Received: October 18, 2019. Revised: January 2, 2020.

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Microorganism Estimation in a Shrimp Pond Using Gaussian Process Regressor and Gradient Tree Boosting

Oskar Natan[1]* **Agus Indra Gunawan[1]** **Bima Sena Bayu Dewantara[1]** **Junaedi Ispianto[2]**

[1]*Politeknik Elektronika Negeri Surabaya, Indonesia*
[2]*Asosiasi Tambak Intensif, Indonesia*
* Corresponding author's Email: theoskarnatan@gmail.com

Abstract: The balance of the aquatic ecosystem is an influential factor in the world of aquaculture, especially in shrimp cultivation. The one that plays a role in that ecosystem is aquatic microorganisms such as vibrio, bacteria, and algae. Therefore, farmers need to know their number and ratio to maintain the shrimp growth. Thus, in this research, models that can estimate vibrio-bacteria ratio and number of algae are developed. These models are formed from aquaculture datasets which are modeled using machine learning algorithms named Gaussian process regressor (GPR) and gradient tree boosting (GTB). Other processing techniques like data pre-processing, feature decomposition, and optimization are also applied to improve model performance. Moreover, these models are also compared to other models which are modeled using another machine learning algorithm like support vector regression (SVR), Lasso, and kernel ridge regression (KRR), so that the best models can be determined. Based on k-fold cross-validation, the GPR model has the best performance in estimating the vibrio-bacteria ratio with mean absolute error (MAE) value of 0.02482 and explained variance score of 0.96515. Then, in the algae estimation, the best performance is achieved by the GTB model with MAE value of 6.55554 and explained variance score of 0.33001.

Keywords: Aquaculture, Shrimp cultivation, Gaussian process regressor, Gradient tree boosting, Machine learning, Algorithm.

1. Introduction

Success in shrimp cultivation is not only influenced by feeding techniques or water quality supervision but also the biological factor and its proper treatment to protect the ecosystem that exists in the pond [1]. In aquatic ecosystems, there are various types of organisms and microorganisms that give influence one another [2]. This ecosystem needs to be balanced so that a large harvest can be achieved. However, the farmer often facing difficulties in estimating the microorganisms which are very crucial since it affects the shrimp growth [3]. Thus, this research is focusing on vibrio, algae, and bacteria that can be a shrimp food and support the shrimp growth but also can be a disease if excess [4]. By knowing the condition of microorganisms, it is expected that the farmer could give proper treatment to balance the ecosystem and prevent massive losses.

Given these problems, an estimation model is needed to assist farmers in estimating the biomass in the pond. In [5], an unmanned aerial vehicle (UAV) is used to take a picture and an image processing algorithm is used to estimate total biomass by using a certain method called green algae attached to nursery-net (GAAN) and green algae attached to rope (GAAR). Similar work in [6] is also using UAV and process the obtained image using the normalized green-red difference index (NGRDI). Another image-based approach also conducted in [7, 8]. However, using an image to make a prediction often facing the problem of uneven illumination. Thus, we used a machine learning approach to create a knowledge model to make a prediction. This technique does not require an image but it needs a set of microorganism data from a statistical yearbook. This model works by processing the input data in the form of easily recognizable microorganism levels

such as green vibrio, yellow vibrio, diatoms, dinoflagellate, protozoa, green algae, and blue-green algae. The model itself is obtained by applying a machine learning algorithm to model an aquaculture dataset. Machine learning has been widely used to create varying knowledge model in many fields. In [9], K-Nearest Neighbour (KNN) classifier is used to predict a shrimp pond condition based on several water parameters. Then, another work is done in [10] by applying SVR to create a model to measure the feasibility of asphalt concrete. The use of SVR is also carried out by [11] for facial expression recognition. In that study, the SVR model is compared with the model of relevance vector regression (RVR) in terms of performance during the testing process. Furthermore, a neural network (NN) is used to estimate the number of end-of-life vehicles in China [12]. Machine learning can also be used for classification problems. In [13], deep learning is used to classify the condition of preeclampsia during the pregnancy process. Another machine learning application in the medical field is also carried out by [14] where fuzzy logic and decision tree are used for decision making.

In this research, the GPR algorithm is used to create VB ratio estimation model and GTB algorithm to make an algae estimation model. Both algorithms have several parameters to be tuned to give a higher performance in modeling the data. Thus, an optimization algorithm called grid search is applied to find the best parameters. In addition, several data pre-processing techniques such as impute missing value, min-max normalization, and feature decomposition using principal component analysis (PCA) are also used in the modeling process. Finally, the GPR and GTB models are also compared with other models formed by another algorithm like SVR, Lasso, and KRR. With varying schemes of processing, it is expected to form models that have good performance in estimating the microorganism. The rest of this paper is organized as follows. In Section 2, we describe our approach in detail from dataset information to model evaluation. In Section 3, we discussed the model's performance in detail. Section 4 concludes this paper.

2. Methods

This study aims to form a model for VB ratio and total algae using GPR and GTB. In this chapter, we explain each processing step in detail, starting from dataset information until knowledge modeling. Then, a comparative test to measure the model performance for each algorithm is performed.

Table 1. Dataset information

Specification	VB Ratio	Algae Estimation
Task	Regression	Regression
Number of attributes	4 input and 1 output	5 input and 1 output
Number of samples	262	188
Data types	Numerical	Numerical
Missing values	Yes	Yes

2.1 Dataset information

There are 2 datasets used for making the model. The first dataset is used to make a model of VB ratio estimation using GPR algorithm. Then, the second dataset is used to make a model of algae estimation using GTB algorithm. Both datasets are acquired from several shrimp ponds in Bulukumba, South Sulawesi. The detailed information of these datasets can be seen in Table 1.

Both datasets contain the condition of the pond which is represented by the amounts of algae and VB ratio. All input attributes (independent variables) are measured through laboratory tests. In the VB ratio dataset, there are 4 inputs namely green vibrio, yellow vibrio, total vibrio count (TVC) and total bacterial count (TBC). Whereas, in the algae estimation dataset there are 5 input attributes namely green algae, blue-green algae, diatoms, dinoflagellates, and protozoa. Both datasets have one output attribute (dependent variable). The detailed explanation of each attribute in these datasets can be seen in Table 2.

2.2 Data preprocessing

As mentioned before, these datasets contain several missing values. Missing values can be caused by measurement error or human error. If this is ignored, it can lead to failure in the modeling process [15]. In this research, a simple impute missing value technique called averaging as seen in Eq. (1) is used to fill the missing values.

$$d_m = \frac{\sum_{k=1}^{K} x_{mk}}{Km} \quad (1)$$

Where d_m is the average value, x_{mk} is the non-missing value of sample k in the attribute m, Km is the total of non-missing value in the attribute m. A dataset can contain one or many attributes where each attribute has its own value that represents the nature where the sample is taken.

Table 2. Attributes information

Attribute	Description	Units
Green Vibrio	Number of *Green Vibrio* / *Vibrio Harveyi* / *pathogen Vibrio*	cfu/ml
Yellow Vibrio	Number of *Yellow Vibrio* / good bacteria	cfu/ml
TVC	Total vibrio in the pond	cfu/ml
TBC	Total bacteria in the pond	cfu/ml
VB Ratio	Ratio of Vibrio and Bacteria	-
Green Algae	Number of *chlorophyta*	%
Blue-Green Algae	Number of *cyanophyta* / *cyanobacteria*	%
Diatom	Number of plankton algae	%
Dinoflagellata	Number of dinoflagellate / protista algae	%
Protozoa	Number of protozoa	%
Total Algae	Number of algae	cell/ml

Therefore, the range of each attribute is possible to be different. This can cause an inequality during the modeling process since the attributes with large range are more influential compared to attributes that have a small range. To overcome this problem, data normalization needs to be done so that each attribute has the same influence on the output attributes [16]. In this research, a feature scaling technique called min-max scaling is used to normalize the data where each attribute value is normalized to 0 to 1. The formula for min-max scaling can be seen in Eq. (2).

$$x'_{mk} = \frac{x_{mk} - \min(x_m)}{\max(x_m) - \min(x_m)} \quad (2)$$

Where x'_{mk} is the normalized value of x_{mk}, x_{mk} is the sample k in attribute m, x_m is all sample values in attribute m. Then, a feature decomposition algorithm called PCA is applied to extract more values in the dataset. This process is performed to make each value in each attribute more differentiate in representing the value in the output attribute. PCA is a technique that implements orthogonal transformation to change a set of sample observations into a set of values that are not linearly correlated or called principal components [17]. Mathematically, PCA as in Eq. (3) is defined by a set of p-dimensional vectors of coefficients $w_{(k)} = (w_1, \dots w_p)_{(k)}$ that map each row vector $x_{(i)}$ of sample X to a new vector $t_{(i)} = (t_1, \dots t_l)_{(i)}$.

$$t_{k(i)} = w_{(k)} x_{(i)} \quad (3)$$
$$for\ i = 1, \dots, n \quad and\ \ k = 1, \dots, l$$

Where l is the number of principal components (usually less than p) and i is vector index in row vector $x_{(i)}$ of sample X. The number of principal components cannot exceed the number of attributes. In other words, the number of principal components (l) is in accordance with $\min(m-1, l)$ where m is the number of observed attributes and l is the number of principal components. The number of principal components used in the VB ratio dataset is 2 and 4 (PCA2 and PCA4) while in the algae dataset is 3 and 5 (PCA3 and PCA5). The detailed schemes for each data pre-processing step before the modeling process can be seen in Fig. 1.

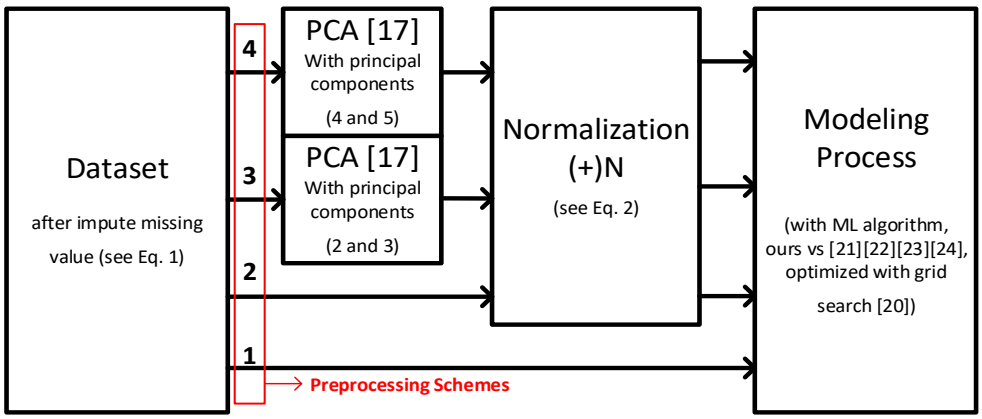

Figure.1 Schemes of data preprocessing

In the first scheme, the dataset is directly modeled without any preprocessing. This scheme is called raw processing. In the second scheme, the dataset is processed by min-max scaling so that each data in the dataset has the same values range. In the third scheme, PCA is used to explore the dataset before modeling. The principal component used is 2 for the VB ratio dataset and 4 for the algae estimation, then the data is normalized again. Furthermore, the fourth scheme is the same as the third scheme, but with different number of principal components used, 4 for the VB ratio and 5 for algae estimation.

2.3 Modeling

The preprocessed datasets are then modeled using the machine learning algorithm. As explained in the previous subchapter, VB ratio dataset is modeled using GPR algorithm. GPR implements the Gaussian process in the regression process on the dataset. The algorithm predicts a value by measuring the closeness of some previous data. In other words, the GPR algorithm computes the weighted average of known values to form a knowledge model [18]. The GPR model can be written as in Eq. (4).

$$\acute{Z}(x_0) = \sum_{k=1}^{K} w_k(x_0)\ Z(x_k) \qquad (4)$$

Where $\acute{Z}(x_0)$ is the estimator function for input x_0, $Z(x_k)$ is the interpretation value of x_k, x_k is the set of attribute values of sample *k*, *K* is the number of samples, w_k is set of weight for each x_k. Then, in the algae estimation dataset, the GTB algorithm is used to create the estimation model. In machine learning, GTB is a kind of ensemble learning that combines various prediction models of decision trees. This algorithm predicts an unknown value of *y* from the input *x* value by using the *F(x)* function [19]. Just like GPR, to form the *F(x)* function model, a dataset with pair input-output $\{(x_1, y_1), \ldots (x_k, y_k)\}$ is required and the algorithm is calculating the specific loss function of $L(y, F(x))$ to evaluate the model. Mathematically, the GTB estimator can be formulated as in Eq. (5).

$$F(x) = \sum_{k=1}^{K} \gamma_k h_k(x) + c \qquad (5)$$

Where $F(x)$ is the estimation model for set of input *x*, *c* is a bias, $\gamma_k h_k(x)$ is multiplication of learning rate and tree model, *K* is the number of samples. As seen in Eq. (4) and Eq. (5), both algorithms have several parameters that need to be tuned to produce a good estimation model. Therefore, an optimization algorithm called grid search is used to find the optimum GPR and GTB parameter that produce the best model performance. Grid search works by trying all the combination numbers in a certain range. Then, it selects the best combination based on the achieved highest performance [20].

In addition, this research also analyzed the comparison between our proposed combination which is the optimized GPR [18, 20] and GTB [19, 20] with another model formed with other machine learning algorithms such as SVR [21], KRR [22], and Lasso [23]. Each algorithm is also tuned with grid search [20] on its parameters. The combination number used for optimization on each machine learning algorithm can be seen in Table 3. The kernel used by the SVR is radial basis function kernel (RBF) [24] so that there are 2 parameters that are optimized namely bias (*c*) and gamma (γ).

2.4 Validation and performance measurement

The most important process that needs to be done after the modeling process is testing the model and measure its performance. In the regression problem, a model is said to have a good performance if the model has a high explained variance (EV) score and low mean absolute error (MAE) [25]. To calculate these scores, a testing or validation mechanism called k-fold cross-validation is needed. This mechanism works by dividing the dataset into k parts and separated between parts for training and parts for testing [26]. In the first iteration, one part is used as a test data and the remaining parts are used as train data. The model is made from the processing of train data and being validated on the test data. Then, a pair of EV and MAE score is calculated. In the next iteration, one another part becomes the test data and the rest becomes the train data. This process is

Table 3. Tuning range for grid search optimization

Algorithm	Parameters	Tuning range
GPR [18]	Alpha (α)	0.001, 0.002, …, 1
GTB [19]	Number of trees (t)	1, 2, …, 100
	Learning rate (lr)	0.005, 0.01, …, 0.1
KRR [22]	Alpha (α)	0.001, 0.002, …, 1
SVR – RBF [21][24]	Gamma (γ)	0.01, 0.02, …, 1
	Bias (c)	0.5, 1.0, …, 5
Lasso [23]	Alpha (α)	0.001, 0.002, …, 1

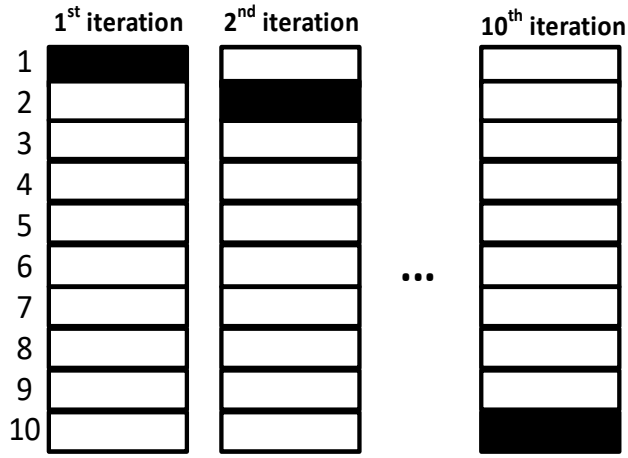

Figure.2 K-fold cross-validation (K = 10)

repeated until k iteration so that each part has been a train and test data.

The k value used in this research is 10 so there will be 10 parts of data to assess how well the performance of the model formed with each algorithm. Illustration of k-fold cross-validation can be seen in Fig. 2 where the black box is the test data and the rest (white boxes) is the train data.

As explained before, to justify how well the performance of a model, the EV and MAE score are calculated during the validation process. The mathematical formula to calculate these scores can be seen in Eq. (6) and Eq. (7) respectively.

$$EV = 1 - \frac{Var|y - y'|}{Var|y|} \tag{6}$$

$$MAE = \frac{\sum_{k=1}^{K} |y - y'|}{K} \tag{7}$$

Where y is the actual value of the output attribute, y' is the predicted value of the output attribute, K is the total samples in the test data, and Var means a varian function. Since k=10 in k-fold cross-validation, there will be 10 scores of EV and MAE for each model. Thus, to calculate the final score of the model performance, all scores are being averaged. All steps in the modeling process can be seen in Fig. 3.

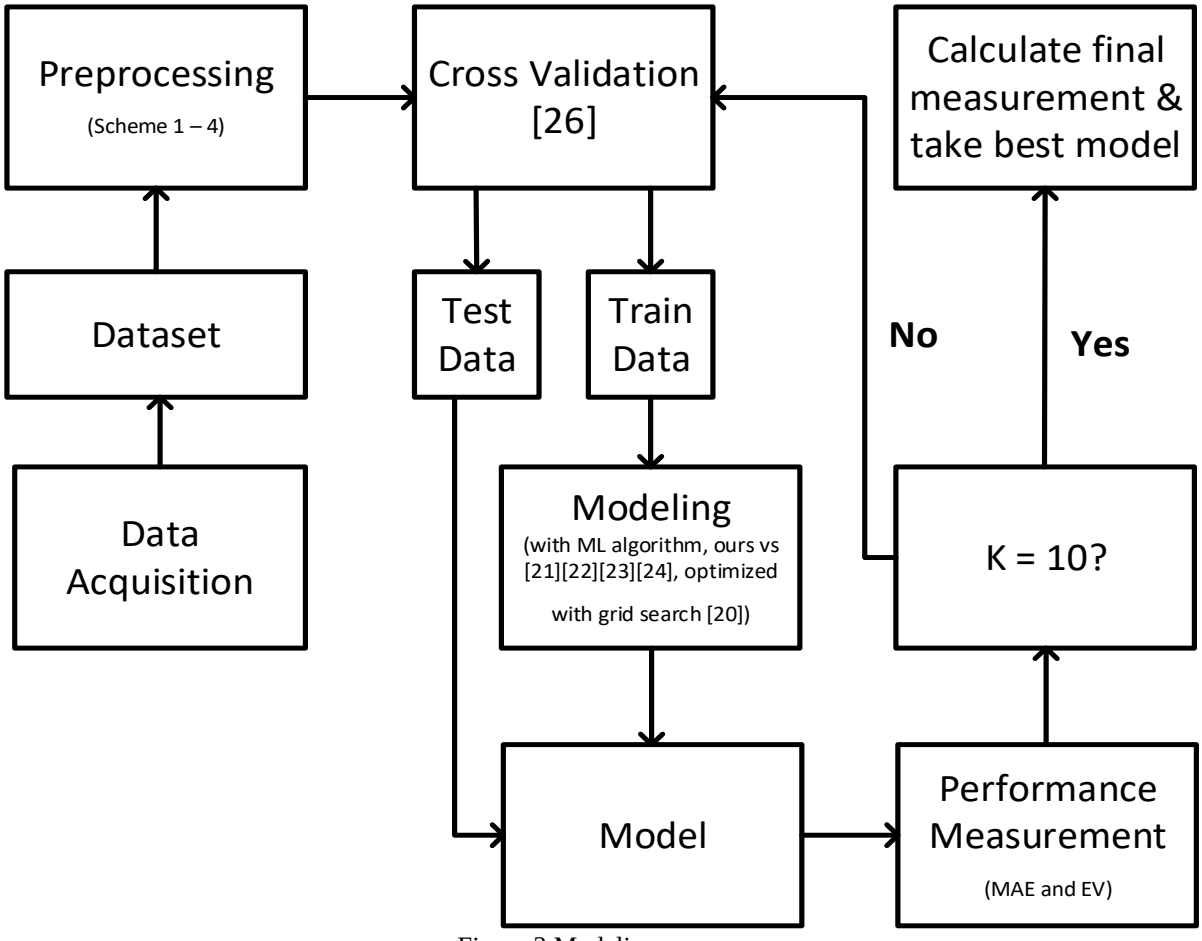

Figure.3 Modeling process

3. Results and discussion

This chapter explains the results of VB ratio and algae estimation modeling using GPR and GTB compared with another machine learning algorithm. The parameters used to measure the performance of the model are EV and MAE. All of the preprocessing schemes are also compared and analyzed. Thus, the total number of models compared in this experiment is 20 for the VB ratio estimation and also 20 for the algae total estimation model.

3.1 MAE comparison

As explained earlier, a model is said to have good performance if it has a small MAE value. MAE shows the level of closeness of the predicted value with the true value. The smaller the MAE, the closer the predicted value to the real value. In accordance with Eq. (7), MAE calculates the absolute value of the difference between all predicted values and the actual values, then calculate its mean. Since this study performs 10-fold cross-validation, there are 10 MAE values calculated and the final MAE is calculated by averaging all MAE values obtained from the validation process. The results of the MAE comparison of the five algorithms along with 4 pre-processing schemes for VB ratio and algae estimation can be seen in Fig. 4 and Fig. 5 respectively. In the VB ratio modeling, the best model is obtained by the GPR algorithm with the third data pre-processing scheme (PCA with 2 principal components and min-max scaling), achieving the smallest MAE of 0.2482.

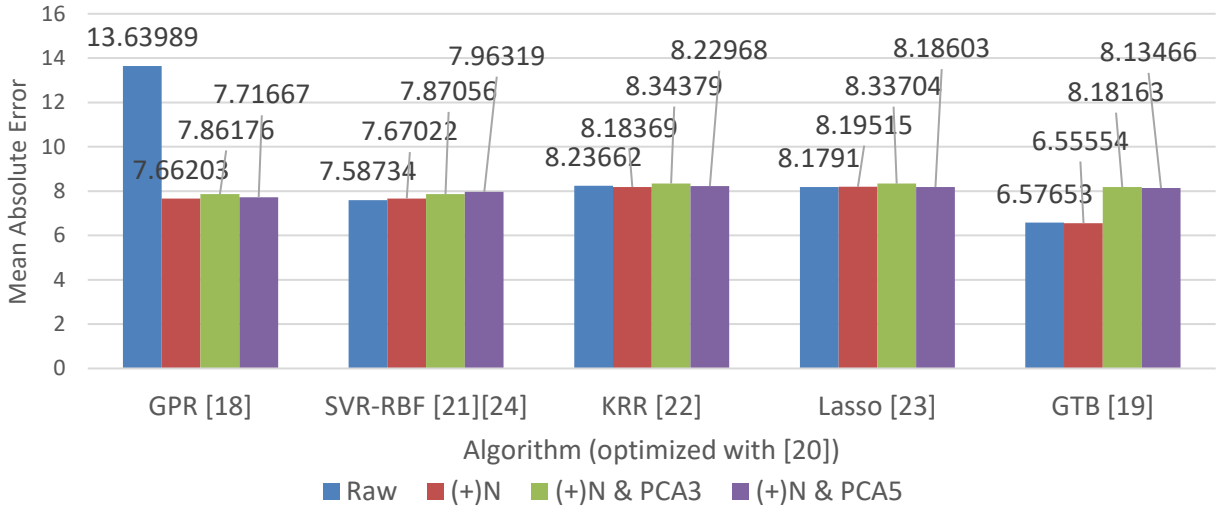

Figure.5 MAE comparison of algae estimation model

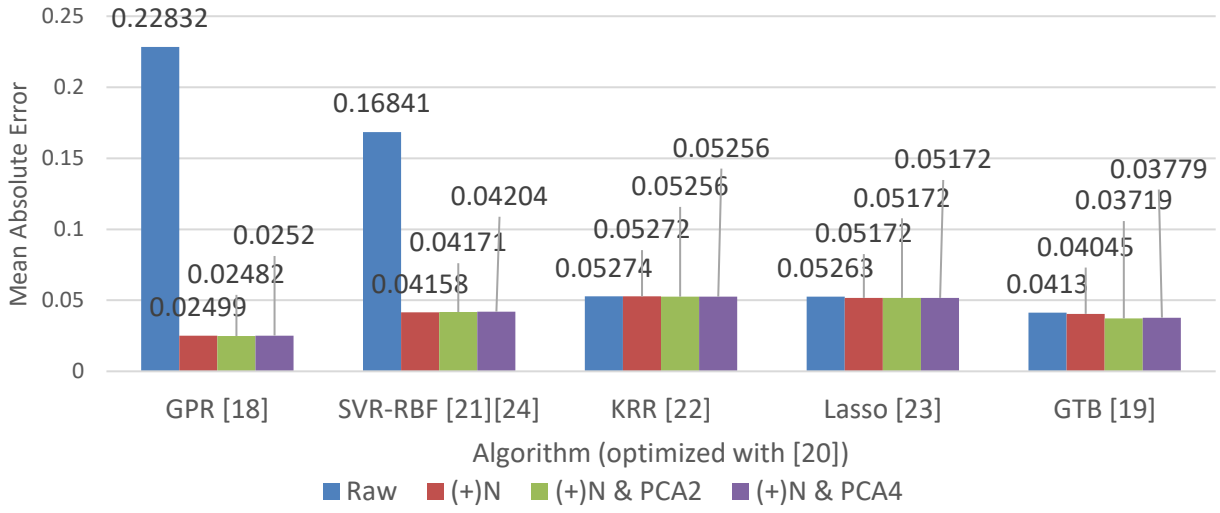

Figure.4 MAE comparison of VB ratio model

The tuned parameter α given by the grid search for GPR optimization is 0.001. Whereas, in the algae estimation modeling, GTB achieved the smallest MAE of 6.55554 with the help of the second pre-processing scheme (min-max scaling only). The optimized parameters for the number of trees and the learning rate of the GTB algorithm are 65 and 0.07 respectively. As shown in Fig. 4 and Fig. 5, it can be concluded that the application of pre-processing just gives a significant influence on the SVR kernel RBF and GPR.

3.2 Explained variance comparison

Besides MAE, a model is said to be good if the variance between the predicted error values is small. This indicates that the model has consistency in estimating a value. If the MAE is analogous to accuracy, then the EV score is can be analogous to precision. From the Eq. (6), if the predicted error variant is smaller, then the subtractor becomes smaller and the EV score becomes higher. Thus, the higher EV score, the better the model. The final EV score is obtained by averaging all EV scores during validation. The EV comparison of VB ratio and algae estimation models can be seen in Fig. 6 and Fig. 7 respectively.

In the VB ratio estimation model, the GPR algorithm succeeded in obtaining the highest EV score of 0.96515 by applying third pre-processing scheme (PCA with 2 principal components and min-max normalization).

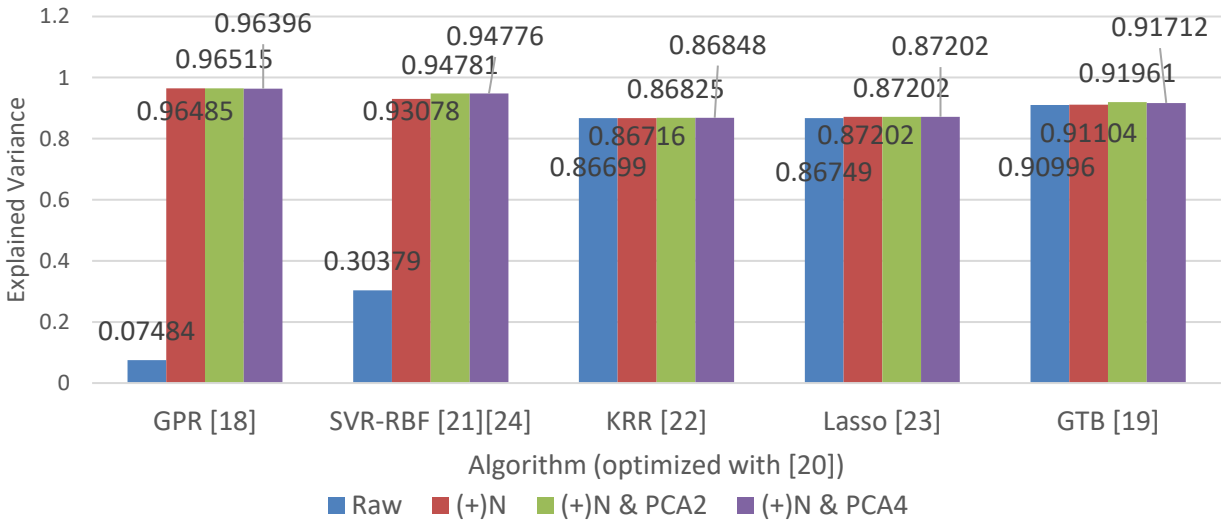

Figure.6 EV comparison of VB ratio model

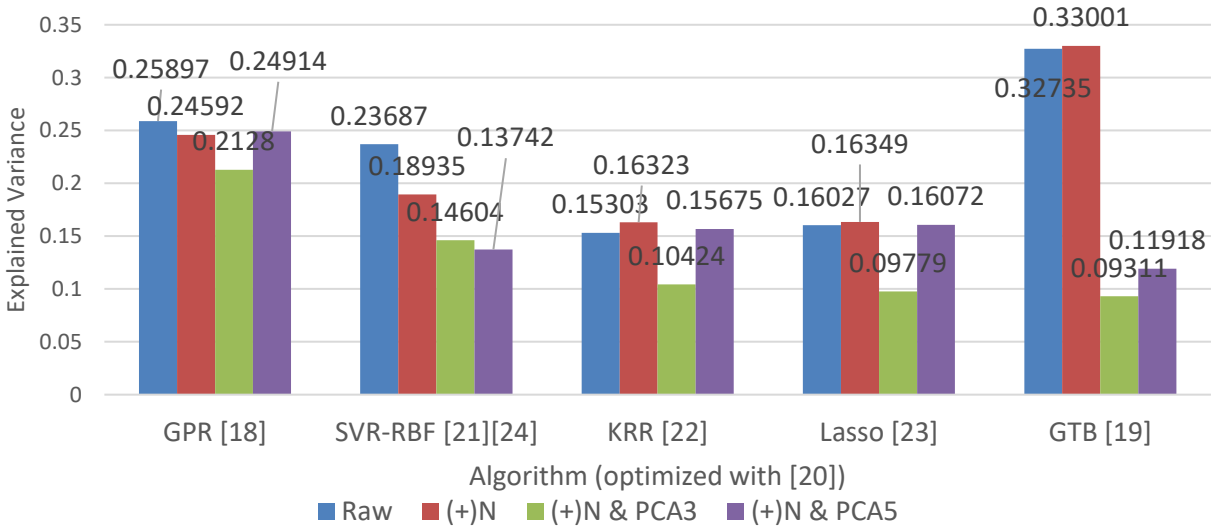

Figure.7 EV comparison of algae estimation model

Table 4. All experimental results

Algorithm and its (parameter)	Preprocessing Scheme	VB ratio			Algae estimation		
		Param. Tuning	MAE	EV	Param. Tuning	MAE	EV
GPR + Grid Search (α)	1	0.669	0.22832	0.07484	0.001	13.63989	0.25897
	2	0.001	0.02499	0.96485	0.191	7.66203	0.24592
	3	0.001	0.02482	0.96515	0.077	7.86176	0.2128
	4	0.001	0.0252	0.96396	0.088	7.71667	0.24914
SVR-RBF + Grid Search (c) and (γ)	1	0.5 and 0.01	0.16841	0.30379	5 and 0.02	7.58734	0.23687
	2	1.5 and 0.8	0.04158	0.93078	5 and 0.94	7.67022	0.18935
	3	5 and 0.51	0.04171	0.94781	5 and 1	7.87056	0.14604
	4	5 and 0.51	0.04204	0.94776	5 and 1	7.96319	0.13742
KRR + Grid Search (α)	1	1	0.05274	0.86699	1	8.23662	0.15303
	2	0.001	0.05272	0.86716	0.086	8.18369	0.16323
	3	0.033	0.05256	0.86825	0.294	8.34379	0.10424
	4	0.049	0.05256	0.86848	0.119	8.22968	0.15675
Lasso + Grid Search (α)	1	0.001	0.05263	0.86749	0.256	8.1791	0.16027
	2	0.001	0.05172	0.87202	0.008	8.19515	0.16349
	3	0.001	0.05172	0.87202	0.025	8.33704	0.09779
	4	0.001	0.05172	0.87202	0.01	8.18603	0.16072
GTB + Grid Search (t) and (lr)	1	53 and 0.09	0.0413	0.90996	50 and 0.095	6.57653	0.32735
	2	55 and 0.09	0.04045	0.91104	65 and 0.07	6.55554	0.33001
	3	86 and 0.075	0.03719	0.91961	21 and 0.09	8.18163	0.09311
	4	78 and 0.095	0.03779	0.91712	15 and 0.1	8.13466	0.11918

The results of the grid search optimization for the α parameter of the GPR algorithm are 0.001. Whereas, in the algae estimation model, the highest EV score of 0.33001 is achieved by the GTB algorithm with the second preprocessing scheme (min-max scaling only). As for the optimum number of trees and the learning rate tuning on GTB is 65 and 0.07 respectively. If we compare the MAE graphs in Fig. 4 and Fig. 5 with the EV graphs in Fig. 6 and Fig. 7, it can be concluded that the smaller the MAE, the higher the EV score. This is in accordance with the Eq. (6) and Eq. (7) about calculating MAE and EV score in the previous explanation. The results of all experiments conducted in this research can be seen in Table 4. If we compare the modeling results of the VB ratio and algae estimation, it can be seen that the difference in MAE and EV scores is quite far away where the VB ratio model is better than the algae estimation model. This can be due to the inconsistent characteristics of the algae estimation dataset where its input attributes are inconsistent in representing the values in the output attribute. This causes a machine learning algorithm having difficulty in making a prediction/estimation.

4. Conclusion

From section 3, it can be concluded that the best model for VB ratio and algae estimation is obtained by using GPR and GTB algorithm respectively combined with pre-processing and optimization algorithm. GPR achieved MAE of 0.02482 and EV score of 0.96515 with the help of third pre-processing scheme where PCA with 2 principal components and min-max normalization is applied. Then, GTB achieved MAE of 6.55554 and EV score of 0.33001 when applying the second pre-processing scheme which is min-max scaling only. Concisely, the pre-processing scheme only gives a significant difference to the GPR algorithm and RBF kernel SVR algorithm. The grid search optimization algorithm also plays an important role in optimum parameters tuning of the machine learning algorithm. From the experimental results, the optimum α for GPR is 0.001 while the number of trees and learning rate for GTB is 65 and 0.07 respectively. In addition, data consistency also influences the modeling process. This is evidenced by the difference performance of the VB ratio model with algae estimation model.

In the future works, another machine learning algorithm such as neural nets, bayesian, k-nearest neighbor and their parameter's tuning can be studied.

References

[1] O. Natan, A. I. Gunawan, and B. S. B. Dewantara, "Design and Implementation of

Embedded Water Quality Control and Monitoring System for Indoor Shrimp Cultivation", *EMITTER International Journal of Engineering Technology*, Vol. 7, No.1, pp.129-150, 2019.

[2] D. E. Alexander, *Encyclopedia of Environmental Science*, Springer, New York, N.Y.1999.

[3] A. A. Malik, Khaeruddin, and Fitriani, "The Effect of Sargassum Extract on Culture Medium to The Growth of Chaetoceros Gracilis", *Aquacultura Indonesiana*, Vol.19, No.1, pp.10-14, 2018.

[4] M. T. Kamble, B. R. Chavan, A. Gabriel, T. Azpeitia, S. V. Medhe, S. Jain, and R. R. Jadhav, "Application of Moringa Oleifera for Development of Sustainable and Biosecure Aquaculture", *Aquacultura Indonesiana*, Vol.15, No.2, pp.64-73, 2014.

[5] X. Jiang, Z. Gao, Q. Zhang, Y. Wang, X. Tian, W. Shang, and F. Xu, "Remote Sensing Methods for Biomass Estimation of Green Algae Attached to Nursery-nets and Raft Rope", In: *Press of Marine Pollution Bulletin*, Vol. 150, 2020.

[6] F. Xu, Z. Gao, X. Jiang, W. Shang, J. Ning, D. Song, and J. Ai, "A UAV and S2A Data-based Estimation of The Initial Biomass of Green Algae in the South Yellow Sea", *Marine Pollution Bulletin*, Vol. 128, No. 1, pp.408-414, 2018.

[7] B. Pan, Z. Shi, Z. An, Z. Jiang, and Y. Ma, "A Novel Spectral-Unmixing-Based Green Algae Area Estimation Method for GOCI Data", *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 10, No. 2, pp.437-449, 2017.

[8] X. Tao, T. Cui, and P. Ren, "Cofactor-Based Efficient Endmember Extraction for Green Algae Area Estimation", *IEEE Geoscience and Remote Sensing Letters*, Vol. 16, No. 6, pp.849-853, 2019.

[9] O. Natan, A. I. Gunawan, and B. S. B. Dewantara, "A New Feature Extraction Algorithm to Extract Differentiate Information and Improve KNN-based Model Accuracy on Aquaculture Dataset", *International Journal on Advanced Science Engineering Information Technology (IJASEIT)*, Vol.9, No.3, pp.999-1007, 2019.

[10] J. Meng, Y. Gao, and Y. Shi, "Support Vector Regression Model for Measuring the Permittivity of Asphalt Concrete", *IEEE Microwave and Wireless Components Letters*, Vol.17, No.12, pp.819-821, 2007.

[11] G. Gupta and N. Rathee, "Performance comparison of Support Vector Regression and Relevance Vector Regression for facial expression recognition", In: *Proc. of International Conference on Soft Computing Techniques and Implementations*, pp.1-6, 2015.

[12] F. Xin, S. Ni, H. Li, and X. Zhou, "General Regression Neural Network and Artificial-Bee-Colony Based General Regression Neural Network Approaches to the Number of End-of-Life Vehicles in China", *IEEE Access*, Vol.6, No.1, pp.19278-19286, 2018.

[13] M. Tahir, T. Badriyah, and I. Syarif, "Classification Algorithms of Maternal Risk Detection for Preeclampsia with Hypertension During Pregnancy using Particle Swarm Optimization", *EMITTER International Journal of Engineering Technology*, Vol.6, No.2, pp.236-253, 2018.

[14] E. Papageorgiou, C. Stylios, and P. Groumpos, "A Combined Fuzzy Cognitive Map and Decision Trees Model for Medical Decision Making", In: *Proc. of International Conference of the IEEE Engineering in Medicine and Biology Society*, pp.6117-6120, 2006.

[15] C. K. Enders, *Applied Missing Data Analysis*, Guilford Press, New York, N.Y. 2010.

[16] B. Liefeng, W. Ling, and J. Licheng, "Feature Scaling for Kernel Fisher Discriminant Analysis Using Leave-one-out Cross-validation", *Neural Computation*, Vol.18, No.4, pp.961–978, 2006.

[17] I. T. Jolliffe, *Principal Component Analysis*, Springer Series in Statistics, New York, N.Y. 2002.

[18] A. Rohani, M. Taki, and M. Abdollahpour, "A novel soft computing model (Gaussian process regression with K-fold cross validation) for daily and monthly solar radiation forecasting (Part: I)", *Renewable Energy*, Vol.115, pp.411-422, 2018.

[19] Y. C. Chang, K. H. Chang, and G. J. Wu, "Application of eXtreme gradient boosting trees

in the construction of credit risk assessment models for financial institutions", *Applied Soft Computing*, Vol.73, pp.914-920, 2018.

[20] C. Marc and B. D. Moor, "Hyperparameter Search in Machine Learning", In: *Proc of Metaheuristics International Conference*, pp.1-5, 2015.

[21] J. P. Karmy and S. Maldonado, "Hierarchical time series forecasting via Support Vector Regression in the European Travel Retail Industry", *Expert Systems with Applications*, Vol. 137, pp.59-73, 2019.

[22] P. Exterkate, P. J. F. Groenen, C. Heij, and D. V. Dijk, "Nonlinear forecasting with many predictors using kernel ridge regression", *International Journal of Forecasting*, Vol.32, No.3, pp.736-753, 2016.

[23] R. Muthukrishnan and R. Rohini, "LASSO: A feature selection technique in predictive modeling for machine learning", In: *Proc. of International Conference on Advances in Computer Applications (ICACA)*, pp.18-20, 2016.

[24] K. Soman, A. Sathiya, and N. Suganthi, "Classification of stress of automobile drivers using Radial Basis Function Kernel Support Vector Machine", *International Conference on Information Communication and Embedded Systems (ICICES)*, pp. 1-5, 2014.

[25] P. I. Good and J. W. Hardin, *Common Errors in Statistics (And How to Avoid Them)*, Hoboken Press, New Jersey, N.J.2009.

[26] T. T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation", *Pattern Recognition*, Vol.48, No.9, pp.2839-2846, 2015.