Journal IJECE Hybrid Model Q2

by Alimuddin Alimuddin

Submission date: 19-May-2022 11:58PM (UTC+0700)
Submission ID: 1839961225
File name: Jurnal_IJECE_Hybrid_2022_Alimuddin.pdf (747.13K)
Word count: 8060
Character count: 43178

International Journal of Electrical and Computer Engineering (IJECE)
Vol. 12, No. 4, August 2022, pp. 4457~4468
ISSN: 2088-8708, DOI: 10.11591/ijece.v12i4.pp4457-4468

Hybrid model in machine learning–robust regression applied for sustainability agriculture and food security

Mukhtar[1,3,4], Majid Khan Majahar Ali[1], Mohd. Tahir Ismail[1], Ferdinand Murni Hamundu[2], Alimud[4,5], Naseem Akhtar[6], Ahmad Fudholi[7,8]

[1]School of Mathematical Sciences, Universiti Sains Malaysia, Gelugor, Pulau Pinang, Malaysia
[2]Department of Computer Science, Faculty of Mathematics and Natural Science, Universitas Halu Oleo, Kendari, Indonesia
[3]Department of Primary Education, Faculty of Teacher Training and Education, University Sultan Ageng Tirtayasa, Serang, Indonesia
[4]Center of Excellence for Local Food Innovation, Universitas Sultan Ageng Tirtayasa, Serang, Indonesia
[5]Departemen of Electrical Engineering, Faculty of Engineering, Universitas Sultan Ageng Tirtayasa, Serang, Indonesia
[6]School of Industrial Technology, Universiti Sains Malaysia, Gelugor, Pulau Pinang, Malaysia
[7]Solar Energy Research Institute, Universiti Kebangsaan Malaysia, Bangi Selangor, Malaysia
[8]Research Centre for Electrical Power and Mechatronics, National Research and Innovation Agency (BRIN), Bandung, Indonesia

Article Info

Article history:

Received Jun 22, 2021
Revised Feb 11, 2022
Accepted Mar 8, 2022

Keywords:

Food security
Machine learning
M-robust regression
Sustainability agriculture

ABSTRACT

A dataset containing 1924 observations used in this study to evaluate the effect of 435 different independent variables on one dependent variable. Big data has some issues such as irrelevant variables and outliers. Therefore, this study focused on analyzing and comparing the impact of three different variable selection based on machine learning techniques, including random forest (RF), support vector machines (SVM), and boosting. Further, the M robust regression was applied to address the outliers using M–bi square, M–Hampel, and M–Huber. Random forest and M-Hampel results revealed the significant comparing from the other methods such as mean absolute error (MAE) 175.33995, mean square error (MSE) 31.8608, mean average percentage error (MAPE) 9.16091, sum of square error (SSE) 89270.45, R–square 0.829511, and R–square adjusted 0.82670. Also, these techniques indicated that the 8 selection criteria were lower than the other techniques including Akaike information criterion (AIC) 47.25915, generalized cross validation (GCV) 47.27169, Hannan-Quinn (HQ) 47.60351, RICE (47.2845), SCHWARZ 51.7099, sigma square (SGMASQ) 46.50605, SHIBATA 47.23489, and final prediction error (FPE) 47.25929. Therefore, the study recommended that the best random forest and M-Hampel models are helpful to show the minimum issues and efficient validation for analyzing and comparing big data.

Corresponding Author:

Majid Khan Bin Majahar Ali
School of Mathematical Sciences, Universiti Sains Malaysia
USM-11800, Penang, Malaysia
Email: majidkhanmajaharali@usm.my

1. INTRODUCTION

Agriculture is an aspect of food security, and its problems have been severe in several regions of the world. Agriculture contributes to reducing poverty alleviation through lowering food rates, creating employment, enhancing farm incomes and rising wages. Food security considers the capacity of the human to produce sufficient food every day, provide nutrition, and minimize environmental impact [1]. Despite the demand for food, productions have been increased because the population raised throughout the world [2]. More food is needed in the context of population increase to fulfil the demands of developing countries [3].

Dealing with agricultural problems requires effective techniques to address the challenges of sustainable agriculture and food security.

Sustainability agriculture and food security h[31] used big data technologies expected to become more widespread in the future and evaluate the issues. Big data technologies are used in agriculture for the accuracy and enhancement of sustainable agriculture and food security [4]. Big data has a combination of many observations and a more significant number of variables [5], [6].

However, using too many variables in regression models becomes a problem, especially if there are irrelevant variables. Irrelevant variables can lead to noise and negatively influence the regression model [7]. Irrelevant variables have implications for model constrain that have higher variance and bias. In addition, the existing model can significantly affect the statistical analysis, which can lead to overfitting. For addressing these issues, we will propose machine learning as variable selection. Further, machine learning and big data are performed to solve the problems of agriculture activity [8].

Machine learning aims to solve the issue of irrelevant variables. Machine learning provides the rank of significant variables. The highest essential variables are the ranking of the independent variables that contribute to the dependent variable [9]. Variable selection is a significant challenge faced in today's world in big data, waiting to be explored. Variable selection results in variable importance [10].

Variable important represents the machine learning relevance (significance) of each variable in the data concerning its effect on the generated model. Variable importance is the process of selecting a suitable variable subset from the original variables. Variable important has been proven capable of improving regression accuracy and reducing the complexity of the learning model [11]. Machine learning has two groups such as supervised and unsupervised learning. Supervised learning is also divided into two sub-groups such as classification and regression. The dependent variable in the classification is discrete and continuous in regression.

The regression analysis has been used in this study for addressing the relationship between independent and dependent variables. Regressions in big data continue to get significant appreciation and attention. However, reversals in big data still have open problems, such as irrelevant variables [12] and outliers [13]. Numerous machine learning has been suggested to handle regression in big data problem[39] such as unsupervised, reinforcement, supervised, and semi-supervis[24] learning in different areas [14]. In this study, we will use machine learning-based variable selections such as random forest (RF), support vector machines (SVM), [40] boosting. This study used supervised machine learning techniques based on variable selections such as random forest, support vector machines, and boosting. Therefore, the subset of the highest 30 influential variables has been taken from each technique. [3]

The second issue in regression-big data is an outlier. An outlier is a data point that is significantly different from surrounding points [15]. The outliers can occur due to various reasons such as human error, mechanical error, and instrument error [16]. The existence of outliers in a regression model is inaccurate and incorrectly identified model. Furthermore, an incorrect estimate model can lead to rising erroneous results and significantly influence the mean and standard deviation and lead to either over-or underestimated values [17].

Ordinary least[2] quare regression can be very sensitive to outliers, and outliers can strongly distort and unreliable results. Several robust-to outliers have been proposed in the statistical literature [18], [19]. M-robust regressions can handle outliers [20]. M-robust regressions include M-bi square Tukey, M-Hampel, and M-Huber.

The objectives of this study were to address the problems based on both irrelevant variables and outli[22] Also, this study used a hybrid model such as machine learning and M-robust regression techniques. The main objectives of this study were to examine the impact on the variable selection of three different techniques of machines learning, including random forest, support vector machine, and boosting. In addition, this study was addressed outliers based on M-robust regression techniques such as M-bi square, M-Hampel, and M-Huber.

2. METHODS

2.1. Regression learning

Regression is a challenging issue in various field of knowledge. This study investigates performances regression in big data. Regression is a method used to build the predictive model [20]. Regression analysis is a supervised machine learning technique for building a model and evaluat[28] its performance for a continuous response based on the relationship among several variables. Regression is one of the main tasks in machine learning and has been successfully applied to many areas[32] ch as sustainable agriculture and food security [21]. The multiple regressions construct and assess the model based on the relationship between independent and dependent variables [22]. The main purpose of these methods is

training the relationship between independent variables, $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, a dependent variable y_i, for n observations, p the number of variables, $(x_i, y_i)_{i=1}^{n}$. i). Determine the casual relationship between independent variables $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and a dependent variable $y_i = (y_1, y_2, \ldots, y_n)_{i=1}^{n}$; ii) predictive y_i based on a set of variables $x_{i1}, x_{i2}, \ldots, x_{ip}$; and iii) screening $x_{i1}, x_{i2}, \ldots, x_{ip}$ to select variables that have a most significant effect than others to describe the dependent variable Y_i. Regression learning tasks can be stated as learning a function $\varphi: x \rightarrow y$ from a learning set $\mathcal{L} = (x, y)$. The purpose of regression learning is to find a model in such that its prediction $\varphi(x)$ which denoted by $\hat{Y}$ that as good as possible and Y_i is continuous [23].

2.2. Random forest

Random forest is used for classification and regression. The prediction in classification is based on the majority votes of the predicted values, and in regression is based on average [24]. Random forest is a tree-based ensemble learning. Random forest takes bootstrapped sample each of the ensemble. Random forest takes several subsets of nominee variables at every node when trees are built. The nominee variables are random subset of m independent variables from p independent variables. The m is a parameter which is controlled by user. In the original article, m equal $\log_2(p+1)$ [25]. Later, researchers applied default as $m \approx \sqrt{p}$ for classification and $m \approx \frac{p}{3}$ for regression [26].

Algorithm [26] random forest refers to Breiman algorithm.
Given, D: dataset with n observations, p independent variables, and one dependent variable.
Procedure:
For $b = 1$ to n
- Generate bootstrapped sample D_b^* from the training set D.
- Grow a tree using a m from bootstrapped sample D_b^*.
  For a given 3 mode i) Randomly choose m variables, ii) Find the best split variables and values, and iii) Split a node using the best split variables and values.

Repeat i) – iii) until stopping rules are met.

Radom forest has advantages: i) low complexity is $o(n\log(n))$, ii) robustness to handle is in unbalanced dataset, and iii) embedded variable selection is to rank variables by important [24]. The issue is important variables which a bias toward correlated independent variables [27].

2.3. Boosting

Boosting aims to improve model's accuracy. It has an idea to find and the average of thumb than to see single learning [28]. Boosting is applied to training data in a step-by-step manner, apply proper methods regularly to place more emphasis on observations [29].

The data $\{x_i, y_i\}_{i=1}^{N}$ of known (x, y) – values. Boosting aims to get an approximation $\hat{F}(x)$. The function $F^*(x)$ aims mapping x to y which minimizes the fitted values for loss function $L(y, F(x))$ include squared error $(y-F)^2$ for $y \in R$ [30].

Algorithm boosting:
Given: $(x_i, y_i), \ldots, (x_n, y_n)$
where $x_i \in X, y_i \in Y$
Initialize $D_1(i) = \frac{1}{m}$
- Train base learner using distribution D_t
- Get base regression $f_t: X \rightarrow \mathbb{R}$
- Choose $\alpha_t \in \mathbb{R}$
- Update

$$D_{t+1}(i) = \frac{D_t(i)exp(-\alpha_t y_i f_t(x_i))}{m}$$

The Z_t is a normalization factor (chosen so that D_{t+1} will be a distribution) output the final regression.

$$F^* = arg\min_F E_{yx} L(y, F(x)) = arg\min_F E_x \left[E_x\left(L(y, F(x))\right)|x\right]$$

2.4. Support vector machines-regression

Support vector machines (SVM) is developed by Vapnick that it is a learning system based on structural risk minimization (SRM) [31]. The traditional empirical risk minimization (ERM) principle minimizes the errors in training data, while SRM minimizes ERM an upper bound on the expected risk. The

SVM more accurate in generalization. The SVM was previously applied to overcome classification issue [32]. However, the SVM can also be applied to overcome regression problems by a loss function.

The ε loss function is frequently applied for regression purposes [33]. If the ε is smaller than predictive error that will be ignored [32]. In most cases, the ε is a small positive number or zero such as 0.001. Support vector machines has the ε – incentive loss function $\|y - f(x)\|_\varepsilon = max\{0, \|y - f(x)\| - \varepsilon\}$. The $\varepsilon > 0$ and creating a tube around the true output.
The primal becomes:

$$t(w,\xi) = \frac{1}{2}\|\omega\|^2 + \frac{C}{m}\sum_{i=1}^{m}(\xi_i + \xi_i^*) \quad (1)$$

$$\text{Subject to } (\langle\phi(x_i), w\rangle + b) - y_i \leq \varepsilon - \xi_i \quad (2)$$
$$y_i - (\langle\phi(x_i), w\rangle + b) \leq \varepsilon - \xi_i^*$$
$$\xi_i^* \geq 0 \qquad (i = 1, .., m)$$

The formulation can estimate the accuracy of support vector machines – regression by computing the scale parameter of a Laplacian distribution on the residuals. The $f(x)$ is the estimation decision function [34], [35].

2.5. M robust-regression

In regression analysis, outliers in a dataset can cause the least squares estimator to distort and produce unreliable results. To deal with this issue, a number of robust to outliers' methods have been proposed in the statistical literature. Numerous robust to outlier's methods have been proposed in the statistical literature to address this issue. The traditional methods usually decrease the efficiency in estimating the population parameters as these methods are sensitive to outliers [36]

Therefore, in the present study we adapt the various robust regression techniques such as M-estimation such as M Tukey bi square, M Hampel, and M Huber. The M in M-estimation is "Maximum likelihood". We consider only the linear model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad (3)$$

$$y_i = x_i^T \beta + \varepsilon_i \quad (4)$$

For the i-*th* of n independent observations. We assume that the model itself is not at issue, therefore $(y|x) = x_i'\beta$, the distribution of the errors may be heavy tailed, producing occasional outliers [37]. Given an estimator b for β, the fitted model is

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip} = x_i^T b \quad (5)$$

$$\varepsilon_i(\beta) = y_i - x_i^T \beta \quad (6)$$

With M-estimation including M–bi square, M–Hampel, and M-Huber; the estimate b determined by minimizing a particular objective function over all b.
The ρ gives the contribution of each residual to objective function [38]. The ρ requires the properties:
- $\rho(\varepsilon) \geq 0$
- Equal to zero when its argument is zero, $\rho(\varepsilon) = 0$
- Symmetric, $\rho(\varepsilon) = \rho(-\varepsilon)$
- Monotone in $|\varepsilon_i|$, $\rho(\varepsilon_i) = \rho(\varepsilon_{i'})$ for $|\varepsilon_i| > |\varepsilon_{i'}|$

M-estimation principle is to minimize the residual function ρ:

$$\hat{\beta}_M = \min_\beta \rho\left(y_i - \sum_{j=0}^{k} x_{ij}\beta_j\right) \quad (7)$$

The (7) has to solve:

$$\min_\beta \sum_{i=1}^{n} \rho\left(\frac{\varepsilon_i}{\sigma}\right) = \min_\beta \rho\left(\frac{y_i - \sum_{j=0}^{k} x_{ij}\beta_j}{\sigma}\right) \quad (8)$$

The function σ is in (9)

$$\hat{\sigma} = \frac{MAD}{0.6745} = \frac{median|\varepsilon_i - median(\varepsilon_i)|}{0.6745} \quad (9)$$

The M-estimator of scale $\hat{\sigma}$ is found by solution of (10):

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{\varepsilon_i}{\hat{\sigma}}\right) = \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{y_i - \beta^T x_i}{\hat{\sigma}}\right) = k \quad (10)$$

The β is the $p \times 1$ parameter vector, then the function ψ – type could be yielding as (11):

$$\sum_i \psi(\varepsilon_i)\frac{\partial \varepsilon_i}{\partial \beta_i}, \text{for } j = 1,2,\dots,p \quad (11)$$

The derivative $\psi(e) = \frac{\partial \rho(\varepsilon_i)}{\partial(\varepsilon_i)}$ is the influence function. Then the weight function colud be defined as (12)

$$w(e) = \frac{\psi(\varepsilon)}{\varepsilon} \quad (12)$$

The $\psi(e)$ – type function becomes:

$$\sum_i w(\varepsilon_i)\varepsilon_i \frac{\partial \varepsilon_i}{\partial \beta_i} = 0, \text{for } j = 1,2,\dots,p \quad (13)$$

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

$$min \sum_i w\left(\varepsilon_i^{(k-1)}\right) e_i^2 \quad (14)$$

The k is the number of iterations. Further, the M robust regressions have been used to eliminate the outliers using M-bi square, M-Hampel, and M-Huber [39], as well as more detail showed in Table 1.

Table 1. Formulas for robust regression M-estimation

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

2.6. Selection models

2.6.1. Phase 1 – all possible models

For this study, the dataset is interacted only second order. Where N is the number of all possible models, k is total number of independent variables and j=1, 2,…, k.

$$N = \sum_{j=1}^{k} j\left(C_j^k\right) \quad (15)$$

A dataset containing 1924 observations will use to study the effect of 29 different independent variables on the on the one dependent variable. Then the data will be interacted with in the second interaction. The data contain the effect of 435 different interaction independent variables on the dependent variable.

2.6.2. Phase 2 – selected models

In this paper, we will analyze the machine learning as variable selection including random forest, support vector machine, and boosting. We will take subset of top 30 highest influential variables from each technique and will apply three M robust regression including M-bi square, M-Hampel, and M-Huber.

2.6.3. Phase 3 – the best model

The next step was to get the best model after a list of selected models was obtained. Eight selection criteria (8SC) were defined for these purposes by [40]. The 8SC formula can be displayed as shown in Table 2. By using mentioned formulas in Table 2, Akaike information criterion (AIC), RICE, final prediction error (FPE), SCHWARZ, generalized cross validation (GCV), sigma square (SGMASQ), SHIBATA, and Hannan-Quinn (HQ) information on the basis of the minimum value obtained from all mentioned criteria.

Table 2. Formula used for 8SC

No	Methods	Formulation	Reference
1.	AIC	$\left(\frac{sse}{n}\right)e^{\frac{2(k+1)}{n}}$	[41]
2.	RICE	$\frac{\left(\frac{sse}{n}\right)}{\left[1-\left(\frac{2(k+1)}{n}\right)\right]}$	[42]
3.	Final prediction error (FPE)	$\left(\frac{(SSE)^2}{n}\right)\frac{n+(k+1)}{n-(k+1)}$	[41]
4.	Schwarz	$\left(\frac{sse}{n}\right)n^{\left(\frac{k+1}{n}\right)}$	[43]
5.	GCV	$\frac{\left(\frac{sse}{n}\right)}{\left[1-\left(\frac{k+1}{n}\right)\right]^2}$	[44]
6.	SGMASQ	$\frac{\left(\frac{sse}{n}\right)}{\left[1-\left(\frac{k+1}{n}\right)\right]}$	[44]
7.	SHIBATA	$\left(\frac{sse}{n}\right)\left(\frac{n+2(k+1)}{n}\right)$	[45]
8.	HQ	$\left(\frac{sse}{n}\right)\ln n^{\frac{2(k+1)}{n}}$	[46]

2.7. Validation models

The validation of model including mean average error (MAE), mean average percentage error (MAPE), sum of square error (SSE), R–Square, and R–Square Adjusted are measured for evaluating the model performances. SSE measure the discrepancy the data and an estimation model. Generally, the lower SSE shows which model can better explain, and the higher SSE shows which model poorly describes the data [47]. Besides that, the value of R–square is common method to explain the goodness of fit in regression model. R-square interprets how many percentages of the variation is explained by the independent variables. R-square=1 interprets that the model is in fitting with real data [48]. Mean square error (MSE) measures the average of the squared deviation between the fitted values with the actual data observation [49]. The validation of model determinations as shown in Table 3.

Table 3. Formulas for validation methods

No	Validation	Formulation	Reference
1	Sum of square error	$SSE=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$	[47]
2	Sum of square total	$SST=\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2$	[47]
3	R-squared	$R^2=\frac{SSR}{SST}=\frac{SST-SSE}{SST}=1-\frac{SSE}{SST}$	[48]
4	Mean absolute error (MAE)	$MAE=\frac{1}{n}\sum_{i=1}^{n}\left\|\frac{Y-\hat{Y}_i}{\hat{Y}_i}\right\|$	[50]
5	Mean square error (MSE)	$MSE=\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y-\hat{Y}_i}{\hat{Y}_i}\right)^2$	[50]
6	Mean average percentage error (MAPE)	$MAPE=\frac{100}{n}\sum_{i=1}^{n}\left\|\frac{Y-\hat{Y}_i}{\hat{Y}_i}\right\|$	[51]

3. RESULTS AND DISCUSSION

3.1. Data

The data was collected from time period of 8:00 am until 5:00 pm starting on 08/04/2017 to 12/04/2017. That is almost four days data. The original data was for each second and then it was converted in hour for data analysis. The variables taken are data contain hourly solar radiation, temperature, humidity, and moisture content. The detailed factor of modelling is shown in Table 4.

In this paper, a dataset containing 1924 observations will use to study the effect of more 29 different independent variables on the one dependent variable. Significance of interaction terms had also been observed in this study. So, T1*T2 represents the interaction between T1 and T2. Another example H1*PY represents the interaction between H1 and PY. The data contain the effect of 435 different interaction independent variables on the one dependent variable.

Table 4. Factors of modelling

Symbols	Factors	Definitions
Y	Dependent	Moisture
H1	Independent	Relative Humidity Ambient
H5	Independent	Relative Humidity Chamber
PY	Independent	Solar Radiation
T1	Independent	Temperature (°C) ambient
T2, T3, T4	Independent	Temperature (°C) before enter solar collector
T5	Independent	Temperature (°C) in front of down v-Groove (Solar Collector)
T6, T8	Independent	Temperature (°C) in front of up v-Groove (Solar Collector)
T7, T14, T15, T16, T21, T22	Independent	Temperature (°C) Solar Collector
T9, T10, T11, T12	Independent	Temperature (°C) behind inside chamber
T13, T17, T18, T19, T23	Independent	Temperature (°C) Infront of (Inside Chamber)
T20, T23, T24, T25, T28	Independent	Temperature (°C) from solar collector to chamber

3.2. Result

The validation metrics are sum square of error (SSE), mean absolute error (MAE), mean square error (RMSE), mean absolute percentage error (MAPE), and R-Square. They are comparing three machines learning-based variable selection and M-Robust regression algorithms-in terms of the best model eight selection criteria (8SC). In this study, we have described the three variable selection techniques which have been used such as random forest, support vector machines, and boosting.

Table 5 shows the final result obtained through each method to obtain the important by ranks. It shows the subset of 30 variable important that is taken by each technique. In order to measure the prediction accuracy, predicted responses with the actual responses are compared of each regression-based model in terms of the validation methods described in Table 6.

Table 5. The 30 highest of variable importance

	Methods	Variable Importance
1	Random Forest	T8, H1*PY, T8*H5, T8*H1, T2*T6, T2*T7, T1*T6, T21*H5, T7*H1, H5*PY, T9, T19*H5, T7, H1*H5, T7*PY, T22*H5, T6*T8, T10*H5, T11*H5, T6*T13, T12*T13, T9*H5, H1, T7*T9, T6*T9, T8*T23, T23*H1, T3*T8, T4*T8
2	Support Vector Machines	T1*T6, T2*T6, T17*H1, T6*T13, T19*H1, T22*H1, T17*PY, T1*T2, T27*PY, T28*PY, T21*H1, T27*H1, T9*PY, T22*PY, T10*PY, T5*PY, T19*PY, T21*PY, T1*PY, T6*T29, T2*PY, T2*T13, T29*PY, T13*PY, T23*PY, T28*H1, T12*PY, T14*PY, T26*H1, T11*PY
3	Boosting	T2*T6, T1*T6, H5*PY, T7*H1, T5*PY, T21*H5, T8*PY, T7*T9, T8, T2*T7, T6*T13, T19*H5, T26*H1, T7*PY, T10*PY, T2*T9, T8*T29, T11*H5, T17*H5, T6*T9, T9, T1*T2, T1*T9, T12*PY, T7, T26*H5, T12*H5, T9*T29, T25*H5, and T1*T8

Table 6 show the validation model metric that the random forest-Hampel exhibited the lowest error data. It could be assumed that random forest-Hampel's is method can rely on the investigation of the accuracy in big data obtained from machine learning-robust regression. Random forest-Hampel obtained significantly better results than others.

Table 6. Results for the validation model

Machine Learning	Robust Regression	MAE	MSE	MAPE	Sum Square of Error	R-square
Random Forest	Bi-Square	235.16695	183.165	12.28667	238913.7	0.543723
	Hampel	175.33995	31.8608	9.160917	87570.9	0.838757
	Huber	221.3641	42.8569	11.56552	89270.45	0.829511
Support Vector Machines	Bi-Square	209.086525	63.4550	10.92406	191406.1	0.634453
	Hampel	249.01216	57.1446	13.01004	134216.8	0.743673
	Huber	237.3297451	52.8000	12.39967	136270.1	0.739752
Boosting	Bi-Square	281.774977	1837.10	14.72179	121532.1	0.767898
	Hampel	184.06188	50.2921	9.616608	86894.18	0.834049
	Huber	187.4855378	64.3844	9.795483	88406.59	0.831161

All possible models have 9 models were machine learning including random forest, support vector machines, and boosting and M robust regression including Tukey-bi square, Hampel, and Huber. The results obtained from 8 selection criteria are observed in Table 7. The minimum value for 8 selection criteria was found for model random forest-Hampel. The minimum value of 8 selection criteria for random forest-Hampel represented the efficient model obtained in phase 3.

Table 7. Results for 8 selection criteria for machines learning–M robust regression

ML	Robust Regression	AIC	GCV	HQ	RICE	SCHWARZ	SGMASQ	SHIBATA	FPE
Random Forest	Bi-Square	128.9339	128.9681	129.8734	129.0031	141.0766	126.8793	128.8677	128.9343
	Hampel	46.90639	47.23564	46.9191	51.3103	46.14667	46.86987	46.89408	46.89395
	Huber	48.17634	48.18912	48.52738	48.20218	52.71347	47.40863	48.15161	48.17648
Support Vector Machines	Bi-Square	103.2956	103.323	104.0483	103.351	113.0237	101.6495	103.2426	103.2959
	Hampel	72.43242	72.45163	72.9602	72.47127	79.25392	71.27817	72.39523	72.43263
	Huber	73.5405	73.56001	74.07638	73.57995	80.46636	72.3686	73.50274	73.54071
Boosting	Bi-Square	49.73711	49.7503	66.06481	49.76379	54.42122	48.94452	49.71157	49.73725
	Hampel	47.25915	47.27169	47.60351	47.2845	51.7099	46.50605	47.23489	47.25929
	Huber	47.71015	47.7228	48.05779	47.73574	52.20336	46.94986	47.68565	47.71028

The random forest-Hampel has the lowest validation of model including MAE (175.33995), MSE (31.8608), MAPE (9.160917), SSE (87570.9), and R-square (0.838757). The random forest-Hampel has the lowest 8 selection criteria including AIC (46.90639), GCV (47.23564), HQ (46.9191), RICE (51.3103), SCHWARZ (46.14667), SGMASQ (46.86987), SHIBATA (46.89408), and FPE (46.89395). In short, we can conclude that random forest-Hampel has generated the lowest error data, which provides the most relevant data in the context of validation model and 8 selection criteria.

The MAE, MSE, MAPE, SSE and R-square are useful measure widely used in validation model. The lowest of validation model is random forest-Hampel than others. The MAE, MSE, and SSE are used in explaining how well the regression model is toward to the model data. In particular, the explained MAE, MSE, and SSE measure the variation for the error between the predicted and actual data. Hence, the random forest-Hampel has the lowest bias and variation. The highest of R-Square (0.829511) is the random forest-Hampel. The R-squares measures variation which was accounted for the predicted data. We suggest that the dependent variable 82.9511% by the independent variables.

Variable selection can build a useful regression model. Variable selection can increase accuracy and reduce model complexity. Variable selection consists of selecting the variables that have the most significant influence on dataset of regression [52]. Variable selection has drawn our attention that the important variable techniques were developed independently in many disciplines [53]. Variable selection has resulted in a subset of important variable. Variable important is rank the variables according to an important variable measure. Variable selection attracts researchers who deal with machine learning.

Random forest has aim to reduce dimensionality [54]. Random forest is easy and fast to implement, provides very accurate predictions and can manage many variables without overfitting [55]. Random forest is well suited for medium to big data. Rand13 forest has good predictive performance in practice. Moreover, random forest provides some measured of the importance of the variables with respect to the prediction of the outcome variable. Random forest is interesting in the machine learning research which concerns ensemble learning which generates regression model. Random forest is broadly accepted in which the performance of set many variables selection is usually more beneficial than others [56]–[59].

M-robust regressions are an analysis that is applied if there are outliers in the dataset [19]. In this study, the result obtained that M-Hampel robust regression gives the lowest in MAE, MSE, MAPE, sum squares of error and gives highest in R square and R square adjusted. M-Hampel robust regression is outperform than others [60].

The random forest-Hampel gives lowest in MAE, MSE, MAPE, SSE, and gives highest in R square. Sustainability agriculture and food security are regarded as two of the most important economical parts of Malaysia. Random forest presents two characteristics, such as high prediction accuracy and information associated with variable importance [61]. Sustainability agriculture and food security are two sectors that are benefiting strongly the development of both machine learning and M-robu7 regression in the latest years.

Machine learning and M-robust regression have emerged with big data technologies and high-performances computing to create new opportunity to unravel, quantity, and understan30ata intensive processes in agriculture operational environments [62], [63]. Machine learning and statistics learning applies in more and more scientific fields such as sustainability agriculture and food security [64].

Machine learning and statistics learning are two core techniques for building precision agriculture systems. Recently, modelling in mathematics has been proposed to promote the modernization of agriculture for increasing both sustainability agriculture and food security greatly. Machine learning and statistics learning are used to analyze the agriculture data for smart decision-making [65]. The sustainability agriculture and food security are more reliable, capable, and help to boost productivity [66]. Random forest has shown a reliable and accurate model to predict paddy showing a very high accuracy, which is aimed for sustainability agricultural and food security [67]. The random forest-Hampel's provides the most relevant data of the result which applied for sustainability agriculture and food security.

4. CONCLUSION

The results show that the random forest-Hampel model provides the best model compared to other existing methods used in the analysis. The proposed hybrid model is found to be better in terms of MAE, MSE, MAPE, SSE, and R-square values than other existing methods. The random forest-Hampel provides the best 8 selection criteria including AIC, GCV, HQ, RICE, SCHWARZ, SGMASQ, SHIBATA, and FPE. The random forest-Hampel's provides the best model which should be applied for Sustainability Agriculture and Food Security.

ACKNOWLEDGEMENTS

The author would like to express his gratitude to the University of Sultan Ageng Tirtayasa Banten Indonesia and University Sains Malaysia for their assistance with this study

REFERENCES

[1] F. Eyhorn *et al.*, "Sustainability in global agriculture driven by organic farming," *Nature Sustainability*, vol. 2, no. 4, pp. 253–255, Apr. 2019, doi: 10.1038/s41893-019-0266-6.
[2] N. Akhtar *et al.*, "Multivariate investigation of heavy metals in the groundwater for irrigation and drinking in garautha tehsil, jhansi district, India," *Analytical Letters*, vol. 53, no. 5, pp. 774–794, Mar. 2020, doi: 10.1080/00032719.2019.1676766.
[3] R. Taghizadeh-Mehrjardi, K. Nabiollahi, L. Rasoli, R. Kerry, and T. Scholten, "Land suitability assessment and agricultural production sustainability using machine learning models," *Agronomy*, vol. 10, no. 4, pp. 573–593, Apr. 2020, doi: 10.3390/agronomy10040573.
[4] N. N. Misra, Y. Dixit, A. Al-Mallahi, M. S. Bhullar, R. Upadhyay, and A. Martynenko, "IoT, big data and artificial intelligence in agriculture and food industry," *IEEE Internet of Things Journal*, vol. 4662, pp. 1–1, 2020, doi: 10.1109/JIOT.2020.2998584.
[5] S. T. McAbee, R. S. Landis, and M. I. Burke, "Inductive reasoning: The promise of big data," *Human Resource Management Review*, vol. 27, no. 2, pp. 277–290, Jun. 2017, doi: 10.1016/j.hrmr.2016.08.005.
[6] S. R. Salkuti, "A survey of big data and machine learning," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 1, pp. 575–580, Feb. 2020, doi: 10.11591/ijece.v10i1.pp575-580.
[7] D. Peralta, S. del Río, S. Ramírez-Gallego, I. Triguero, J. M. Benitez, and F. Herrera, "Evolutionary feature selection for big data classification: A MapReduce approach," *Mathematical Problems in Engineering*, vol. 2015, pp. 1–11, 2015, doi: 10.1155/2015/246139.
[8] A. Sharma, "A comparison of machine learning algorithms using an insufficient number of labeled observations," pp. 1-7, 2012.
[9] A. Cai, R. S. Tsay, and R. Chen, "Variable selection in linear regression with many predictors," *Journal of Computational and Graphical Statistics*, vol. 18, no. 3, pp. 573–591, Jan. 2009, doi: 10.1198/jcgs.2009.06164.
[10] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "Recent advances and emerging challenges of feature selection in the context of big data," *Knowledge-Based Systems*, vol. 86, pp. 33–45, Sep. 2015, doi: 10.1016/j.knosys.2015.05.014.
[11] B. Tran, B. Xue, and M. Zhang, "A new representation in PSO for discretization-based feature selection," *IEEE Transactions on Cybernetics*, vol. 48, no. 6, pp. 1733–1746, Jun. 2018, doi: 10.1109/TCYB.2017.2714145.
[12] J. A. Doornik and D. F. Hendry, "Statistical model selection with 'big data,'" *Cogent Economics & Finance*, vol. 3, no. 1, Dec. 2015, doi: 10.1080/23322039.2015.1045216.
[13] J. Zhu, Z. Ge, Z. Song, and F. Gao, "Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data," *Annual Reviews in Control*, vol. 46, pp. 107–133, 2018, doi: 10.1016/j.arcontrol.2018.09.003.
[14] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," *Neurocomputing*, vol. 237, pp. 350–361, May 2017, doi: 10.1016/j.neucom.2017.01.026.
[15] A. Ayadi, O. Ghorbel, A. M. Obeid, and M. Abid, "Outlier detection approaches for wireless sensor networks: A survey," *Computer Networks*, vol. 129, pp. 319–333, Dec. 2017, doi: 10.1016/j.comnet.2017.10.007.
[16] S. Abghari, V. Boeva, N. Lavesson, H. Grahn, S. Ickin, and J. Gustafsson, "A minimum spanning tree clustering approach for outlier detection in event sequences," in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2018, pp. 1123–1130, doi: 10.1109/ICMLA.2018.00182.
[17] H. Perez and J. H. M. Tah, "Improving the accuracy of convolutional neural networks by identifying and removing outlier images in datasets using t-SNE," *Mathematics*, vol. 8, no. 5, Apr. 2020, doi: 10.3390/math8050662.
[18] G. Bogale Begashaw and Y. Berihun Yohannes, "Review of outlier detection and identifying using robust regression model," *International Journal of Systems Science and Applied Mathematics*, vol. 5, no. 1, pp. 4–11, 2020, doi: 10.11648/j.ijssam.20200501.12.
[19] V. Verardi and C. Croux, "Robust regression in Stata," *Stata Journal*, vol. 9, no. 3, pp. 439–453, 2009, doi: 10.1177/1536867x0900900306.
[20] W. H. Nugroho, N. W. S. Wardhani, A. A. R. Fernandes, and S. Solimun, "Robust regression analysis study for data with outliers at some significance levels," *Mathematics and Statistics*, vol. 8, no. 4, pp. 373–381, Jul. 2020, doi: 10.13189/ms.2020.080401.
[21] A. Zlotnik and V. Abraira, "A general-purpose nomogram generator for predictive logistic regression models," *The Stata Journal: Promoting communications on statistics and Stata*, vol. 15, no. 2, pp. 537–546, Jun. 2015, doi: 10.1177/1536867X1501500212.
[22] M. T. Shakoor, K. Rahman, S. N. Rayta, and A. Chakrabarty, "Agricultural production output prediction using Supervised Machine Learning techniques," in *2017 1st International Conference on Next Generation Computing Applications (NextComp)*, Jul. 2017, pp. 182–187, doi: 10.1109/NEXTCOMP.2017.8016196.
[23] C. Xiao, "Using machine learning for exploratory data analysis and predictive models on large datasets," *Master's Thesis*, pp. 1–78, 2015.
[24] M. Shahhosseini, G. Hu, and H. Pham, "Optimizing ensemble weights and hyperparameters of machine learning models for regression problems," *Machine Learning with Applications*, vol. 7, Mar. 2022, doi: 10.1016/j.mlwa.2022.100251.
[25] P. A. A. Resende and A. C. Drummond, "A survey of random forest based methods for intrusion detection systems," *ACM Computing Surveys*, vol. 51, no. 3, pp. 1–36, May 2019, doi: 10.1145/3178582.
[26] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–23, 2001, doi: 10.1023/A:1010950718922.

[27] S. Han and H. Kim, "On the optimal size of candidate feature set in random forest," *Applied Sciences*, vol. 9, no. 5, Mar. 2019, doi: 10.3390/app9050898.
[28] Y. Qi, *Ensemble machine learning*. Boston, MA: Springer US, 2012.
[29] R. E. Schapire, "The boosting approach to machine learning: An overview," in *Lecture Notes in Statistics*, 2003, pp. 149–171.
[30] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," *Journal of Animal Ecology*, vol. 77, no. 4, pp. 802–813, Jul. 2008, doi: 10.1111/j.1365-2656.2008.01390.x.
[31] K. K. Htike, "Efficient determination of the number of weak learners in AdaBoost," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 29, no. 5, pp. 967–982, 2017, doi: 10.1080/0952813X.2016.1266038.
[32] V. N. Vapnik, *Statistics for Engineering and Information Science Springer Science+Business Media, LLC*. 2000.
[33] S. Huang, "Predictive modeling and analysis of student academic performance in an engineering dynamics course," Dissertation, Utah State University, 2011.
[34] J. Weston, A. Gammerman, and M. Stitson, "Density estimation using support vector machiens," *Computer Science*, pp. 1–11, 1998.
[35] A. Karatzoglou, D. Meyer, and K. Hornik, "Support vector algorithm in R," *Journal of Statistical Software*, vol. 15, no. 9, pp. 1–28, 2006.
[36] R. Muthukrishnan, R. Reka, and E. D. Boobalan, "Robust regression procedure for model fitting with application to image analysis," *International Journal of Statistics and Systems*, vol. 12, no. 1, pp. 79–92, 2017.
[37] M. Subzar, C. N. Bouza, and A. I. Al-Omari, "Utilization of different robust regression techniques for estimation of finite population mean in SRSwor in case of presence of outliers through ratio method of estimation," *Investigacion Operacional*, vol. 40, no. 5, pp. 600–609, 2019.
[38] M. Riani, A. Cerioli, A. C. Atkinson, and D. Perrotta, "Monitoring robust regression," *Electronic Journal of Statistics*, vol. 8, no. 1, pp. 646–677, Jan. 2014, doi: 10.1214/14-EJS897.
[39] E. Mohamed Almetwally and H. Mohamed Almongy, "Comparison between M-estimation, S-estimation, and Mm estimation methods of robust estimation with application and simulation," *International Journal of Mathematical Archive*, vol. 9, no. 11, pp. 55–63, 2018.
[40] H. J. Zainodin and G. Khuneswari, "Model-building approach in multiple binary Logit model for coronary heart disease," *Malaysian Journal of Mathematical Sciences*, vol. 4, no. 1, pp. 107–133, 2010.
[41] H. Akaike, "Fitting autoregressive models for prediction," *Annals of the Institute of Statistical Mathematics*, vol. 21, no. 1, pp. 243–247, Dec. 1969, doi: 10.1007/BF02532251.
[42] J. Rice, "Bandwidth choice for nonparametric regression," *Annals of Statistics*, vol. 12, no. 4, pp. 243–247, 1969, doi: 10.1007/BF02532251.
[43] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, Mar. 1978, doi: 10.1214/aos/1176344136.
[44] G. H. Golub, M. Heath, and G. Wahba, "Generalized a method cross-validation for choosing parameter," *Technometrics*, vol. 21, no. 2, pp. 215–223, 1979.
[45] R. Shibata, "An optimal selection of regression variables," *Biometrika*, vol. 68, no. 1, pp. 45–54, 1981, doi: 10.1093/biomet/68.1.45.
[46] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 41, no. 2, pp. 190–195, Jan. 1979, doi: 10.1111/j.2517-6161.1979.tb01072.x.
[47] S. Kim and H. Kim, "A new metric of absolute percentage error for intermittent demand forecasts," *International Journal of Forecasting*, vol. 32, no. 3, pp. 669–679, 2016, doi: 10.1016/j.ijforecast.2015.12.003.
[48] Z. Mo, "An empirical evaluation of OLS hedonic pricing regression on Singapore private housing market," Royal Institute of Technology, Stockholm, Sweden, 2014.
[49] H. Pham, "A new criterion for model selection," *Mathematics*, vol. 7, no. 12, pp. 1215–1227, 2019, doi: 10.3390/math7121215.
[50] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? – arguments against avoiding RMSE in the literature," *Geoscientific Model Development*, vol. 7, no. 3, pp. 1247–1250, Jun. 2014, doi: 10.5194/gmd-7-1247-2014.
[51] J. Rougier, "Ensemble averaging and mean squared error," *Journal of Climate*, vol. 29, no. 24, pp. 8865–8870, 2016, doi: 10.1175/JCLI-D-16-0012.1.
[52] H. Omara, M. Lazaar, and Y. Tabii, "Effect of feature selection on gene expression datasets classification accurac," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 5, pp. 3194–3203, Oct. 2018, doi: 10.11591/ijece.v8i5.pp3194-3203.
[53] R. Zhang, F. Nie, X. Li, and X. Wei, "Feature selection with multi-view data: A survey," *Information Fusion*, vol. 50, pp. 158–167, Oct. 2019, doi: 10.1016/j.inffus.2018.11.019.
[54] A. Farrell *et al*., "Machine learning of large-scale spatial distributions of wild turkeys with high-dimensional environmental data," *Ecology and Evolution*, vol. 9, no. 10, pp. 5938–5949, 2019, doi: 10.1002/ece3.5177.
[55] G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, no. 2, pp. 197–227, Jun. 2016, doi: 10.1007/s11749-016-0481-7.
[56] P. Zhang, "A novel feature selection method based on global sensitivity analysis with application in machine learning-based prediction model," *Applied Soft Computing*, vol. 85, Dec. 2019, doi: 10.1016/j.asoc.2019.105859.
[57] V. F. Rodriguez-Galiano, J. A. Luque-Espinar, M. Chica-Olmo, and M. P. Mendes, "Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods," *Science of The Total Environment*, vol. 624, pp. 661–672, May 2018, doi: 10.1016/j.scitotenv.2017.12.152.
[58] N. Arora and P. D. Kaur, "A bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment," *Applied Soft Computing*, vol. 86, Jan. 2020, doi: 10.1016/j.asoc.2019.105936.
[59] P. Zhang, Z.-Y. Yin, Y.-F. Jin, and T. H. T. Chan, "A novel hybrid surrogate intelligent model for creep index prediction based on particle swarm optimization and random forest," *Engineering Geology*, vol. 265, Feb. 2020, doi: 10.1016/j.enggeo.2019.105328.
[60] A. M. Qura, M. E. and Gad, "Regression estimation in the presence of outliers: A comparative study," *International Journal of Probability and Statistics*, vol. 5, no. 3, pp. 65 – 72, 2016, doi: 10.5923/j.ijps.20160503.01.
[61] S. Raghavendra and J. Santosh Kumar, "Performance evaluation of random forest with feature selection methods in prediction of diabetes," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 10, no. 1, pp. 353–359, 2020, doi: 10.11591/ijece.v10i1.pp353-359.
[62] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," *Sensors (Switzerland)*, vol. 18, no. 8, pp. 1–29, 2018, doi: 10.3390/s18082674.
[63] A. Cravero and S. Sepúlveda, "Use and adaptations of machine learning in big data-applications in real cases in agriculture,"

Electronics, vol. 10, no. 5, Feb. 2021, doi: 10.3390/electronics10050552.
[64] C. Maione and R. M. Barbosa, "Recent applications of multivariate data analysis methods in the authentication of rice and the most analyzed parameters: A review," *Critical Reviews in Food Science and Nutrition*, vol. 59, no. 12, pp. 1868–1879, Jul. 2019, doi: 10.1080/10408398.2018.1431763.
[65] J. M. Sánchez, J. P. Rodríguez, and H. E. Espitia, "Review of artificial intelligence applied in decision-making processes in agricultural public policy," *Processes*, vol. 8, no. 11, Oct. 2020, doi: 10.3390/pr8111374.
[66] K. Pawlak and M. Kołodziejczak, "The role of agriculture in ensuring food security in developing countries: Considerations in the context of the problem of sustainable food production," *Sustainability*, vol. 12, no. 13, Jul. 2020, doi: 10.3390/su12135488.
[67] P. Ekanayake, W. Rankothge, R. Weliwatta, and J. W. Jayasinghe, "Machine learning modelling of the relationship between weather and paddy yield in Sri Lanka," *Journal of Mathematics*, vol. 2021, pp. 1–14, May 2021, doi: 10.1155/2021/9941899.

BIOGRAPHIES OF AUTHORS

Mukhtar Ph.D student at the University Sains Malaysia, Malaysia. His research uses application of big data, statistics and machine learning. He can be contacted at email: mukhtar@untirta.ac.id.

Majid Khan Majahar Ali is a researcher and appointed fellow working in the field of seaweed cultivation, solar drying systems, processing, modelling and simulation. His research uses application of Iot, big data and simulation methods to improve model predictions of moisture losses during drying in control and uncontrolled environment. He is also interested in modelling the problems in engineering and other biological systems such as tissue culture and aquamarine. He employs modified hieracally multiple regressions using 8 selection criteria approach and robustness regression to overcome multicollinearity and outlier problems. He uses the techniques from statistical theory, approach and existing application tools to develop mathematical model and finally to transform the model in industry application and to answer a range of inspired questions. He can be contacted at email: majidkhanmajaharali@usm.my.

Mohd. Tahir Ismail is Associate Professor and researcher in the School of Mathematical Sciences, Universiti Sains Malaysia (USM), Malaysia. The research area is on financial time series. Particularly he is keen in the modeling and forecasting in time series analysis. He also analyses the economics issues. As of now, he has published more than 100 publications in reviewed journals and proceedings (some of them are listed in ISI, Scopus, Zentralblatt, MathSciNet and ther indices). He can be contacted at email: m.tahir@usm.my, Researcher ID: E-1242-2012.

Ferdinand Murni Hamundu Industrial Engineer from University of Indonesia in 2006. M.Sc and Ph.D degrees in Service Science and Innovation from School of Computer Sciences, Universiti Sains Malaysia, Malaysia in 2011 and 2021, respectively. His area of interest includes management information systems, technopreneurship, technology management, and big data analytics for information systems. He can be contacted at email: ferdinandmurni@gmail.com.

Alimuddin is Associate Professor**.** He received the B.S. Degree in Electrical Engineering from Moeslem University of Indonesia in 1999, Master of Engineering in Electrical Engineering University of Hasanuddin in 2003 and Master of Management Moeslem University of Indonesia in 2000 and PhD Degree in Agriculture Engineering Science Field Bioinformatic and Control Instrumentation IPB University-Sandwich Research Program PhD in Tsukuba University Japan 2012. He is Head of Smart Control System for Muti-Agent (SCEMA) and Head Vice of CELOFI Universitas Sultan Ageng Tirtayasa. His research interests include smart control, nonlinear control, adaptive control, multivariable control, process control, optimal control, artificial intelligence, power stability, model system identification, Engineering Optimation and Smart Farming. He associate Professor in Electrical Engineering, Faculty of Engineering Universitas Sultan Ageng Tirtayasa, Indonesia. He can be contacted at email: alimuddin@untirta.ac.id, alimuddineeuntirta@yahoo.com.

Naseem Akhtar pursuing PhD at the School of Industrial Technology, Universiti Sains Malaysia (USM), mainstream is hydrology, hydrogeology, and environment science. He is expert in groundwater exploration, water quality, water pollution, hydrochemistry and sustainality issues solving with the geophysical survey, pumping test, life cycle assessment, and machine learning. He can be contacted at email: naseemamu6@gmail.com.

Ahmad Fudholi joined the SERI as a lecturer in 2014. He involved more than USD 500,000 worth of research grant (24 grant/project). He supervised and completed more than 30 M.Sc and Ph.D students. To date, he has managed to supervise eleven Ph.D (seven as main supervisors and four as co-supervisor). He was also an examiner (three Ph.D and one M.Sc). His current research focus is renewable energy, particularly solar energy technology, micropower systems, solar drying systems and advanced solar thermal systems (solar-assisted drying, solar heat pumps, PVT systems). He has published more than 300 peer-reviewed papers, of which 105 papers are in the WoS index (50 Q1, impact factor of 5-14) and more than 200 papers are in the Scopus index. In addition, he has published more than 80 papers in international conferences. He has a total citation more than 3100 and a h-index of 27 in Scopus (Author ID: 57195432490). He has a total citation of 4630 and a h-index of 32 in Google Scholar. He has been appointed as reviewer of high-impact (Q1) journals, such as Renewable and Sustainable Energy Reviews, Energy Conversion and Management, Applied Energy, Energy and Buildings, Applied Thermal Engineering, Energy, Industrial Crops and Products and so on. He has also been appointed as editor of journals. He has received several awards. He owns one patent and two copyrights. He joined the LIPI as a researcher in 2020. He can be contacted at email: a.fudholi@gmail.com.

Journal IJECE Hybrid Model Q2

ORIGINALITY REPORT

9%	6%	7%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Sunwoo Han, Hyunjoong Kim. "On the Optimal Size of Candidate Feature Set in Random forest", Applied Sciences, 2019 Publication	1%
2	Vincenzo Verardi, Christophe Croux. "Robust Regression in Stata", The Stata Journal: Promoting communications on statistics and Stata, 2018 Publication	1%
3	shahrooz-abghari.github.io Internet Source	<1%
4	1library.net Internet Source	<1%
5	Siti Hasliza Ahmad Rusmili, Norizan Mohamed, Nor Azlida Aleng, Nur Farahana Zainudin. "Modeling of robust regression in breast tissue data", Applied Mathematical Sciences, 2015 Publication	<1%
6	iieta.org Internet Source	<1%

7	Konstantinos Liakos, Patrizia Busato, Dimitrios Moshou, Simon Pearson, Dionysis Bochtis. "Machine Learning in Agriculture: A Review", Sensors, 2018 Publication	<1 %
8	Submitted to University of Duhok Student Paper	<1 %
9	www.writingforlegacy.com Internet Source	<1 %
10	journal.biotrop.org Internet Source	<1 %
11	www.math.hkbu.edu.hk Internet Source	<1 %
12	Jiri Maslan, Ludek Cicmanec. "Setting the Flight Parameters of an Unmanned Aircraft for Distress Detection on the Concrete Runway", 2021 International Conference on Military Technologies (ICMT), 2021 Publication	<1 %
13	www.hydrol-earth-syst-sci-discuss.net Internet Source	<1 %
14	Paulo Angelo Alves Resende, André Costa Drummond. "A Survey of Random Forest Based Methods for Intrusion Detection Systems", ACM Computing Surveys, 2018 Publication	<1 %

15	fse.studenttheses.ub.rug.nl Internet Source	<1 %
16	media.neliti.com Internet Source	<1 %
17	José Luis Rodríguez–Gutiérrez,, Lady Johana Correa-Higuera, Andrés Enrique Alvarado, Jorge Alberto Chaparro-Pesca. "Evaluación de la actividad alelopática de extractos crudos de Copaifera pubiflora (Benth), sobre la germinación de Mimosa pudica (Lineo)", Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales, 2016 Publication	<1 %
18	"New Statistical Developments in Data Science", Springer Science and Business Media LLC, 2019 Publication	<1 %
19	rboutaba.cs.uwaterloo.ca Internet Source	<1 %
20	tel.archives-ouvertes.fr Internet Source	<1 %
21	www.pubfacts.com Internet Source	<1 %
22	Rei Sonobe, Yuta Miura, Tomohito Sano, Hideki Horie. "Monitoring Photosynthetic Pigments of Shade-Grown Tea from	<1 %

Hyperspectral Reflectance", Canadian Journal of Remote Sensing, 2018
Publication

23	ojs.upsi.edu.my Internet Source	<1 %
24	vtechworks.lib.vt.edu Internet Source	<1 %
25	www.cs.man.ac.uk Internet Source	<1 %
26	www.mdpi.com Internet Source	<1 %
27	Amin Hashemi, Mohammad Bagher Dowlatshahi, Hossein Nezamabadi-pour. "MGFS: A multi-label graph-based feature selection algorithm via PageRank centrality", Expert Systems with Applications, 2020 Publication	<1 %
28	Hossein Sadr, Mir Mohsen Pedram, Mohammad Teshnehlab. "Multi-View Deep Network: A Deep Model Based on Learning Features From Heterogeneous Neural Networks for Sentiment Analysis", IEEE Access, 2020 Publication	<1 %
29	Judong Shen, Z. J. Pei, G. R. Fisher, E. S. Lee. "Modelling and analysis of waviness reduction in soft-pad grinding of wire-sawn silicon	<1 %

wafers by support vector regression", International Journal of Production Research, 2006
Publication

30	Submitted to KDU College Sdn Bhd Student Paper	<1 %
31	Md Nazirul Islam Sarker, Min Wu, Bouasone Chanthamith, Shaheen Yusufzada, Dan Li, Jie Zhang. "Big Data Driven Smart Agriculture: Pathway for Sustainable Development", 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD), 2019 Publication	<1 %
32	dokumen.pub Internet Source	<1 %
33	mafiadoc.com Internet Source	<1 %
34	philpapers.org Internet Source	<1 %
35	pjsor.com Internet Source	<1 %
36	sith.itb.ac.id Internet Source	<1 %
37	Baharuddin Hamzah, Noor Jalaluddin, Abdul Wahid Wahab, Ambo Upe. "Copper(II) Extraction from Nitric Acid Solution with 1-	<1 %

Phenyl-3-methyl-4-benzoyl-5-pyrazolone as a Cation Carrier by Liquid Membrane Emulsion", E-Journal of Chemistry, 2010
Publication

38 Ismail, R.. "A comparison of regression tree ensembles: Predicting Sirex noctilio induced water stress in Pinus patula forests of KwaZulu-Natal, South Africa", International Journal of Applied Earth Observations and Geoinformation, 201002 <1%
Publication

39 Inayatullah, Huma Qayyurn. "An improved comparative model for chronic kidney disease (CKD) prediction", 2020 14th International Conference on Open Source Systems and Technologies (ICOSST), 2020 <1%
Publication

40 Neha Chaudhuri, Gaurav Gupta, Vallurupalli Vamsi, Indranil Bose. "On the platform but will they buy? Predicting customers' purchase behavior using deep learning", Decision Support Systems, 2021 <1%
Publication

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		