Research Article

Machine learning-based variable selection: An evaluation of Bagging and Boosting

Mukhtar[1,3], M.K.M. Ali[1], Mohd. Tahir Ismail[1], Ferdinand Murni Hamundu[2*], and Alimuddin[4]

[1] School of Mathematical Sciences, Universiti Sains Malaysia, USM-11800, Pulau Penang, Malaysia
[2] Faculty of Mathematics and Natural Science, Universitas Halu Oleo, Kendari, Indonesia
[3] I-CEFORY (Local Food Innovation), Universitas Sultan Ageng Tirtayasa
[4]Departemen of electrical Engineering, Faculty of Engineerimg, Universitas Sultan Ageng Tirtayasa
[*]**Corresponding author's e-mail:** mukhtar@untirta.ac.id ; majidkhanmajaharali@usm.my

Article History: Received: 24 January 2021; Revised: 25 February 2021; Accepted: 28 March 2021; Published: 4 June 2021

Abstract

Variable selection is a necessary step to build a useful regression. In this paper, an evaluation of different methods (variable selections) including Bagging and Boosting were performed. Large datasets from 1924 observations were taken and the second interaction data which contains 435 variables were employed. In big data, there is no single variable selection technique that is robust towards different families of regression algorithm. The existing variables techniques produce different results with different predictive models. Variable selections only provide the rank of important variables which means that the techniques did not have rules in selecting the suitable range of variable importance. Each of 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, and 120 highest variable important were selected. Several validations such as Sum Square of Error (SSE), R-square, and Mean Square Error (MSE) were used to compare its performances. As the result, bagging for the 90 highest variable important was better than others SSE (31077.8295), R-square (0.9210), and MSE (17.8344), respectively. Hence, the variable selection using bagging has been considered as the best model.

Keywords: Machine Learning, Bagging, Boosting, Variable Selection, and Important Variables

Introduction

Machine learning is a scientific method that focuses on design and development. It employs algorithms to produce worth models based on empirical data with the purpose to generate knowledge (Alpaydin, 2020). In practice, the most challenging aspect of machine learning is variable selection due to the possibility of the presence of unimportant variables (Matin et al., 2018). Machine learning-based variable selection has attracted the attention of researchers, particularly in today's big data era (Bagherzadeh-Khiabani et al., 2016; H. H. Kim & Swanson, 2018).

Numerous machine learning has been suggested to handle big data problems (Saidulu & Sasikala, 2017; Zhou et al., 2017), but the limitation of machine learning cannot provide how many important and unimportant variables. Machine learning only provides the rank of important variables (Drobnič et al., 2020), which means that the techniques did not have rules in selecting the suitable range of variable importance. The important variable is the ranking of the independent variables that contribute to the dependent variable. Important variable is a suitable of the variable from original variables (Gómez-Verdejo et al., 2019; Thi et al., 2017).

Researchers have developed several techniques in variables selection, and thus it should be further explored for practical data analysis. In general, the process of variable selection aims to identify a subset of predictors categorized as important variables. In big data, there is no single variable selection technique that is robust towards different families of regression algorithm. The existing variables selection techniques produce different results with different predictive models. It can be a problem in determining the best predictive model while working with big data (Xu, 2012).

Besides that, limited studies have compared and evaluated the performance of multiple machine learning techniques for regression models. Researchers have concerned and interested in the relationship between the dependent and independent variables. An important issue in the regression models is the variable selection, and the selection is most relevant to the regression task, which provides a fundamental step in the data analysis. The accuracy can be improved (Cai et al., 2009)

This study will provide employee seaweed data with several variables including hourly solar radiation, temperature, humidity, and moisture content. The dataset containing 1924 observations, will be used to study the effect of 29 different independent variables on the one dependent variable. The second interaction data, which contains 435 different interactions of independent variables on the one dependent variable will be implemented. The more detailed table for each interaction variable with all computed scores is attached in [**Appendix** 1]. We will compare subsets of the number of important variables. After comparing the subset, the important variables

then calculating its validation the determining the important optimal variable. The primary focus of this study is to analyse and compare the impact of two different important variable ranking techniques regression algorithms such as Bagging and Boosting on each the 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, and 120 highest important variables.

Materials and methods

Literature reviews

Machine Learning

The purpose of machine learning is to learn from the data (Qiu et al., 2016). Several machine learning algorithms are available to construct predictive models. Machine learning is a field in data analytic that focuses on the development of mathematical algorithms to predict future (Najafabadi et al., 2015). Computer or system in machine learning can learn from the past data. The computer or system analyses big data and finds patterns and rules hidden in the data. Machine learning requires cross-disciplinary proficiency in several areas such as data mining, theory of probability, cognitive science, pattern recognition, and theory of computer science. Two major categories of machine learning such as classification with the dependent variable is discrete (classes) and regression with the dependent variable is continuous.

Variable Selection

Concerning the regression, it is beneficial to choose and maintain a subset of variables with a predictable ability. The purpose of variable selection usually are:

1) To enhance the capability of predictive model,
2) To avoid the obstacle correlated with measuring all the variables and
3) To present a broader understanding of the predictive model, and with data expansion, by reducing unimportant variables (Guyon & Elisseeff, 2003).

Several variables in the regression model can be an issue if there are unimportant variables. Unimportant variables can lead to overfitting, in which the unimportant variables influence on the wrong decision in the regression model. The presence of unimportant variables in the empirical analysis must be addressed since unimportant variables does not have a contribution and will create noise to the regression model (Omara et al., 2018). Variable selection is address for unimportant variables. Variable selection is to determine the best subset to use in regression model for large number of variables, and thus the proper methods are needed to identify the important variables.

Variable selection results important variables. Measuring important variable for computational models or measured data is an important task in many applications. Important variable represents each variable's machine learning important in the data concerning its effect on the generated model. Important variable relates to the dataset that effects the generated model. Important variables are the ranking of the independent variables that contribute to the modelling. Important variables are suitable subset of variable from original variables (Tran et al., 2018).

Regression

Dataset $D = \{(X_i, Y_i): X_i \in \mathbb{R}^{p \times n}, Y_i \in \mathbb{R}\}$ are learning algorithm with underlying function $Y = f(x)$ where the where the $X_i s$ are independent variables and Y_i is dependent variables with p is the number of independent variables and n observations (Botta et al., 2014). The dependent variable can be written as $Y = (Y_1, \dots, Y_n)$. Regression learning tasks can be stated as learning a function $\varphi: X \to Y$ from a learning set $\mathcal{L} = (X, Y)$. The purpose of regression learning is to find a model in such that its prediction $\varphi(\boldsymbol{x})$ which denoted by $\hat{Y}$ that as good as possible and Y_i is continuous (Geurts et al., 2006; Shahhosseini et al., 2019).

Bagging

The data is (X_i, Y_i) $(i = 1, \dots, n)$, where $X_i \in \mathbb{R}^d$ are n-observations the independent variables and $Y_i \in \mathbb{R}$ is the dependent variable. The function estimator is $\hat{g}(\cdot) = h_n\big((X_i, Y_i), \dots, (X_i, Y_i)\big)(\cdot): \mathbb{R}^d \to \mathbb{R}$ where the function $h_n(\cdot)$ defines the estimator as a function of the data (Li & Chen, 2020).

Algorithm Bagging

Input: D - training set, ES – number of the sampled subsets or base models, L – base learner
Output: M – a set of base models, B - bagging ensemble

1. For $i \in \{1,2, \dots, ES\}$ **do**:
2. Randomly generate a subset $D_i = Bootstrap(D)$

3. Base model $g_i = L(D_i)$ is established using base regression L trained on the subset D_i
4. $g_i = g \cup (g_i)$
5. The outcome $g(x)$ of a test sample x predicted by the ensemble model g is given as follows: $g(x) = \frac{1}{N}\sum_{T=1}^{N} h_T(x)$

Boosting

The dataset sample $\{\boldsymbol{x_i}, y_i\}_{i=1}^{N}$ of known $(\boldsymbol{x}, y)$- values. This aim is to get an approximation $\hat{F}(\boldsymbol{x})$. The function $F^*(x)$ aims mapping $\boldsymbol{x}$ **to** $\boldsymbol{y}$ which minimizes the fitted value for loss function $L\big(y, F(x)\big)$ over the distribution of $(\boldsymbol{x}, y)$ (Friedman, 2001). Frequently employed loss function $L(y, F)$ include squared error $(y - F)^2$ and absolute error $|y - \boldsymbol{F}|$ for $y \in R$.

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 $f_t: X \to \mathbb{R}$
- Choose $\alpha_t \in \mathbb{R}$
- Update:

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

Where Z_t is a normalization factor (chosen so that D_{t+1} will be a distribution)

Output the final regression:

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

Data

Data were taken from the experimental drying process of seaweed drier and have optimized for modelling analysis by using machine learning such as Bagging and Boosting. The data was collected from 8.00 am until 5.00 pm starting on 08/04/2017 to 12/04/2017. The original data was for each second and then it was converted in an hour for data analysis. The variables taken are data that contain hourly solar radiation, temperature, humidity, and moisture content. The detailed factor of modelling is shown in Table 1.

Table 1. 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 (℃) ambient
T2, T3, T4	Independent	Temperature (℃) before enter solar collector
T5	Independent	Temperature (℃) in front of down v-Groove (Solar Collector)
T6, T8	Independent	Temperature (℃) in front of up v-Groove (Solar Collector)
T7, T14, T15, T16, T21, T22	Independent	Temperature (℃) Solar Collector
T8, T9, T10, T11, T12	Independent	Temperature (℃) behind inside chamber
T13, T17, T18, T19, T23	Independent	Temperature (℃) Infront of (Inside Chamber)
T20, T23, T24, T25, T28	Independent	Temperature (℃) from solar collector to chamber

The dataset containing 1924 observations will use to study the effect of 31 different independent variables on the one dependent variable. Significance of interaction terms had also been observed in this study. Thus, 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 interactions of independent variables on the one dependent variable. The more detailed tables for each variable interaction are attached in [Appendix 1].

Flowchart

The flowchart is depicted as in Figure 1 for the complete view of the building model regression algorithm. In this study will compare the validation models such as Sum Square of Error, R-square, and Mean Square Error by the implementation of the framework both Bagging and Boosting on each the 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120 highest variable important for determining the best model.

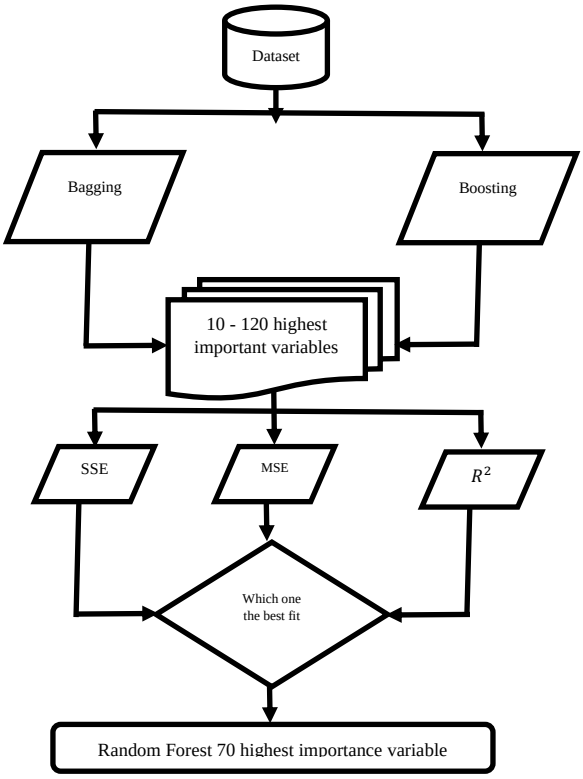

Figure 1. Flow Chart of Modelling

Validation of model

Evaluation model metrics are required to assess the model's correctness. It is important to verify whether the model is adequate, that is, whether the model correctly predicts the target (dependent) variable within a reasonable range of accuracy (Hallman, 2019). The metrics validation including Sum Square of Error (SSE), R-Square, and Mean Square Error (MSE) are measured for evaluating the model performances. The formula of the metrics is shown in Table 2.

Table 2. Validation Model Metric

Validation	**Formulation**	Reference
Sum of Square Error (SSE)	$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$	(S. Kim & Kim, 2016)
Sum of Squared Total (SST)	$SST = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2$	(S. Kim & Kim, 2016)
Mean Squared Error (MSE)	$MSE = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{Y - \hat{Y}_i}{\hat{Y}_i}\right)^2$	(S. Kim & Kim, 2016)

Validation	**Formulation**	Reference
R-square	$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$	(Hallman, 2019; Schneider et al., 2010)

Results and discussion

Results

The primary focus of this paper is to analyze and compare the impact of the three different variable importance ranking techniques over three different regression algorithms for the data seaweed drying. In the methodology section, we have described the three variable important ranking techniques that have been used in this variable important ranking experiment.

Table 3. The 10 highest for variable important

No	Methods	Variable Importance
2	Bagging	T5,T4,T1,T3,T2,T6,H5,T7,T10,T8
3	Boosting	T2*T6,T1*T6,H5*PY,T7*H1,T5*PY,T21*H5,T8*PY,T7*T9,T8,T2*T7

Table 3 shows the final results that was obtained by each variable important ranking technique. All the important variable was ranked according to their importance score computed by their respective techniques. The more detailed tables for each variable important ranking technique with all computed scores are attached in the [Appendix 2].

The results will compare the validation model such as SSE, R-square, and MSE by the implementations of the framework both Bagging and Boosting on each the 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, and 120 highest important variables. Table 4 depicts these comparisons for each method.

Table 4. Result of Validation for Machine Learning

Range of Important Variable (Highest)	Bagging			Boosting		
	SSE	R^2	MSE	SSE	R^2	MSE
10	43027.5588	0.8907	28.8051	267242.5185	0.4967	140.2975
20	36135.8887	0.9082	21.3855	264954.6439	0.4989	139.0456
30	39141.8210	0.9007	22.6504	266185.3358	0.4939	139.7178
40	36877.8917	0.9065	20.5602	250684.5671	0.5183	131.9103
50	37227.5883	0.9056	20.7513	230633.0044	0.5487	121.4154
60	36382.5806	0.9078	20.3124	223362.1246	0.5606	117.6250
70	32155.8191	0.9182	18.6355	232947.3212	0.5467	122.4461
80	31768.4721	0.9192	18.2617	235921.0238	0.5431	123.8871
90	31077.8295	0.9210	17.8344	240921.7012	0.5323	126.6276
100	31269.8772	0.9205	18.0012	238582.3347	0.5364	125.2482
110	31230.9307	0.9206	17.9111	249589.3707	0.5220	130.8553
120	32966.4901	0.9162	19.1018	247648.0966	0.5240	129.9450

Predefined validation model for Bagging and Boosting are given in table 4. All validation model measures such as SSE, R-square, and MSE indicate that significantly better results were obtained by Bagging for the 90 highest variables in comparison to others.

Discussion

According to the table 3 and of all the methods are given. The bagging for the 90 highest important variables with SSE (31077.8295), R-square (0.9210), and MSE (17.8344), respectively. Boosting is the 60 highest important variables with the SSE (223362.1246), R-square (0.5606), and MSE (117.6250), respectively. In short,

we can conclude that the Bagging for the 90 highest important variables has generated the lowest error data, which provides the most relevant data in the context of validation such as SSE, R-square, and MSE.

The SSE, R-square, and MSE are useful measure widely used in validation model. The lowest of SSE and MSE are bagging than boosting. The SSE and MSE are used in explaining how well the regression model is toward to the model data. In particular, the explained SSE and MSE measure the variation for the error between the predicted and actual data. The MSE and SSE measure the discrepancy the data and an estimation model. Generally, the lower MSE and SSE show which model can better explain, and the higher MSE and SSE show which model poorly describes the data (H.-Y. Kim, 2018).

The R – square is a statistics measure for measuring a regression model's validity. The R – square could be interpreted as a proportion of variance of a predicted outcome. The R – square has ranges from 0 to 1 (Hallman, 2019; Schneider et al., 2010). The R-square measures variation which was accounted for the predicted data. The highest R – square of bagging (0.9210) suggest that the dependent variable was predicted 92.10% by the independent variables.

The issue of boosting is to select the right weak learner which is applying the number of weak learners M. If M is too small that boosting regressor will not learn the complexities of data well and will result in underfitting. If M is too large that will overfitting and it will learn the noises and the distribution bias than the true general patterns (Htike, 2017; Li & Chen, 2020)

Bagging (bootstrapping aggregation) improves prediction accuracy and reduces variance and solves overfitting issues. Bagging is a sampling method (Momparler et al., 2016). Bagging is to develop various training sets with the bootstrap sampling and the last model will be achieved by aggregating these base learners. Bagging has two essential components: bootstrap sampling and model aggregation. Bootstrap sampling is to take n samples with the replacement to assure the autonomy of different sampling training. In addition, the prediction for regression is $\mathbb{E}[Y|X = x]$ (Li & Chen, 2020). The advantage of bagging is its ability to reduce the variance of inaccuracy which relates to the degree of instability in regression. The issue is a small change in training: the more unstable in regression (Kotsiantis, 2011).

Conclusions

In this study, we compared the validation models such as SSE, R-Square, and MSE by the implementation of the framework Bagging and Boosting on each the 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, and 120 highest variable important. In comparisons, the bagging for the 90 highest important variables was the most accuracy results of SSE (31077.8295), R-square (0.9210), and MSE (17.8344), respectively. The bagging exhibited the lowest error data which provides the most relevant data of the result.

Acknowledgements

We acknowledged Universitas Sultan Ageng Tirtayasa and Universitas Sains Malaysia

References

1. Bagherzadeh-Khiabani, F., Ramezankhani, A., Azizi, F., Hadaegh, F., Steyerberg, E. W., & Khalili, D. (2016). A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. *Journal of Clinical Epidemiology*, *71*, 76–85. https://doi.org/10.1016/j.jclinepi.2015.10.002
2. Botta, V., Louppe, G., Geurts, P., & Wehenkel, L. (2014). Exploiting SNP correlations within random forest for genome-wide association studies. *PLoS ONE*, *9*(4). https://doi.org/10.1371/journal.pone.0093379
3. Cai, A., Tsay, R. S., & Chen, R. (2009). Variable selection in linear regression with many predictors. *Journal of Computational and Graphical Statistics*, *18*(4), 573–591. https://doi.org/10.1198/jcgs.2009.06164
4. Drobnič, F., Kos, A., & Pustišek, M. (2020). On the interpretability of machine learning models and experimental feature selection in case of multicollinear data. *Electronics (Switzerland)*, *9*(5). https://doi.org/10.3390/electronics9050761
5. Friedman, J. (2001). Greedy Function Approximation : A Gradient Boosting Machine Author (s): Jerome H . Friedman Source : The Annals of Statistics , Vol . 29 , No . 5 (Oct ., 2001), pp . 1189-1232 Published by : Institute of Mathematical Statistics Stable URL : http://www. *The Annals of Statistics*, *29*(5), 1189–1232.
6. Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning*, *63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

7. Gómez-Verdejo, V., Parrado-Hernández, E., & Tohka, J. (2019). Sign-Consistency Based Variable Importance for Machine Learning in Brain Imaging. *Neuroinformatics*, *17*(4), 593–609. https://doi.org/10.1007/s12021-019-9415-3
8. Guyon, Isabelle., & Elisseeff, Andre. (2003). An Introduction to Variable and Feature Selection. *Analytica Chimica Acta*, *703*(2), 152–162. https://doi.org/10.1016/j.aca.2011.07.027
9. Hallman, J. (2019). *A comparative study on Linear Regression and Neural Networks for estimating order quantities of powder blends*.
10. Htike, K. K. (2017). Efficient determination of the number of weak learners in AdaBoost. *Journal of Experimental and Theoretical Artificial Intelligence*, *29*(5), 967–982. https://doi.org/10.1080/0952813X.2016.1266038
11. Kim, H. H., & Swanson, N. R. (2018). Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. *International Journal of Forecasting*, *34*(2), 339–354. https://doi.org/10.1016/j.ijforecast.2016.02.012
12. Kim, H.-Y. (2018). Statistical notes for clinical researchers: simple linear regression 2 – evaluation of regression line. *Restorative Dentistry & Endodontics*, *43*(3), 1–5. https://doi.org/10.5395/rde.2018.43.e34
13. Kim, S., & Kim, H. (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting*, *32*(3), 669–679. https://doi.org/10.1016/j.ijforecast.2015.12.003
14. Kotsiantis, S. (2011). Combining bagging, boosting, rotation forest and random subspace methods. *Artificial Intelligence Review*, *35*(3), 223–240. https://doi.org/10.1007/s10462-010-9192-8
15. Li, Y., & Chen, W. (2020). A comparative performance assessment of ensemble learning for credit scoring. *Mathematics*, *8*(10), 1–19. https://doi.org/10.3390/math8101756
16. Matin, S. S., Farahzadi, L., Makaremi, S., Chelgani, S. C., & Sattari, G. (2018). Variable selection and prediction of uniaxial compressive strength and modulus of elasticity by random forest. *Applied Soft Computing Journal*, *70*, 980–987. https://doi.org/10.1016/j.asoc.2017.06.030
17. Momparler, A., Carmona, P., & Climent, F. (2016). Banking failure prediction: a boosting classification tree approach. *Revista Espanola de Financiacion y Contabilidad*, *45*(1), 63–91. https://doi.org/10.1080/02102412.2015.1118903
18. Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1–21. https://doi.org/10.1186/s40537-014-0007-7
19. Omara, H., Lazaar, M., & Tabii, Y. (2018). Effect of Feature Selection on Gene Expression Datasets Classification Accurac. *International Journal of Electrical and Computer Engineering (IJECE)*, *8*(5), 3194. https://doi.org/10.11591/ijece.v8i5.pp3194-3203
20. Qiu, J., Wu, Q., Ding, G., Xu, Y., & Feng, S. (2016). A survey of machine learning for big data processing. *Eurasip Journal on Advances in Signal Processing*, *2016*(1). https://doi.org/10.1186/s13634-016-0355-x
21. Saidulu, D., & Sasikala, R. (2017). Machine learning and statistical approaches for big data: Issues, challenges and research directions. *International Journal of Applied Engineering Research*, *12*(21), 11691–11699.
22. Schneider, A., Hommel, G., & Blettner, M. (2010). Lineare regressionsanalyse - Teil 14 der serie zur bewertung wissenschaftlicher publikationen. *Deutsches Arzteblatt*, *107*(44), 776–782. https://doi.org/10.3238/arztebl.2010.0776
23. Shahhosseini, M., Hu, G., & Pham, H. (2019). *Optimizing Ensemble Weights and Hyperparameters of Machine Learning Models for Regression Problems*. 1–22. http://arxiv.org/abs/1908.05287
24. Thi, H. A. le, Le, H. M., Phan, D. N., & Tran, B. (2017). Stochastic DCA for the large-sum of non-convex functions problem and its application to group variable selection in classification. *34th International Conference on Machine Learning, ICML 2017*, *7*, 5211–5220.
25. Tran, B., Xue, B., & Zhang, M. (2018). A New Representation in PSO for Discretization-Based Feature Selection. *IEEE Transactions on Cybernetics*, *48*(6), 1733–1746. https://doi.org/10.1109/TCYB.2017.2714145
26. Zhou, L., Pan, S., Wang, J., & Vasilakos, A. v. (2017). Machine learning on big data: Opportunities and challenges. *Neurocomputing*, *237*(December 2016), 350–361. https://doi.org/10.1016/j.neucom.2017.01.026