Published by
NIGERIAN SOCIETY OF PHYSICAL SCIENCES
Available online @ https://journal.nsps.org.ng/index.php/jnsps

J. Nig. Soc. Phys. Sci. 5 (2023) 1137

Journal of the Nigerian Society of Physical Sciences

Robust M-Estimators and Machine Learning Algorithms for Improving the Predictive Accuracy of Seaweed Contaminated Big Data

O. J. Ibidoja[a,b], F. P. Shan[b], Mukhtar[c], J. Sulaiman[d], M. K. M. Ali[b,*]

[a] *Department of Mathematics, Federal University Gusau, Gusau, Nigeria*
[b] *School of Mathematical Sciences, Universiti Sains Malaysia 11800 USM, Penang, Malaysia*
[c] *I-CEFORY (Local Food Innovation), Universitas Sultan Ageng Tirtayasa Indonesia*
[d] *School of Science and Technology, Universiti Malaysia Sabah, Kota Kinabalu, Sabah, Malaysia*

Abstract

A common problem in regression analysis using ordinary least squares (OLS) is the effect of outliers or contaminated data on the estimates of the parameters. A robust method that is not sensitive to outliers and can handle contaminated data is needed. In this study, the objective is to determine the significant parameters that determine the moisture content of the seaweed after drying and develop a hybrid model to reduce the outliers. The data were collected with sensors from the v-Groove Hybrid Solar Drier (v-GHSD) at Semporna, South-Eastern Coast of Sabah, Malaysia. After the second order interaction, we have 435 drying parameters, each parameter has 1914 observations. First, we used four machine learning algorithms, such as random forest, support vector machine, bagging and boosting to determine the significant parameters by selecting 15, 25, 35 and 45 parameters. Second, we developed the hybrid model using robust methods such as M. Bi-Square, M. Hampel and M. Huber. The results show that there is a significant improvement in the reduction of the number of outliers and better prediction using hybrid model for the contaminated seaweed big data. For the highest variable importance of 45 significant drying parameters of seaweed, the hybrid model bagging M Bi-square performs better because it has the lowest percentage of outliers of 4.08 %.

DOI:10.46481/jnsps.2023.1137

Keywords: Robust method, Hybrid model, Machine learning, Outliers, Big data.

Article History :
Received: 22 October 2022
Received in revised form: 08 January 2023
Accepted for publication: 08 January 2023
Published: 04 February 2023

Communicated by: Tolulope Latunde

1. Introduction

The purpose of regression analysis is to study the relationship between two or more independent variables and a dependent variable. Consider a multiple regression model:

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

where y is an $n \times 1$ vector of response variables, X is known as the design matrix of order $n \times p$, β is a $p \times 1$ vector of unknown parameters and ε is an $n \times 1$ 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 [1, 2], the ordinary least squares (OLS) estimator of β is

*Corresponding author tel. no: +60 14-9543405
Email address: majidkhanmajaharali@usm.my (M. K. M. Ali)

obtained as:

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

Observations that deviate from the distribution's general shape or pattern are called outliers [3]. The relationship between the observed and the dependent variable can be estimated by OLS regression, by minimizing the sum of squares [4]. OLS also has limitations when the assumptions are violated [5]. Estimates from OLS are not precise due to the high variances and covariances [6]. The presence of outliers in the data makes the LS estimator unstable, inefficient, and unreliable [7]. Agricultural data has outliers because of factors that cannot be regulated, and these outliers will increase the standard errors [4, 8]. The presence of outliers affects the performance of OLS, and a robust regression is used [9].

When modelling data using regression analysis, various assumptions are tested but these assumptions are violated. This model needs to be tested on the error structure for the necessary assumptions before prediction [10]. The researcher can transform the variables to fulfil the assumptions, but this cannot eradicate the outliers in the data that affect the forecast and estimate of the parameters [11]. Data with outliers is common in the field of agriculture [11, 12].

To overcome this problem, robust estimators have been introduced. M-estimation is the most common method of robust regression, it was introduced by [13], it is a generality to the method of maximum likelihood estimation. Before we used the robust methods to reduce the outliers, four machine learning algorithms such as random forest, support vector machine, boosting and bagging are used to select the significant parameters that determine the moisture content of the seaweed.

The major contributions of this study are:

i. To determine the significant parameters for the moisture content removal of seaweed during drying and reduce the number of outliers.
ii. To propose a hybrid model that combines robust M-estimators and machine learning models to improve the prediction accuracy.

2. Flowchart of the study

Figure 1 shows the flowchart of the various stages in the study.

2.1. Stage I

This involves the inclusion of all possible models.

$$\frac{n!}{(n-r)!r!} + \text{number of single factor}, \tag{3}$$

where n is the number of single factors, r is the number of orders. Equation (3) can be used to compute the total number of all possible models.

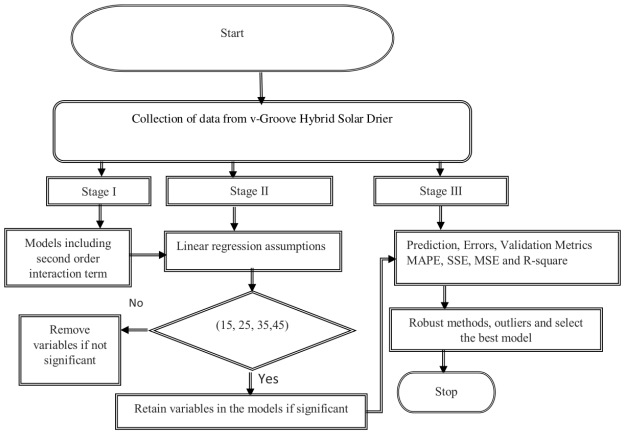

Figure 1: Flowchart of the procedure for the hybrid model

2.2. Stage II

Test for the assumptions of linear regression. The residual vs fitted plot, normal Q-Q plot are Kolmogorov-Smirnov test are used to verify the assumptions. Next, each machine learning model is used to select 15, 25, 35 and 45 highest important variables for optimization and easy comparison, to determine the moisture content removal of the seaweed after drying. We selected the number of variables because features selection can only provide the rank of important variables and does not tell us the number of significant factors [14]. Similarly, there is no rule to decide the number of parameters to be included in a prediction model [15]. Furthermore, the algorithms cannot tell us the number of significant variables except the ranks [16].

2.3. Stage III

After the selection of the significant parameters, the prediction is done and the validation metrics such as MAPE, SSE, MSE and R-square are computed. The outliers are also computed, and the robust method is introduced to build the hybrid model.

3. Materials and Methods

3.1. Data Description

The data were collected from 8th April 2017 to 12th April 2017, between the hours of 8:00 am to 5:00 pm during the drying of seaweed by using v-Groove Hybrid Solar Drier (v-GHSD) at Semporna, South-Eastern Coast of Sabah, Malaysia. There are 435 parameters after the inclusion of the second order interaction in this study.

3.2. Machine learning algorithms

Machine learning can learn from data and use the algorithms to understand and forecast the future [17]. Machine learning algorithms can be used to determine the rank of significant explanatory variables that contribute significantly to the response variable. These high-ranking variables selected using variable importance can reduce the training time, complexity

of the model and improve accuracy [18]. Four machine learning algorithms such as random forest, support vector machine, bagging and boosting are used in this study, to determine the significant parameters that determine the moisture content removal of the seaweed.

3.2.1. Random Forest

A random forest (RF) is a mixture of classification and regression trees (CARTs). It uses the highest number of votes (classification) or the mean forecasts (regression) of all the trees [19]. It uses the idea of bagging, and it is an ensemble learning method [20], [21].

If $\mathcal{L}$ is a learning set ,with a group of $\mathbb{N}$ pairs of features, with the output $(x_1, y_1), (x_2, y_2), (x_3, y_3) \ldots (x_N, y_N)$, if $x_i \in X$ and $y_i \in Y$. A class of p-features x_i $(for\ i = 1, 2, \ldots, N)$ is a $N \times p$ matrix X,where the rows $i = 1, 2, ..., N$ relates as x_i, with columns $j = 1, 2, 3, ..., p$ as x_j.
Algorithm:
For $b = 1$ to n

1. Create a bootstrapped sample D_b^* from the training set D.
2. Grow the tree by using the m from the bootstrapped sample D_b^*.

For a specific mode

i. Select m variables randomly.
ii. Identify the top split variables and values.
iii. Divide a node using the top divided variables and values.

Replicate the steps 1–3 till the stopping conditions are satisfied.

3.2.2. Suppor Vector Machine (SVM)

Support vector machine can be used for regression and classification problems [22]. SVM has the capacity to reveal nonlinear connections with kernel function [20, 23]. The SVM was developed by Cortes & Vapnik [24]. A good tutorial and explanations were given by [25, 26]. In support vector regression, the ϵ loss function is usually minimized. Beyond this particular bound, a straightforward linear loss function is applied, and any loss less than ϵ is set to zero:

$$L_\epsilon = f(x) = \begin{cases} 0, & \text{if } |y_i - f(x_i)| < \epsilon \\ y_i - f(x_i) - \epsilon|, & \text{otherwise} \end{cases} \quad (4)$$

For instance, suppose $f(x)$ is a linear function $f(x) = \beta_0 + x_i^t\beta$, then the loss function is given as

$$\sum_{i=1}^{n} \max\left(y_i - x_i^t\beta - \beta_0 - \epsilon, 0\right) \quad (5)$$

The ϵ is the tuning parameter and can be written as the constrained optimization problem:
Minimize

$$\frac{1}{2}\|\beta\|^2 \quad (6)$$

Subject to

$$\begin{cases} y_i - x_i^t\beta - \beta_0 \le \varepsilon \\ -\left(y_i - x_i^t\beta - \beta_0\right) \le \varepsilon \end{cases}. \quad (7)$$

If there are observations who do not lie within the ε band around that regression line,then there is no solution to the problem. The slack variables ζ_i and ζ_i^* are used ,this allows the observations to fall outside the ε band around that regression line.
Minimize

$$\frac{1}{2}\|\beta\|^2 + K\sum_{i=1}^{n}(\zeta_i + \zeta_i^*) \quad (8)$$

Subject to

$$\begin{cases} y_i - x_i^t\beta - \beta_0 \le \varepsilon + \zeta_i \\ -\left(y_i - x_i^t\beta - \beta_0\right) \le \varepsilon + \zeta_i^* \\ \zeta_i, \zeta_i^* \ge 0 \end{cases} \quad (9)$$

3.2.3. Boosting

Boosting is used to improve the accuracy of algorithms [27]. Boosting starts with an algorithm or method to discover the rough rules of thumb. It is called the "base" or "weak" learning algorithm many times. The base learning algorithm creates a new weak prediction rule each time it is called, and after many rounds, the boosting algorithm must merge these weak rules into a singular forecast rule that, ideally, will be significantly more precise than any of the weak rules [28]. Suppose we have this model matrix $X = \left[X_1, X_2, \ldots, X_p\right]\epsilon\mathbb{R}^{n\times p}$, outcomes variable vector $y \in \mathbb{R}^{n\times 1}$. The regression coefficients vector is given as $\beta \in \mathbb{R}^p$, the value of predicted for the outcome variable is denoted by $X\beta$, and the residuals are denoted by $\varepsilon = y - X\beta$. For regression purposes, least squares boosting (LSB(ε)) gives an accurate description of the data and regularization [27].

The algorithm for LSB(ε) is as follows:
Algorithm: LSB (ε)
Choose the rate of learning $\varepsilon > 0$ and iterations number N.
Define at $\hat{\beta}^0 = 0$, $\hat{r}^0 = y$, $k = 0$.

1. Do this for $0 \le k \le N$
2. Establish the covariates $\widetilde{u}_{j_k}$ and j_k as below:

$$\hat{u}_n = \underset{u\in\mathbb{R}}{\operatorname{argmin}}\left(\sum_{i=1}^{n}\left(\hat{r}_i^k - x_{\text{in}}u\right)^2\right)^2 \quad \text{for } n = 1, 2, 3, \ldots, p,$$

$$j_k \in \underset{1\le n\le p}{\operatorname{argmin}}\sum_{i=1}^{n}\left(\hat{r}_i^k - x_{\text{in}}\widetilde{u}_n\right)^2$$

3. Revise the present errors and regression coefficients as:

$$\hat{r}^{k+1} \leftarrow \hat{r}^k - \varepsilon\widetilde{u}_{j_k}$$
$$\hat{\beta}_{j_k}^{k+1} \leftarrow \hat{\beta}_{j_k}^k + \varepsilon\widetilde{u}_{j_k} \text{ and } \hat{\beta}_j^{k+1} \leftarrow \hat{\beta}_j^k,\ j \ne j_k$$

3.2.4. Bagging

Breiman [29] introduced bagging (bootstrap aggregating) to decrease the variance of classification and regression tree models. It is used to improve the present method and leads to an improvement in the accuracy. Bagging is used as an intensive methods to enhance erratic estimation. For a high - dimensional

data problems, bagging can be used to find a good model. Suppose we have a feature $\varphi(\mathbf{x}, \mathcal{L})$ to predict y from x, if there is a training sequence $\{\mathcal{L}_k\}$ consisting of N objects , from $\mathcal{L}$ distribution, the aim here is to use the $\{\mathcal{L}_k\}$ to build a more accurate predictor than $\varphi(\mathbf{x}, \mathcal{L})$ as a specific training set predictor $\varphi(\mathbf{x}, \mathcal{L})$ [29]. If y is not discreet and we put $\varphi(x, \mathcal{L}_k)$ with the mean of $\varphi(x, \mathcal{L}_k)$ over k. We get continually many samples via the bootstrap $\left\{\mathcal{L}^{(A)}\right\}$, an from $\mathcal{L}$, and form $\left\{\varphi\left(x, \mathcal{L}^{(A)}\right)\right\}$. If y is continuous, then φ_A as $\varphi_A(\mathbf{x}) = \text{average}\varphi_A\left(\mathbf{x}, \mathcal{L}^{(A)}\right)$. The $\left\{\mathcal{L}^{(A)}\right\}$ will form replicate datasets with M cases are randomly chosen from $\mathcal{L}$ and by applying replacement. Each (y_m, x_m) can appear many times in a any specific $\mathcal{L}^{(A)}$. The technique to construct φ is an important factor to know if bagging improves precision or reliability.

Theoretically bagging is described as follows:

i. Build a bootstrap sample $L_i^* = \left(Y_i^*, X_i^*\right)$ $(i = 1, 2, 3, \ldots, m)$ centred on an empirical distribution of these pairs $L_i = (Y_i, X_i)$ $(i = 1, 2, 3, \ldots, m)$.
ii. Use the plug-in principle to ascertain the bootstrapped predictor $\hat{\theta}_m^*(x)$; which is, $\hat{\theta}_m^*(x) = g_m(L_1, L_2, L_3, \ldots, L_m)(x)$.
iii. $\hat{\theta}_{m;B}(x) = \mathbb{E}^*\left[\hat{\theta}_m^*(x)\right]$ means the bagged predictor.

The bagging algorithm is as follows:
Input: Data $D = \{(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_m, y_m)\}$;
Learning algorithm base $\mathcal{L}$;
Base learner's numbers j.
Process:
For $j = 1, 2, \ldots, J$:
$\text{bs}_j = \text{bootstrap}(D)$; %Create the bootstrap sample from D
$\theta_j = \mathcal{L}\left(\text{bs}_j\right)$ % Train the base learner θ_j from the bootstrap sample
End
Output: $\frac{1}{J}\sum_{j=1}^{J}\theta_j(x)$ % For regression studies

3.3. Robust Estimation Method

Outliers are common with contaminated data and how to determine the observations is a challenge. A robust method can deal with the influence of outliers. Contaminated data can be analyzed using robust estimation [6], [30, 31, 32]. A robust method is used to solve the problems of traditional methods because of these outliers. To know the best method for the robust estimation methods, M estimation methods M Huber, M Hampel and M Bi-Square are compared.

The M-estimation method attempts to minimise that the function $\rho(\bullet)$ operates on the residual. M-estimators define:

$$\hat{\beta}_M = \underset{\beta}{\text{argmin}}\sum_{i=1}^{n}\rho(e_i(\beta)). \tag{10}$$

The ρ is $\rho-$type M-estimation. Assume σ is known and the residuals approximate β be $e_i = y_i - \beta^T x_i$. The β in M-estimate minimizes the objective function:

$$\sum_{i=1}^{n}\rho\left\{\frac{e_i(\beta)}{\sigma}\right\}. \tag{11}$$

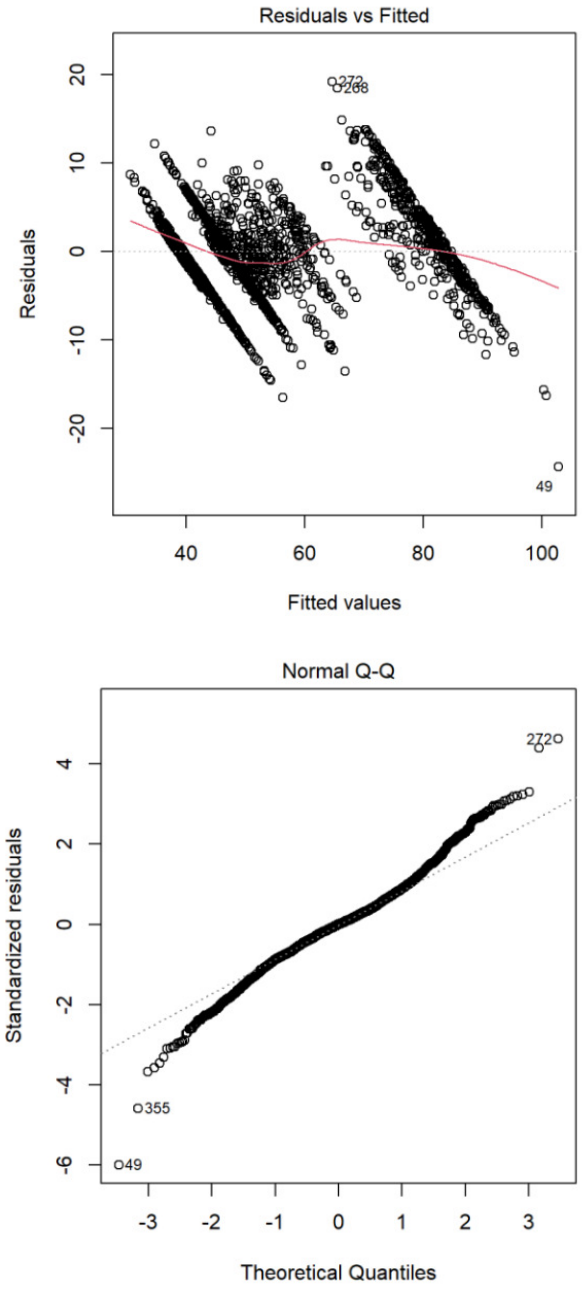

Figure 2: (a) Residuals vs Fitted (b) Residuals vs Normal Q-Q

The σ robustly estimate and the scale $\widetilde{\sigma}_M$ in M-estimator has solution:

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{e_i}{\sigma}\right) = \frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{y_i - \beta^T x_i}{\sigma}\right) = k, \tag{12}$$

where the β has the $p\times 1$ parameter vector, and then the function ψ yields:

$$\sum_i \psi(e_i)\frac{\partial e_i}{\partial \beta_i}, \quad \text{for} \quad j = 1, 2, \ldots, p. \tag{13}$$

The function $\psi(e) = \frac{\partial\rho(e)}{\partial(e)}$ derivatives the influence function.

Table 1: Robust Regression M-estimation Description

Methods	Objective Function	Weight Function
Bi-Square	$\begin{cases} \frac{k^2}{6}\left\{1-\left[1-\left(\frac{e}{k}\right)^2\right]^3\right\} & \text{for } \lvert e\rvert \le k \\ \frac{k^2}{6} & \text{for } \lvert e\rvert > k \end{cases}$	$\begin{cases} \left[1-\left(\frac{e}{k}\right)^2\right]^2 & \text{for } \lvert e\rvert \le k \\ 0 & for \lvert e\rvert > k \end{cases}$
Huber	$\begin{cases} \frac{1}{2}e^2 & \text{for } \lvert e\rvert \le k \\ k\lvert e\rvert - \frac{1}{2}k^2 & \text{for } \lvert e\rvert > k \end{cases}$	$\begin{cases} 1 & for\ \lvert e\rvert \le k \\ \frac{k}{\lvert e\rvert} & \text{for } \lvert e\rvert < k \end{cases}$
Hampel	$\begin{cases} \frac{e^2}{2}, & 0 < \lvert e\rvert < a \\ a\lvert e\rvert - \frac{e^2}{2}, & b < \lvert e\rvert \le c \\ \frac{-a}{2(c-b)}(c-e)^2 + \frac{a}{2}(b+c-a), & b < \lvert e\rvert \le c \end{cases}$	$\begin{cases} 1 & for\ 0 < \lvert e\rvert < a \\ \frac{a}{\lvert e\rvert} & \text{for } b < \lvert e\rvert \le c \\ a\frac{\frac{c}{\lvert e\rvert}-1}{c-b} & \text{for } b < \lvert e\rvert \le c \end{cases}$

Table 2: Kolmogorov-Smirnov Test for Normality

Test Statistic Value	P-value	Remarks
0.1641	2.2e-16	The residuals do not come from a normal distribution.

Then the weight function defines:

$$w(e) = \frac{\psi(e)}{e}, \tag{14}$$

where function $\psi(e)$ states:

$$\sum_i w(e_i)\, e_i \frac{\partial e_i}{\partial \beta_i} = 0, \text{for } \; j = 1, 2, \ldots, p \tag{15}$$

and the object becomes to obtain the following iterated re-weighted least square problem:

$$\min \sum_i w\left(e_i^{(k-1)}\right) e_i^2, \tag{16}$$

where k indicates the iterate number.

Table 1 shows the summary of the M - estimators and their respective weight function.

4. Results and Discussion

From the plot in Figure 2a, the residuals vs fitted plot shows that there is no pattern since the residuals did not spread out. There is evidence of non-linearity and heterogeneity. Figure 2b shows the normal Q-Q plot, the residuals are not normally distributed, this also supports the result of Kolmogorov-Smirnov test in Table 2. The possible outliers are the observations 272 and 355. The observation 272 determine more the moisture content removal of the seaweed than the model predict. Though, it is an extreme case, but still affect the moisture content removal. The observation 355 has a negative residual and determine less the moisture content removal of the seaweed than the model predicts.

The normality assumption is checked with the Kolmogorov-Smirnov test for a two-taied test. From the results in Table 2, the p-value =2.2e-16, which is less than 0.05, it means we have enough evidence to say that the residuals do not come from a normal distribution. This also explains why we have this type of QQ plot in Figure 2.

The results in Table 3 are the evaluation of each machine learning algorithm for 15, 25, 35 and 45 high - ranking variables that determine the moisture content removal of the seaweed. Based on the mean absolute percentage error (MAPE), mean squared error (MSE), R^2 and sum of squared error (SSE), random forest outperforms support vector machine, bagging and boosting for the 15, 25, 35 and 45 significant parameters. This also confirms the results of [33], where random forest absolutely performed better than the other methods.

Random forest when 45 significant parameters that determine the moisture content of the seaweed were selected gave MAPE of 2.125891, MSE of 7.330011, R^2 of 0.9732063 and SSE of 14029.64 gave the best performance. All the validation measures such as MAPE, MSE, R-square and SSE imply that significantly better results are obtained by random forest to the determine the moisture content removal of the seaweed.

Table 4 is the summary of the original model without using robust method and the hybrid models ,which combines machine learning models and robust estimation techniques. It also shows the number and percentage of outliers using 2-sigma limit.The percentage for the outliers is the number of observations outside the 2-sigma limit. It shows the percentage of outliers outside the 2-sigma limit for the original model without using robust method and the hybrid model. This sigma limit can improve the outputs quality and eliminate the source of deficiencies [34].

Based on the results in Table 4 for the original model, for

Table 3: Evaluation metrics for the 15, 25, 35 and 45 high-ranking important variables

Machine Learning Model	High-ranking important variables selected	High-ranking important variables selected			
		MAPE	MSE	R^2	SSE
Random Forest	15	2.458969	9.910512	0.9637737	18968.72
	25	2.337353	9.010273	0.9670644	17245.66
	35	2.174667	7.790909	0.9715216	14911.80
	45	**2.125891**	**7.330011**	**0.9732063**	**14029.64**
Support Vector Machine	15	8.614626	45.25618	0.8347612	86620.32
	25	7.980399	35.80985	0.8691446	68540.05
	35	7.568951	34.00095	0.8757802	65077.81
	45	7.351331	32.38644	0.8816661	61987.65
Bagging	15	12.25897	74.29053	0.7284423	142192.10
	25	9.778194	47.33173	0.8269861	90592.93
	35	8.413645	36.41955	0.8668739	69707.02
	45	8.151903	33.65611	0.8769752	64417.80
Boosting	15	8.168942	142.4542	0.5310293	272657.30
	25	8.697362	136.3236	0.5543729	260923.30
	35	8.183671	140.1463	0.5368431	268240.10
	45	8.203304	134.0864	0.5569358	256641.30

Table 4: Percentage of outliers outside 2 - sigma limits for hybrid models

Machine Learning Model	Robust Regression Method	15 highest important variables $\mu \pm 2\sigma$(%)	25 highest important variables $\mu \pm 2\sigma$ (%)	35 highest important variables $\mu \pm 2\sigma$ (%)	45 highest important variables $\mu \pm 2\sigma$ (%)
Random Forest	Original	118(6.17)	113(5.90)	112(45.85)	118(6.17)
	M Bi-Square	118(6.17)	117(6.11)	75(3.92)	99(5.17)
	M Hampel	72(3.76)	88(4.60)	92(4.81)	93(4.86)
	M Huber	83(4.34)	90(4.70)	88(4.60)	102(5.33)
Support Vector Machine	Original	108(5.64)	98(5.12)	86(4.49)	87(4.55)
	M Bi-Square	64(3.34)	18(0.94)	84(4.39)	89(4.65)
	M Hampel	66(3.45)	62(3.24)	85(4.44)	86(4.49)
	M Huber	81(4.23)	83(4.34)	96(5.02)	99(5.17)
Bagging	Original	98(5.12)	96(5.02)	97(5.07)	**84(4.39)**
	M Bi-Square	126(6.58)	97(5.07)	95(4.96)	**78(4.08)**
	M Hampel	101(5.28)	97(5.07)	90(4.70)	85(4.44)
	M Huber	113(5.90)	99(5.17)	97(5.07)	89(4.65)
Boosting	Original	193(10.10)	168(8.78)	194(10.12)	194(10.12)
	M Bi-Square	77(4.02)	77(4.02)	133(6.95)	79(4.12)
	M Hampel	76(3.97)	76(3.97)	72(3.76)	80(4.18)
	M Huber	83(4.34)	81(4.23)	67(3.50)	85(4.44)

15 highest important variables, the maximum is boosting with 193 (10.1%) outliers, while the minimum is bagging with 98 (5.12%). For the 25 highest important variables, the maximum is boosting 168 (8.78%) , while the minimum is bagging with 96 (5.02%). For the 35 highest important variables, the maximum is boosting 194 (10.12%), while the minimum is

support vector machine with 86 (4.49%). For the 45 highest important variables, the maximum is boosting 194 (10.12%) , while the minimum is bagging with 84 (4.39%). Based on this results, bagging with 45 variables importance gave the best performance because it has the lowest number of outliers of 84.

Based on the results in Table 4 for the hybrid model, for the 15 highest important variables, bagging M Bi-square has the highest number of outliers of 126 with 6.58% ,while support vector machine M Bi-square has the lowest number of outliers with of 64 with 3.34%. For the 25 highest important variable random forest M Bi-square has the highest number of outliers of 117 with 6.11% ,while support vector machine M Bi-square has the lowest number of outliers with of 18 with 0.94%. For the 35 highest important variable boosting M Bi-square has the highest number of outliers of 133 with 6.95% ,while boosting M Huber has the lowest number of outliers with of 67 with 3.50%. For the 45 highest important variable random forest M Huber has the highest number of outliers of 102 with 5.33% ,while bagging M Bi-square has the lowest number of outliers with of 78 with 4.08%. Based on this result, bagging M Bi-square gave the best performance because it had the lowest number of outliers of 78 and used the highest number of high ranking variables.

5. Conclusion

The aim of this study is to develop a hybrid model, to forecast seaweed drying parameters that determine the moisture content removal that would enhance the quality of the seaweed. Four predictive models such as random forest, support vector machine, bagging and boosting were built with M Huber, M Hampel and M Bi-Square to develop a hybrid model that can improve the predictive accuracy of the seaweed contaminated data. In summary, the best model to determine the moisture content removal of the seaweed big data is the bagging M Bi-square, it gave the best performance because it had the lowest number of outliers of 78 and used the highest number of high - ranking variables. For future study, a hybrid model with imbalanced data or missing values can be investigated.

Acknowledgement

The authors are grateful to the Ministry of Higher Education Malaysia for Fundamental Research Grant Scheme with Project Code RGS/1/2022/STG06/USM/02/13 for their assistance. We are also grateful to the Editor, associate editor, and anonymous reviewers for their insightful comments and suggestions to improve the quality and clarity of the paper.

References

[1] D. N. Gujarati & D. N. Porter, *Basic econometrics*, 4th ed. New York, USA: The McGraw-Hill Companies, (2004).

[2] O. G. Obadina, A. F. Adedotun, & O. A. Odusanya, "Ridge Estimation's Effectiveness for Multiple Linear Regression with Multicollinearity: An Investigation Using Monte-Carlo Simulations", Journal of the Nigerian Society of Physical Sciences **3** (2021) 278, doi: 10.46481/jnsps.2021.304.

[3] A. B. Yusuf, R. M. Dima, & S. K. Aina, "Optimized Breast Cancer Classification using Feature Selection and Outliers Detection," Journal of the Nigerian Society of Physical Sciences **3** (2021) 298, doi: 10.46481/jnsps.2021.331.

[4] H. Y. Lim, P. S. Fam, A. Javaid, & M. K. M. Ali, "Ridge regression as efficient model selection and forecasting of fish drying using v-groove hybrid solar drier", Pertanika J Sci Technol. **28** (2020) 1179, doi: 10.47836/pjst.28.4.04.

[5] A. Javaid, M. T. Ismail, & M. K. M. Ali, "Comparison of Sparse and Robust Regression Techniques in Efficient Model Selection for Moisture Ratio Removal of Seaweed using Solar Drier", Pertanika J. Sci. & Technol **28** (2020) 609.

[6] A. Javaid, M. T. Ismail, & M. K. M. Ali, "Efficient Model Selection of Collector Efficiency in Solar Dryer using Hybrid of LASSO and Robust Regression", Pertanika J. Sci. & Technol **28** (2020) 210.

[7] I. Dawoud & M. R. Abonazel, "Robust Dawoud–Kibria estimator for handling multicollinearity and outliers in the linear regression model", J. Stat. Comput. Simul. **91** (2021) 3678, doi: 10.1080/00949655.2021.1945063.

[8] A. Rajarathinam & B. Vinoth, "Outlier Detection in Simple Linear Regression Models and Robust Regression-A Case Study on Wheat Production Data", International Journal of Scientific Research **3** (2014) 531.

[9] S. L. Jegede, A. F. Lukman, K. Ayinde, & K. A. Odeniyi, "Jackknife Kibria-Lukman M-Estimator: Simulation and Application", Journal of the Nigerian Society of Physical Sciences **4** (2022) 251, doi: 10.46481/jnsps.2022.664.

[10] B. T. Tan, P. S. Fam, R. B. R. Firdaus, T. Mou Leong, & M. S. Gunaratne, "Impact of climate change on rice yield in malaysia: A panel data analysis", Agriculture (Switzerland) **11** (2021), doi: 10.3390/agriculture11060569.

[11] Y. Susanti, H. Pratiwi, H. Sulistijowati, & T. Liana, "M Estimation, s estimation, and mm estimation in robust regression", International Journal of Pure and Applied Mathematics **3** (2014) 349, doi: 10.12732/ijpam.v91i3.7.

[12] Y. Susanti & D. Pratiwi, "MODELING OF SOYBEAN PRODUCTION IN INDONESIA USING ROBUST REGRESSION", Bionatura **14** (2012) 148.

[13] P. J. Huber, "Robust Estimation of a Location Parameter", The Annals of Mathematical Statistics **35** (1964) 73.

[14] F. Drobnic, A. Kos, & M. Pustisek, "On the interpretability of machine learning models and experimental feature selection in case of multicollinear data", Electronics (Switzerland) **9** (2020), doi: 10.3390/electronics9050761.

[15] M. Z. I. Chowdhury & T. C. Turin, "Variable selection strategies and its importance in clinical prediction modelling", Fam Med Community Health **8** (2020), doi: 10.1136/fmch-2019-000262.

[16] H. Kaneko, "Examining variable selection methods for the predictive performance of regression models and the proportion of selected variables and selected random variables", Heliyon **7** (2021) 1, doi: 10.1016/j.heliyon.2021.e07356.

[17] Mukhtar, M. K. M. Ali, M. T. Ismail, M. H. Ferdinand, & Alimuddin, "Machine learning-based variable selection: An evaluation of Bagging and Boosting", Turkish Journal of Computer and Mathematics Education **12** (2021) 4343.

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

[19] S. Georganos, T. Grippa, A.N. Gadiaga, C. Linard, M. Lennert, S. Vanhuysse, N. Mboga, E. Wolff., & S. Kalogirou, "Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling", Geocarto Int **36** (2021) 121, doi: 10.1080/10106049.2019.1595177.

[20] D. O. Oyewola, E. G. Dada, N. J. Ngozi, A. U. Terang, & S. A. Akinwumi, "COVID-19 Risk Factors, Economic Factors, and Epidemiological Factors nexus on Economic Impact: Machine Learning and Structural Equation Modelling Approaches", Journal of the Nigerian Society of Physical

Sciences **3** (2021) 395, doi: 10.46481/jnsps.2021.173.

[21] V. Umarani, A. Julian, & J. Deepa, "Sentiment Analysis using various Machine Learning and Deep Learning Techniques", Journal of the Nigerian Society of Physical Sciences **3** (2021) 385, doi: 10.46481/jnsps.2021.308.

[22] R. Gandhi, "Support Vector Machine — Introduction to Machine Learning Algorithms", Towards Data Science, (2018).

[23] H. H. Rashidi, N. K. Tran, E. V. Betts, L. P. Howell, & R. Green, "Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods", Acad Pathol **6** (2019) 1, doi: 10.1177/2374289519873088.

[24] C. Cortes & V. Vapnik, "Support-Vector Networks", Mach Learn **20** (1995) 273.

[25] A. J. Smola, B. Scholkopf, & S. Scholkopf, "A tutorial on support vector regression", Kluwer Academic Publishers, (2004).

[26] N. Guenther & M. Schonlau, "Support vector machines", The Stata Journal **3** (2016) 917.

[27] Y. Freund, "Boosting a weak learning algorithm by majority", Inf Comput **121** (1995) 256.

[28] R. E. Schapire, "The Boosting Approach to Machine Learning an Overview", MSRI Workshop on Nonlinear Estimation and Classification, (2002).

[29] L. Breiman, "Bagging Predictors", Mach Learn **24** (1996) 123.

[30] Ö. G. Alma, "Comparison of Robust Regression Methods in Linear Regression", Int. J. Contemp. Math. Sciences **6** (2011) 409.

[31] A. E. Mohamed, H. M. Almongy, & A. H. Mohamed, "Comparison Between M-estimation, S-estimation, And MM Estimation Methods of Robust Estimation with Application and Simulation", International Journal of Mathematical Archive **9** (2018) 55.

[32] Mukhtar, M. K. M. Ali, A. Javaid, M. T. Ismail, & A. Fudholi, "Accurate and Hybrid Regularization - Robust Regression Model in Handling Multicollinearity and Outlier Using 8SC for Big Data", Mathematical Modelling of Engineering Problems **8** (2021) 547, doi: 10.18280/mmep.080407.

[33] R. C. Chen, C. Dewi, S. W. Huang, & R. E. Caraka, "Selecting critical features for data classification based on machine learning methods", J Big Data **17** (2020) 1, doi: 10.1186/s40537-020-00327-4.

[34] C. Njeru & A. Amayo, *Evaluation of Quality Control in Clinical Chemistry Using Sigma Metrics*, (2022).