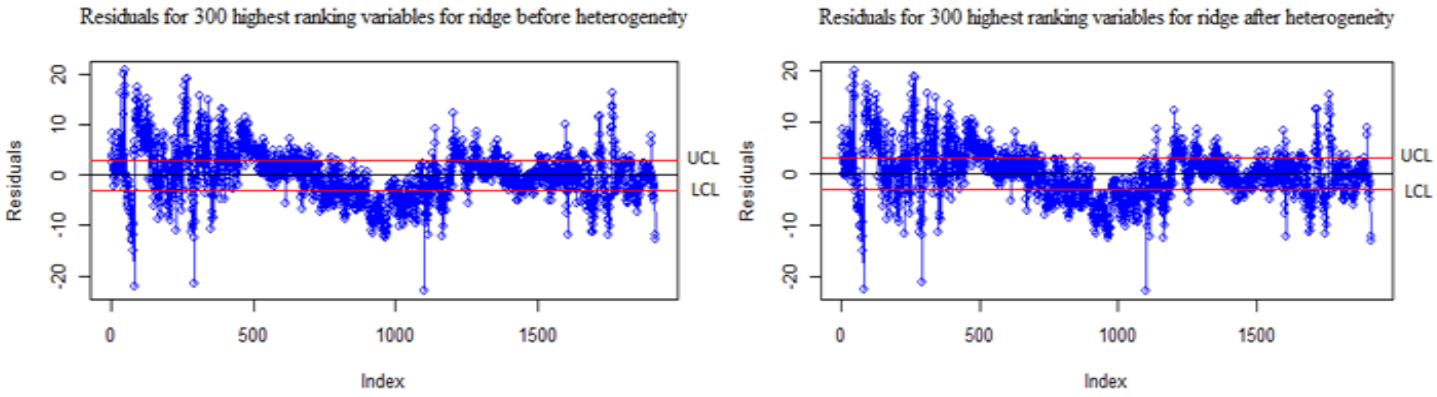

Figure 5. Comparison of standardized residuals for the top 300 variables ranked by Ridge before and after accounting for heterogeneity

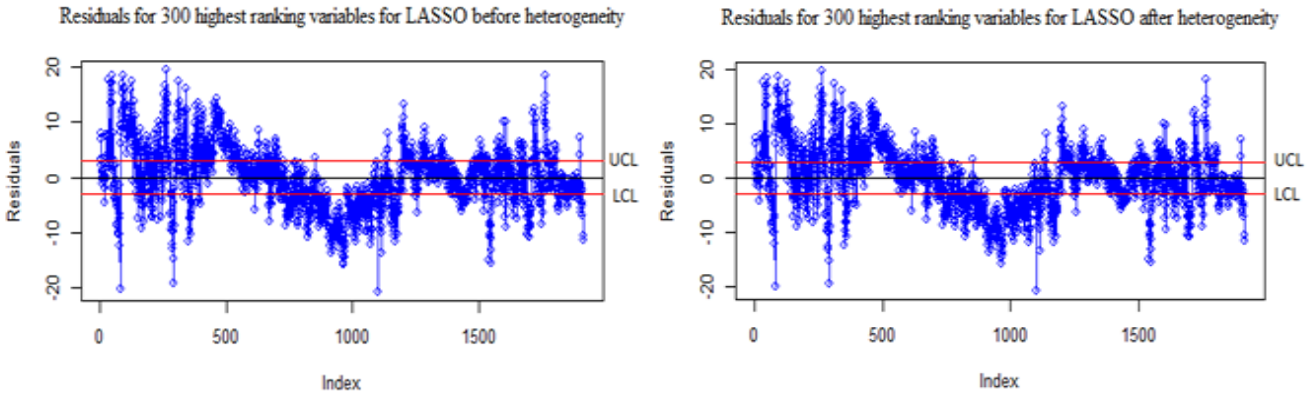

Figure 6. Comparison of standardized residuals for the top 300 variables ranked by LASSO before and after accounting for heterogeneity

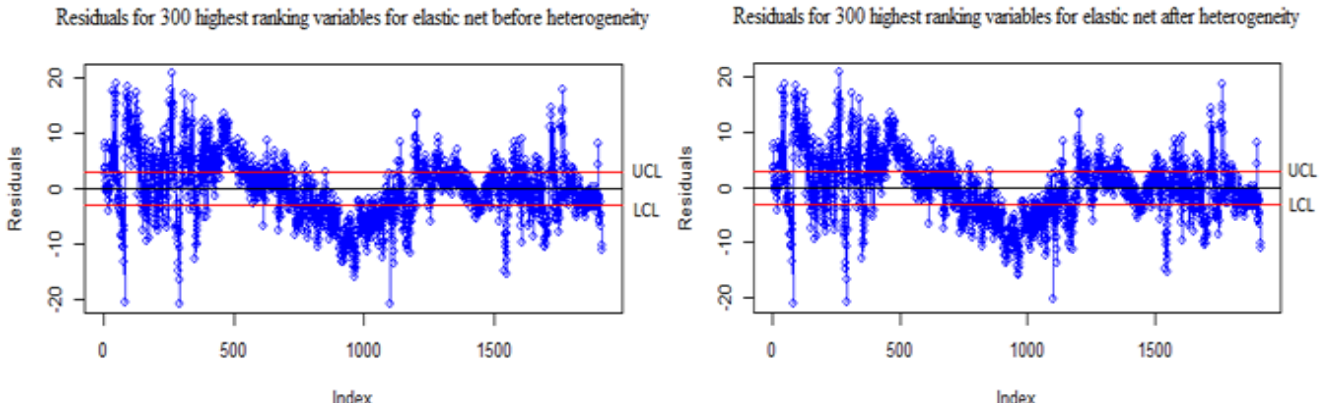

Figure 7. Comparison of standardized residuals for the top 300 variables ranked by elastic net before and after accounting for heterogeneity

Figures 5, 6, and 7 show standardized residuals plots for Ridge, LASSO, and elastic net before and after heterogeneity. Outliers outside of the 3-sigma limit can be noticed. UCL and LCL are the upper -and lower-class limits of 3 sigma and -3 sigma, respectively.

Figure 5 presents residual plots that demonstrate the impact of addressing heterogeneity in a ridge regression model for the 300 highest-ranking variables. Before heterogeneity, the residuals exhibit a wider spread with many points fallings outside the control limits, indicating potential outliers and a lack of model fit. After heterogeneity, the spread of residuals appears more compact, suggesting an improvement in the model's performance. However, some residuals still exceed the control limits, and residual patterns persist, implying that while the adjustment has led to some improvements, further refinement or alternative modeling techniques may be necessary to fully resolve the underlying issues.

Figure 6 displays the residual plots comparing the LASSO regression model's performance before and after addressing heterogeneity, focusing on the 300 highest-ranking variables. Initially, for before heterogeneity, the residuals are widely dispersed, with many points fallings outside the control limits, suggesting potential outliers and inadequate model fit. After heterogeneity, the residuals are more tightly clustered around the zero line, indicating improved model accuracy. Despite this improvement, some residuals still exceed the control limits, and persistent patterns suggest that while the adjustment has enhanced the model, further refinement or alternative techniques may be needed to fully address these issues. Figure 7 shows residual plots for 300 highest-ranking variables using elastic net regression before and after heterogeneity. Both plots display the residuals scattered around zero, with the upper and lower control limits (UCL and LCL)

indicated by red lines. The residuals appear more dispersed before addressing heterogeneity, with several points outside the control limits. After accounting for heterogeneity, the residuals are more tightly clustered within the control limits, suggesting that the model fit improved by accounting for heterogeneity, resulting in a reduction in unexplained variability.

All three regularization methods—LASSO, Ridge, and Elastic Net—show improvements in model performance after addressing heterogeneity, as evidenced by tighter clustering of residuals within the control limits (UCL and LCL). Before heterogeneity, the residuals for all methods were widely scattered, with several points exceeding the control limits, indicating that the models did not fully capture the data's variability. After heterogeneity, the residuals are more contained within the control limits, suggesting a better model fit. The reduction in residual spread is noticeable across all methods, with Elastic Net and LASSO showing particularly strong improvements. Ridge regression, while improved, still exhibits a few more outliers compared to the others, hinting that it might be slightly less effective in this scenario. Overall, addressing heterogeneity leads to better residual behavior and enhances model accuracy across all three techniques.

Conclusions

This study explores the variability in drying parameters and introduces a heterogeneity model within machine learning algorithms to enhance the accuracy of moisture content prediction. Ridge regression, LASSO, and Elastic Net were utilized for variable selection, and the performance of these predictive models was assessed using metrics like MSE, SSE, MAPE, and R-squared. The results show that the Ridge model outperforms the other models in predictive accuracy, both before and after accounting for heterogeneity. Future research should consider exploring additional machine learning algorithms, such as support vector machines, bagging, boosting, and random forests, for variable selection. These approaches could be used to analyse the effects of heterogeneity before and after adjusting for heterogeneity parameters. Furthermore, since this study did not address outliers, robust regression techniques such as M Huber, M Hampel, M Bi-Square, MM, and S estimators should be considered to manage this issue effectively. The developed model could also be applied to various fields, including medicine, engineering, and agriculture.

Conflicts of Interest

The author(s) indicate no conflicts related to their interests for the publication of this paper.

Acknowledgement

The authors are grateful to the “Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with Project Code: FRGS/1/2023/STG06/USM/02/6” for their fnancial support.

References

[1] IBM. (n.d.). *Smart farming*. IBM. https://www.ibm.com/topics/smart-farming
[2] Shanthakumari, G., Vignesh, A., Siva Harish, R. V., & Karthick, R. (2024). Advancements in smart agriculture: A comprehensive review of machine learning and IoT approaches. *Proceedings of the International Conference on Computing, Communication, and Internet of Things (IC3IoT)*. https://doi.org/10.1109/ic3iot60841.2024.10550268
[3] Sharanangat, K. (2024). Automated irrigation system in farming by solar energy. *Indian Scientific Journal of Research in Engineering and Management*, https://doi.org/10.55041/ijsrem34461
[4] Devi, T. B., & Kalnar, Y. (2021). Design consideration of smart solar dryer for precision drying: Smart solar dryer for precision drying. *Journal of Agricultural Sciences, 8*(2). https://doi.org/10.21921/JAS.V8I2.7297
[5] Villa-Medina, J. F., Porta-García, M. Á., Gutiérrez, J. M., & Porta-Gandara, M. (2023). Solar forced convection dryer for agriproducts monitored by IoT. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4513800
[6] Senthil Kumar, K. L., & Saravanan, B. (2020). Design and fabrication of solar dryer for dehydration of vegetables. *AIP Conference Proceedings*. https://doi.org/10.1063/5.0019403
[7] Sharma, K. P., Kothari, S., Panwar, N. L., Ram, M., & Patel, M. (2022). Influences of a novel cylindrical solar dryer on farmer’s income and its impact on environment. *Environmental Science and Pollution Research*. https://doi.org/10.1007/s11356-022-21344-1
[8] Agricorn. (2023, July). *Drying and dehydration material handling equipment*. https://www.agricorn.in/2023/07/drying-and-dehydration-material-handling-equipment.html
[9] Tagnamas, Z., Idlimam, A., & Lamharrar, A. (2023). Predictive models of beetroot solar drying process through

machine learning algorithms. *Renewable Energy, 219*(Part 2), 119522. https://www.elsevier.com/locate/renene

[10] Udomkun, P., Romuli, S., Schock, S., Mahayothee, B., Sartas, M., Wossen, T., Njukwe, E., Vanlauwe, B., & Müller, J. (2020). Review of solar dryers for agricultural products in Asia and Africa: An innovation landscape approach. *Journal of Environmental Management, 268*, 110730. https://doi.org/10.1016/j.jenvman.2020.110730

[11] Rizawi, J. A., Naqvi, S. M. D., Salehsulieman, O., Merhawikidanegebreslassie, J., Jemalosmansaleh, Y., Kahsay, Y. B., & Danaityonasamanuel, Y. (2022). Design of solar tray tomato drier. *Journal of Eco-Friendly Agriculture*, https://doi.org/10.5958/2582-2683.2022.00077.6

[12] Dhande, H. K., Shelare, S. D., & Khope, P. B. (2020). Developing a mixed solar drier for improved postharvest handling of food grains. *Agricultural Engineering International: The CIGR Journal*.

[13] Kristianto, F. P., & Salim, M. G. (2023). Simulation of a solar drier for Iroko wood (Chlorophora excelsa) in a tropical environment. *Eksergi, 20*(1). https://doi.org/10.31315/e.v20i1.8166

[14] Schmid, B., Navalho, S., Schulze, P. C., Van De Walle, S., Van Royen, G., Schüler, L., Maia, I. B., Bastos, C. R. V., Baune, M.-C., Januschewski, E., Coelho, A., Pereira, H., Varela, J., Navalho, J., & Cavaco Rodrigues, A. M. (2022). Drying microalgae using an industrial solar dryer: A biomass quality assessment. *Foods*. https://doi.org/10.3390/foods11131873

[15] Kang, H., Zhang, G., Mu, G., Zhao, C., Huang, H., Kang, C., Li, X., & Zhang, Q. (2022). Design of a greenhouse solar-assisted heat pump dryer for kelp (Laminaria japonica): System performance and drying kinetics. *Foods*. https://doi.org/10.3390/foods11213509

[16] Culaba, A. B., Atienza, A. H., Ubando, A. T., Mayol, A. P., & Cuello, J. L. (2021). Energy and exergy evaluation of an onshore solar dryer for seaweeds. *IOP Conference Series: Materials Science and Engineering*. https://doi.org/10.1088/1757-899X/1109/1/012042

[17] Del Rosario, E. Z., & Mateo, W. (2019). Hot water blanching pre-treatments: Enhancing drying of seaweeds (Kappaphycus alvarezii S.). *Open Science Journal, 4*(1). https://doi.org/10.23954/OSJ.V4I1.2076

[18] Del Rosario, E. Z., & Mateo, W. (2019). Hot water blanching pre-treatments: Enhancing drying of seaweeds (Kappaphycus alvarezii S.). *Open Science Journal, 4*(1). https://doi.org/10.23954/OSJ.V4I1.2076

[19] Tsukii, R., Imanishi, T., Iriyama, M., Ono, M., Suyama, A., & Miura, E. (2016). Drier and drying system.

[20] Ibidoja, O. J., Shan, F. P., Sulaiman, J., & Ali, M. K. M. (2023). Detecting heterogeneity parameters and hybrid models for precision farming. *Journal of Big Data, 10*(130). https://doi.org/10.1186/s40537-023-00810-8

[21] Mishra, N., Jain, S. K., Agrawal, N., Jain, N. K., Wadhawan, N., & Panwar, N. L. (2023). Development of drying system by using internet of things for food quality monitoring and controlling. *Energy Nexus, 11*, 100219. https://doi.org/10.1016/j.nexus.2023.100219

[22] Lu, C., Ge, M., Song, L., Wu, J., Pan, G., & Wang, H. (2023). Energy efficiency evaluation study on the air source heat pump drying system based on internet of things. *2023 IEEE IAS Global Conference on Renewable Energy and Hydrogen Technologies (GlobConHT)*, 1–8. https://doi.org/10.1109/GlobConHT56829.2023.10087498

[23] Nalendra, A. K., Wahvudi, D., Mujiono, T., & Fuad, N. (2022). IoT-Agri: IoT-based environment control and monitoring system for agriculture. *2022 International Conference on Industrial Cyber-Physical Systems (ICPS)*. https://doi.org/10.1109/ICIC56845.2022.10006964

[24] IoT-Agri: IoT-based environment control and monitoring system for agriculture. (2022). *2022 International Conference on Industrial Cyber-Physical Systems (ICPS)*. https://doi.org/10.1109/icic56845.2022.10006964

[25] Fusion of multiple sensors to implement precision agriculture using IoT infrastructure. (2023). *Preprints*. https://doi.org/10.20944/preprints202304.0119.v1

[26] Bhoyar, N. C. (2023). Smart agriculture system using IoT-based technology. *International Journal for Science Technology and Engineering*. https://doi.org/10.22214/ijraset.2023.50651

[27] Future IoT applications using artificial intelligence-based sensors: Agriculture. (2022). *2022 IEEE International Conference on Intelligent and Resilient Computing Applications (ICIRCA)*. https://doi.org/10.1109/icirca54612.2022.9985712

[28] Redondo, J. M., Siqueiros-García, J. M., Bustamante-Zamudio, C., Seara-Pereira, M. F., & Trujillo, H. (2022). Heterogeneity: Method and applications for complex systems analysis. *Journal of Physics: Conference Series, 2159*(1), 012013. https://doi.org/10.1088/1742-6596/2159/1/012013

[29] Javaid, A., Muthuvalu, M. S., Sulaiman, J., Ismail, M. T., & Ali, M. K. M. (2019). Forecast of the moisture ratio removal during the seaweed drying process using solar drier. *AIP Conference Proceedings, 2184*(1), 050016. https://doi.org/10.1063/1.5136404

[30] Javaid, A., Ismail, M. T., & Ali, M. K. M. (2019). Model selection for collector efficiency of seaweed drier by using LASSO and multiple regression analysis using 8SC. *AIP Conference Proceedings, 2184*(1), 050032. https://doi.org/10.1063/1.5136420

[31] Javaid, A., Ismail, M. T., & Ali, M. K. M. (2020). Efficient model selection of collector efficiency in solar dryer using hybrid of LASSO and robust regression. *Pertanika Journal of Science & Technology, 28*(1), 193–210. https://www.researchgate.net/publication/341089637

[32] Lim, H. Y., Fam, P. S., Javaid, A., & Ali, M. K. M. (2020). Ridge regression as efficient model selection and forecasting of fish drying using V-groove hybrid solar drier. *Pertanika Journal of Science & Technology, 28*(4), 1179–1202. https://www.researchgate.net/publication/344873493

[33] Javaid, A., Ismail, M. T., & Ali, M. K. M. (2021). Efficient model selection for moisture ratio removal of seaweed using hybrid of sparse and robust regression analysis. *Pakistan Journal of Statistics and Operational Research, 17*(3), 669–681. https://doi.org/10.18187/pjsor.v17i3.3641

[34] Mukhtar, M., Ali, M. K. M., Ismail, M. T., Hamundu, F. M., Alimuddin, Akhtar, N., & Fudholi, A. (2022). Hybrid model in machine learning–robust regression applied for sustainability agriculture and food security. *International Journal of Electrical and Computer Engineering, 12*(4), 4457–4468. https://doi.org/10.11591/ijece.v12i4.pp4457-4468

[35] Mukhtar, M., Ali, M. K. M., Javaid, A., Ismail, M. T., & Fudholi, A. (2021). Accurate and hybrid regularization–robust regression model in handling multicollinearity and outlier using 8SC for big data. *Mathematical Modelling of Engineering Problems, 8*(4), 547–556. http://iieta.org/journals/mmep

[36] Ibidoja, O. J., Shan, F. P., Mukhtar, Sulaiman, J., & Ali, M. K. M. (2023). Robust M-estimators and machine learning algorithms for improving the predictive accuracy of seaweed contaminated big data. *Journal of the Nigerian Society of Physical Sciences, 5*, 1137. https://doi.org/10.46481/jnsps.2022.1137

[37] Usman, M., Doguwa, S. I., & Alhaji, B. B. (2022). Comparing the prediction accuracy of ridge, lasso, and elastic net regression models with linear regression using breast cancer data. *Bayero Journal of Pure and Applied Sciences*. https://doi.org/10.4314/bajopas.v14i2.16

[38] Elastic gradient descent, an iterative optimization method approximating the solution paths of the elastic net. (2022). *arXiv*. https://doi.org/10.48550/arxiv.2202.02146

[39] Performance of lasso and elastic-net methods on non-invasive blood glucose measurement calibration modeling. (2023). *Barekeng: Journal of Mathematics and Its Applications, 17*(1), 37–42. https://doi.org/10.30598/barekengvol17iss1pp0037-0042

[40] Zhang, J., Nai, W., Luo, K., Leng, P., Yang, Z., Li, D., & Zhang, C. (2021). Elastic network regression based on differential evolution dragonfly algorithm with T-distribution parameters. *2021 IEEE International Conference on Artificial Intelligence and Big Data (ICAIBD)*. https://doi.org/10.1109/ICAIBD51990.2021.9459070

[41] Ali, M. K. M., Sulaiman, J., Md Yasir, S., & Ruslan, M. (2017). Cubic spline as a powerful tool for processing experimental drying rate data of seaweed using solar drier. *Malaysian Journal of Mathematical Sciences, 11*, 159–172.

[42] Gujarati, D. N., & Porter, D. C. (2004). *Basic econometrics* (5th ed.). McGraw-Hill/Irwin.

[43] Obadina, A., Oyewole, O., Sanni, L., & Abiola, S. S. (2006). Fungal enrichment of cassava peels proteins. *African Journal of Biotechnology, 5*(3), 302–304.

[44] Gormley, T. A., & Matsa, D. A. (2014). Common errors: How to (and not to) control for unobserved heterogeneity. *The Review of Financial Studies, 27*(2), 617–661.

[45] Cheng, J., Sun, J., Yao, K., Xu, M., & Cao, Y. (2022). A variable selection method based on mutual information and variance inflation factor. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 268*, 120652.

[46] Dorugade, A. V. (2014). New ridge parameters for ridge regression. *Journal of the Association of Arab Universities for Basic and Applied Sciences, 15*, 94–99.

[47] Khalaf, G., & Iguernane, M. (2014). Ridge regression and ill-conditioning. *Journal of Modern Applied Statistical Methods, 13*(2), 18.

[48] Dorugade, A. V., & Kashid, D. N. (2010). Alternative method for choosing ridge parameter for regression. *Applied Mathematical Sciences, 4*(9), 447–456.

[49] Pal, D., Bhattacharyya, T., Bhattacharyya, A., Biswas, S., Gangadharan, D., Raha, S., & Sinha, B. (2003). The extent of strangeness equilibration in quark-gluon plasma. *Pramana, 60*(5), 1083–1087.

[50] Kibria, B. M. G. (2003). Performance of some new ridge regression estimators. *Communications in Statistics-Simulation and Computation, 32*(2), 419–435.

[51] Muniz, G., & Kibria, B. M. G. (2009). On some ridge regression estimators: An empirical comparison. *Communications in Statistics—Simulation and Computation, 38*(3), 621–630.

[52] Khalaf, G., Månsson, K., & Shukur, G. (2013). Modified ridge regression estimators. *Communications in Statistics-Theory and Methods, 42*(8), 1476–1487.

[53] Troskie, C. G., & Chalton, D. O. (1996). A Bayesian estimate for the constants in ridge regression. *South African Statistical Journal, 30*(2), 119–137.

[54] Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73*(3), 273–282.

[55] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 58*(1), 267–288.

[56] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association, 101*(476), 1418–1429.

[57] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(2), 301–320.

[58] https://blog.lifeqisystem.com/control-limits-in-spc-chart

[59] Scott, M. R., & Willmott, C. J. (2023). Decomposition of the mean absolute error (MAE) into systematic and unsystematic components. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0279774

[60] Set, F. N., Low, H. C., & Quah, S. H. (2008). Mean squared error: A tool to evaluate the accuracy of parameter estimators in regression.

[61] Ng, S. F., Low, H. C., & Quah, S. H. (2008). Mean squared error: A tool to evaluate the accuracy of parameter estimators in regression (Min Ralat Kuasa Dua - Satu Kaedah untuk Menilai Kejituan Penganggar Parameter dalam Regresi).

[62] Schluchter, M. D. (2014). Mean square error. In *Wiley StatsRef: Statistics Reference Online*. https://doi.org/10.1002/9781118445112.STAT05906

[63] Bach, F., Cornacchia, E., Pesce, L., & Piccioli, G. (2023). Theory and applications of the sum-of-squares technique. *arXiv*. https://doi.org/10.48550/arXiv.2306.16255

[64] Theory and applications of the sum-of-squares technique. (2023). *arXiv*. https://doi.org/10.48550/arxiv.2306.16255

[65] Zhang, D. (2017). A coefficient of determination for generalized linear models. *The American Statistician, 71*(4), 310–316.

[66] On the use of percent change within rehabilitative ultrasound imaging research: A systematic review with Monte Carlo simulations. (2022). https://doi.org/10.31219/osf.io/k9qg4

[67] Cox, K. S., & Holcomb, Z. C. (2017). Box and whisker plot. In *Encyclopedia of Biostatistics*. https://doi.org/10.4324/9781003096764-22
[68] Banacos, P. C. (2011). Box and whisker plots for local climate datasets: Interpretation and creation using Excel 2007/2010.
[69] Pranatha, M. D. A., Pramaita, N., Sudarma, M., & Widyantara, I. M. O. (2018). Filtering outlier data using box whisker plot method for fuzzy time series rainfall forecasting. *2018 International Conference on Wireless and Telematics (ICWT)*. https://doi.org/10.1109/ICWT.2018.8527734
[70] Hall, B. (2006). Box and whisker plots.
[71] Lai, A., Menezes, E., Bennett, A. P., & Triantafyllou, M. S. (2022). Whisker sensor calibration and replication. *2022 IEEE OCEANS Conference*. https://doi.org/10.1109/OCEANS47191.2022.9977014
[72] Ghareeb, Z., Ali, S., & Al-Temimi, S. (2023). A comparative study between shrinkage methods (ridge-lasso) using simulation. *Periodicals of Engineering and Natural Sciences (PEN)*. https://doi.org/10.21533/pen.v11i2.3472
[73] Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data, 7*(1), 1–26.
[74] Ghareeb, Z., Ali, S., & Al-Temimi, S. (2023). A comparative study between shrinkage methods (ridge-lasso) using simulation. *Periodicals of Engineering and Natural Sciences (PEN)*. https://doi.org/10.21533/pen.v11i2.3472
[75] Usman, M., Doguwa, S. I., & Alhaji, B. B. (2022). Comparing the prediction accuracy of ridge, lasso, and elastic net regression models with linear regression using breast cancer data. *Bayero Journal of Pure and Applied Sciences*. https://doi.org/10.4314/bajopas.v14i2.16
[76] Autcha, A. (2022). The penalized regression and penalized logistic regression of Lasso and elastic net methods for high-dimensional data: A modelling approach. *IST Transactions on Applied Mathematics & Modeling*. https://doi.org/10.9734/bpi/ist/v3/1695b
[77] Khan, H. R., Bhadra, A., & Howlader, T. (2019). Stability selection for Lasso, Ridge, and Elastic Net implemented with AFT models. *Statistical Applications in Genetics and Molecular Biology*. https://doi.org/10.1515/SAGMB-2017-0001
[78] García-Nieto, J. P., García-Gonzalo, E., & Paredes-Sánchez, J. P. (2021). Prediction of the critical temperature of a superconductor by using the WOA/MARS, Ridge, Lasso, and Elastic-net machine learning techniques. *Neural Computing and Applications*. https://doi.org/10.1007/S00521-021-06304-Z
[79] Ahrens, A., Hansen, C., & Schaffer, M. E. (2019). LASSOPACK: Stata module for Lasso, square-root Lasso, Elastic Net, Ridge, adaptive Lasso estimation and cross-validation. *Research Papers in Economics*.
[80] Nie, R. (2022). Analysis of influencing factors of fiscal revenue in Beijing based on Ridge regression and Lasso regression model. *International Journal of New Developments in Engineering and Society*. https://doi.org/10.25236/ijndes.2022.060201