Proceedings of the 10th International Conference on Fisheries and Aquaculture, Vol. 10, Issue 1, 2023, pp 45-63

ISSN 2386-1282 online
DOI: https://doi.org/10.17501/23861282.2023.10105

KERNEL PCA AND ENSEMBLE LEARNING FOR PREDICTING WATER CHEMISTRY AND MICROBIOLOGY PROPERTIES OF PONDS IN PENAEUS VANNAMEI CULTIVATION

Lukman H*, Syauqy NA and Liris M

JALA TECH Pte Ltd, Indonesia

Abstract: Water quality is one of the important factors that determine shrimp cultivation yields. It determines shrimp growth and survival rate. Hence water quality monitoring is one of the important activities in shrimp farming. Despite its importance, monitoring water quality during shrimp farming can be costly. This research was conducted to develop prediction models that would give farmers insight about water quality of their ponds. The prediction models used temperature, dissolved oxygen, salinity, and pH as input to predict chemical and microbiological properties of the water. The chemical properties included hardness, magnesium, calcium, and total ammonia whereas the microbiological properties included total organic matter and total plankton. The prediction model was built by combining Kernel Principal Component Analysis and machine learning algorithms (Random Forest and Gradient Boosting separately). The method was tested on the data collected from 31 ponds. The results showed that the algorithm can predict biological and chemical conditions of water (Total Organic Matter, Hardness, Calcium, Magnesium) quite well with R2 score higher than 0.8 on most parameters. Further the result also showed that the combination of Kernel PCA (configured with 2 order polynomial kernel) and Gradient Boosting had best prediction accuracy. These findings show that the method can be used as an alternative to laboratory tests. This would help the farmer in monitoring their pond's condition in a faster and less expensive way. This also would help farmers who don't have access to laboratory facilities in monitoring the water quality condition.

Keywords: Water quality prediction, monitoring in aquaculture, machine learning for aquaculture

Introduction

Aquaculture is one of the food sectors with the fastest growth rate. Amongst the various branches of aquaculture, shrimp culture has expanded rapidly across the world because of faster growth rate of shrimps, short culture period, high export value and demand in the market (Rahman et al., 2015)

Indonesia is one of largest shrimp producers in Southeast Asia (FAO, 2020). The shrimp farming industry developed in Indonesia starting in the late 1980s, initially in East Java, then spreading throughout the country. As in other major shrimp farming countries, the presence of bacterial and viral diseases poses a threat to the sustainable development of shrimp farming with the potential for severe economic losses affecting yield and survival rate (Sunarto et al., 2004; Walker et al., 2009; Ali et al., 2018).

Water quality monitoring is one of several methods used to control the risk in shrimp farming. Several studies recommend increasing farmer awareness of the importance of recording consistency and imputation accuracy to ensure reliable modeling of longitudinal data used for improving production

*Corresponding Author's Email: *lukman@jala.tech*

outcomes and mitigating crop failures (Walker et al., 2009; Emilie et al., 2022). This research was conducted to develop prediction models that would give farmers insight into the water quality characteristics of their ponds. The research focused on developing predictive models that predict water hardness, concentration of magnesium calcium, nitrate, carbonate, and bicarbonate.

Table 1: Variables in the Dataset

No.	Parameters	Description	Roles
1.	Pond area	Measured in square meter	Independent Variable
2.	Total seed	Total seed of shrimp	Independent Variable
3.	Total feed usage	Total feed given to the shrimp during 1 cycle cultivation (kg)	Independent Variable
4.	Day of cultivation	The age of cultivation in days	Independent Variable
5.	Temperature	Measured in Celsius in the morning (3 to 9 am) and evening (17 to 21 pm)	Independent Variable
6.	Dissolved Oxygen	Measured in ppm in the morning (3 to 9 am) and evening (17 to 21 pm)	Independent Variable
7.	Salinity	Measured in ppm in the morning (3 to 9 am) and evening (17 to 21 pm)	Independent Variable
8.	pH	Measured in the morning (3 to 9 am) and evening (17 to 21 pm)	Dependent Variable
9.	Total Organic Matter	Measured in ppm with varying frequency of measurements	Dependent Variable
10.	Water hardness	Measured in ppm with varying frequency of measurements	Dependent Variable
11.	Calcium	Measured in ppm with varying frequency of measurements	Dependent Variable
12.	Magnesium	Measured in ppm with varying frequency of measurements	Dependent Variable

MATERIALS AND METHODS

Dataset

The dataset used in this study was collected from several locations in Indonesia. It contains 867 measurements samples from 146 cultivation cycles. The cultivation cycles were done from July 2021 to June 2022 in 138 shrimp ponds. The cultivation cycles were done in different times and with different duration during those periods. There are several ponds with 2 cultivation cycles. The dataset has several significant parameters. The parameters that this dataset has listed in Table 1. Total organic

matter (TOM), water hardness, calcium, and magnesium were used as dependent variables. Meanwhile the other variables were used as predictors.

Data Cleaning

Data cleaning was done to make sure that the data has good quality. In this work the dataset was cleaned through two steps, outlier handling and data imputation.

Outlier handling

The process of identifying outliers was implemented to ensure the validity of all data and to eliminate any anomalous conditions. This study employed a univariate Gaussian distribution to spot outliers in each variable. The probability distribution function (PDF) of the Gaussian distribution, which uses the mean (μ) and standard deviation (π), is depicted in Equation 1, as per Zhang, X. (2011). For each variable, the two parameters were calculated using maximum likelihood estimation (MLE).

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \quad (1)$$

The obtained mean and standard deviation then used to estimate quantiles 5% and 95% of every variable. The obtained quantile values are then used to filter the data. In this research we only use the data if the value is within the range of quantile 5% and 95%.

Data imputation

Data imputation is the process of filling the missing values in a dataset. Missing data create a number of potential challenges for statistical analysis. Fundamentally, missing values call into question the validity of the dataset to represent the observed cases and, ultimately, the sampling frame. From a statistical standpoint, missing values can increase the chances of making Type I and Type II errors, reduce statistical power, and limit the reliability of confidence intervals (Streiner, 2002). In this research we use regression imputation. A better approach to missing value imputation uses regression analysis to predict missing values with conditional means based on relationships among observed variables. This method is also referred to as conditional mean imputation (Schafer & Collins, 2002), and involves two steps: (1) estimating a regression equation and (2) calculating conditional means from the regression equation.

Feature engineering

During feature engineering the data was transformed into new features that can be used in building predictive models. During this phase, seed density and 4-days windowed moving average was calculated.

Stocking density

To measure the density of shrimp in a pond this study used Stocking density. The stocking density plays a role in growth and survival rate of shrimp (Marlina, e., et al, 2020), This parameter calculated as follows:

$$Stocking\ density = \frac{Total\ Seed}{Pond\ area} \tag{2}$$

With:

$$Total\ Seed = Number\ of\ stocked\ seed\ (Tails)$$

$$Pond\ area = Area\ of\ shrimp\ pond\ (m^2)$$

$$Stocking\ density = Number\ of\ stocked\ shrimp\ per\ squared\ area\ (Tails/m^2)$$

4-days windowing

We adopted data windowing from time series analysis. It involves creating a sliding window of fixed or variable size that moves through the data and extracts segments of observations as input variables. The input is a sequence of current and previous time steps. Data windowing can help capture the temporal dependencies and patterns in time series data, as well as reduce the dimensionality and noise of the data. Data windowing can also be used to resample the data at different frequencies, such as hourly, daily, or weekly, depending on the analysis objective and the available data.

Kernelized Principal Component Analysis (KPCA)

Classical PCA algorithm aims at finding a linear subspace of lower dimension than the original space. KPCA is an extension of Principal Component Analysis (PCA). Unlike normal PCA, KPCA achieves non-linear dimensional reduction of data through kernel function. In this research, a polynomial kernel as explained in Shaft-Taylor (2011) was used. Steps of PCA can be found at Ezuwokwe (2019).

Z-score Normalization

In some datasets there are different ranges of values for each attribute. The difference in the range of the value might cause the malfunction of the attribute which has a much smaller value compared to other attributes (Henderi., et al, 2021). Hence transformation toward the dataset such as normalization is needed. Normalization is a way to adjust values measured in different scales to a notationally common scale. Z-score normalization normalize values by using mean (μ) and standard deviation (σ). It can be calculated with following formula (Aldhyani et al., 2020):

$$Z\ Score = \frac{(x - \mu)}{\sigma} \tag{3}$$

Random Forest Regression

Random forest is a group of un-pruned classification or regression trees made from the random selections of samples of the training data. Random features are selected in the induction process based on the selected samples. Prediction is made by averaging the prediction of the ensemble from all of the trees (Ali and Ahmad 2012). The basic steps of random forest algorithm are follows (Xu., et al, 2021):

1. From training set data, K sets of data are generated by bootstrap sampling with put-back. Each dataset is divided into two sampled and un-sampled-data. Sampled data is used during the training phase while un-sampled is used during testing phase, Each data set will generate a decision tree from the training phase.
2. Each decision tree is trained by training data. At each node, m features are randomly selected. The optimal features are selected based on the gini metric.
3. Each generated decision tree is tested using un-sampled data. The prediction error during this phase is used to determine the best decision tree.
4. Determined decision tree models from each dataset are used for prediction.The prediction value is generated by taking the average value of prediction results generated by determined decision tree models.

Gradient Boosting Regression

Gradient boosting is a family of powerful machine-learning techniques that have shown considerable success in a wide range of practical applications.. The main idea of boosting is to add new models to the ensemble sequentially. At each particular iteration, a new weak, base-learner model is trained with respect to the error of the whole ensemble learnt so far. In gradient boosting machines, or simply, GBMs, the learning procedure consecutively fits new models to provide a more accurate estimate of the response variable. The principal idea behind this algorithm is to construct the new base-learners to

be maximally correlated with the negative gradient of the loss function, associated with the whole ensemble (Natekin and Knoll, 2013).

Model error diagnostic

We used Mean Absolute Percentage Error (MAPE) and R2 score as model performance indicators. MAPE can be calculated with the following equation (Myttenaere et al., 2016):

$$MAPE = \frac{100}{N}\sum_{i=i}^{n} \left|\frac{(\hat{y} - y)}{y}\right| \tag{4}$$

R^2 (Coefficient of determination) is a regression score. It is a statistical measure indicating how close the data are to the fitted regression line (Ostasevicius, V., et all, 2022). The value of R2 score is calculated using these equations:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$R2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{4}$$

Results and Discussion

To verify the effectiveness of the proposed method. We conducted an experiment in two phases. In the first phase we conducted correlation analysis to find out how independent variables relate to the dependent variables. Then in the second phase we conducted an experiment to build prediction models for each independent variable separately.

Descriptive statistics

In descriptive statistics we explored statistical properties of the dependent variables. Descriptive statistics is needed to know the "natural" distribution of the data. We conducted descriptive statistics starting from alkalinity and water hardness related variables and then continuing to nitrogen and organic matter related variables.

Table 2 shows statistical properties of alkalinity and water hardness related variables. From the statistics, we can see that the alkalinity, calcium, and bicarbonate tend to have central tendency. It means that those variables tend to be symmetric. As for hardness and carbonate, those variables tend to be more skewed.

Table 2: Statistical properties of alkalinity and water hardness related variables

Statistics	**Alkalinity**	**Hardness**	**Calcium**	**Magnesium**	**Carbonate**	**Bicarbonate**
Mean	127.16	4649.08	587.95	1760.96	6.45	115.65
Std	26.90	1334.92	332.51	1318.15	9.15	26.12
Min	82.00	796.54	158.00	479.00	0.00	60.00
Q25%	105.00	3548.84	255.50	795.00	0.00	95.00
Q50%	128.00	5000.00	597.00	1110.00	0.00	116.00
Q75%	144.00	5749.61	880.00	2900.00	12.00	135.00
Max	202.00	6500.00	1200.00	4600.00	37.00	178.00

Table 3 displays the statistical characteristics of variables related to alkalinity and water hardness. The statistics reveal that alkalinity, calcium, and bicarbonate tend to have central tendency, meaning that these variables tend to be symmetric. On the other hand, hardness and carbonate tend to be more skewed.

Table 3: Statistical properties of nitrogen and organic matter related variables

Statistics	Ammonia	Nitrate	Nitrite	TOM	Total plankton
Mean	0.03	3.52	0.22	75.74	326394.99
Std	0.03	4.26	0.52	19.44	457479.91
Min	0.00	0.00	0.00	34.81	0.00
Q25%	0.01	1.00	0.02	59.00	0.00
Q50%	0.02	3.00	0.05	75.84	140000.00
Q75%	0.03	3.00	0.13	91.01	493750.00
Max	0.20	51.00	3.00	139.00	2060000.00

Correlations

Before modeling water quality parameters, correlation analysis was conducted to find out whether independent variables that were used have correlation with dependent variables or not. Figure 1. show correlation between independent and dependent variables. Correlation analysis showed that most dependent variables have weak correlation with independent variables. Those weak correlation

Prediction model

To validate the model, the dataset was split into two subsets: 70% for training and 30% for testing. The performance of Kernel PCA + GB model and Kernel PCA + RF Model with various hyperparameter configurations was compared. The models were tuned with KFold cross-validation on the training dataset. This was done for each of the predicted variables (TOM, water hardness, calcium, and magnesium).

To validate the model, the dataset was split into two subsets: 70% for training and 30% for testing. The performance of Kernel PCA + GB model and Kernel PCA + RF Model with various hyperparameter configurations was compared. The models were tuned with KFold cross-validation on the training dataset. This was done for each of the predicted variables (TOM, water hardness, calcium, and magnesium).

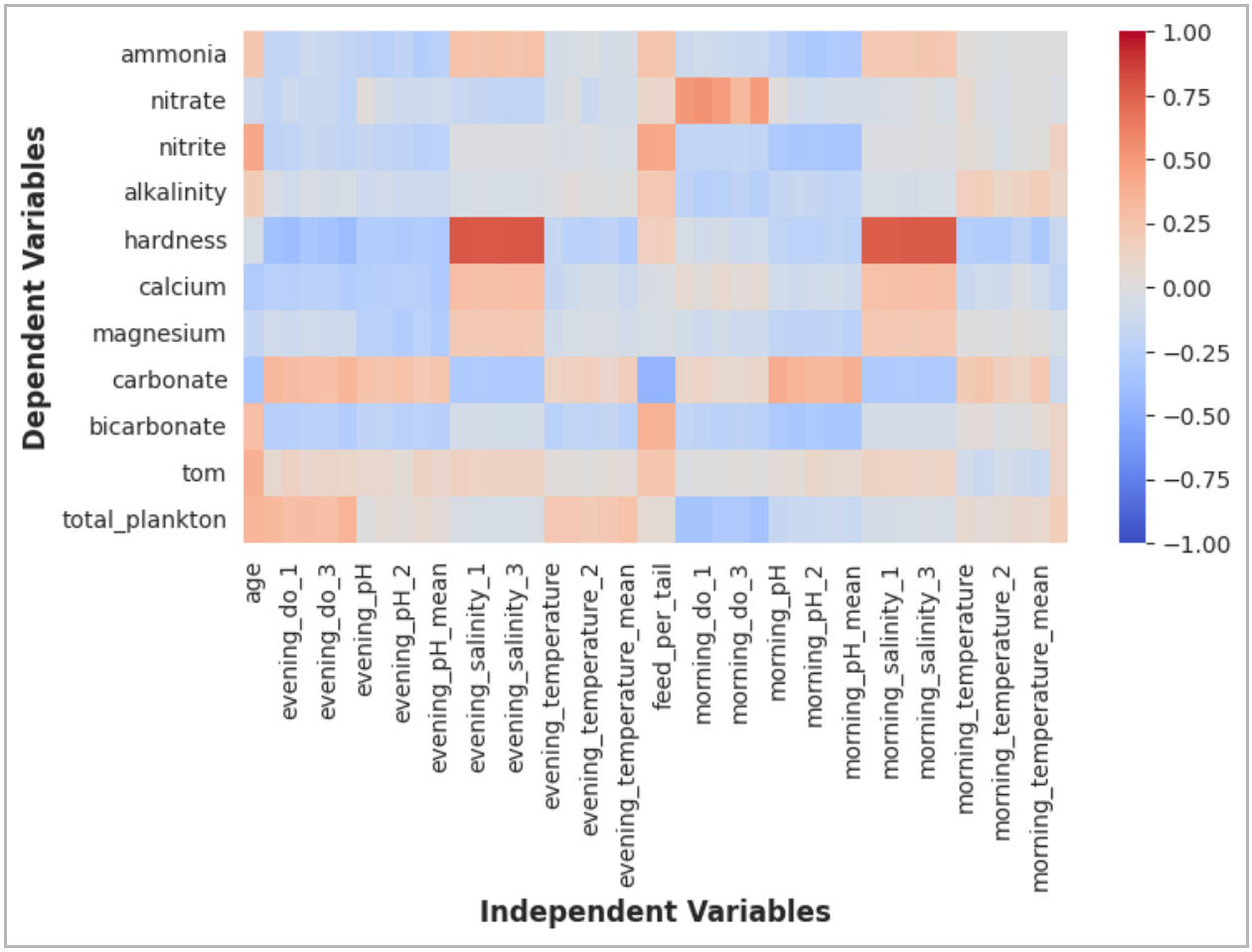

Figure 1: Correlation map between independent variables and dependent variables

Hardness

Figure 2 illustrates the performance of model predictions in performing hardness prediction. It is comparing predictions from KPCA+RF and KPCA+GB models. Error value displayed in the left plot calculated by subtracting actual values from the dataset with predicted values. The right plot in futures illustrates the relationship between actual and predicted values.

As shown in the figures, there is a good alignment between predicted values and actual values. The KPCA+GB model has better accuracy in predicting hardness than the KPCA+RF model. It is shown by lower MAPE values of the models. The KPCA+GB model also has better precision that is shown by lower standard deviation of prediction error. As comparison, R^2 Score was also used as performance metrics. Table 3 and Table 4 listed R2 scores of GB and RF Model respectively with different configurations. KPCA+GB models have better performance with the R^2 score of the best model reaching 0.912 during validation phase whereas KPCA+RF model only reached 0.87 during validation phase.

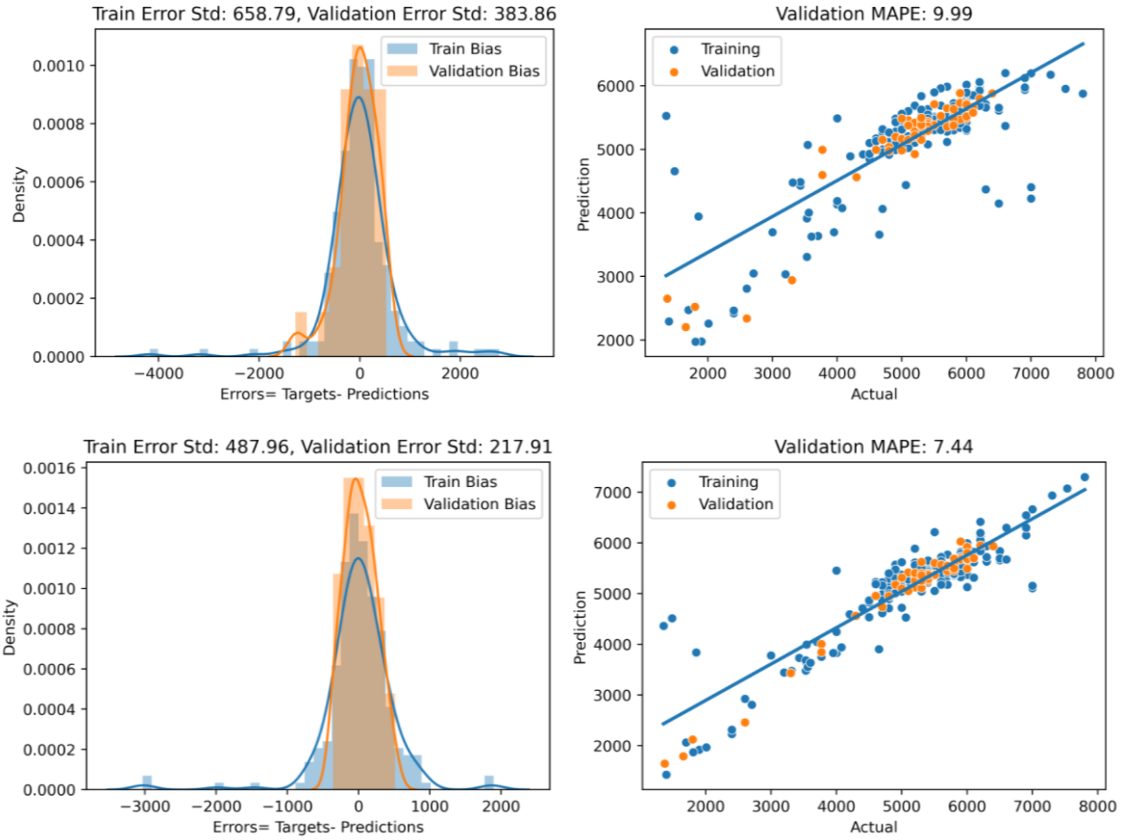

Figure 2. Model performance of (Top) Kernel PCA + RF Model and (Bottom) Kernel PCA + GB Model in Predicting Water Hardness

Table 3: Results of Kernelized GB Model Hyperparameter Tuning with Water Hardness as Predicted Variable

Kernel PCA + GB Hyperparameter			**R^2 Score Validation Phase**
GB - Min Samples of Leaf	**Kernel PCA - Polynomial Degree**	**Kernel PCA - N Component**	
9	1	37	0.912
9	3	37	0.902
9	5	37	0.890
7	1	37	0.874
7	5	37	0.834

Table 4: Results of Kernelized RF Model Hyperparameter Tuning with Water Hardness as Predicted Variable

Kernel PCA + RF Hyperparameter			R^2 Score Validation Phase
RF - Min Samples of Leaf	Kernel PCA - Polynomial Degree	Kernel PCA - N Component	
4	3	35	0.87
4	5	37	0.86
4	1	37	0.86
4	1	35	0.86
4	3	37	0.86

To be more firm with the results, we conducted cross validation with 10 different data splits towards best configuration. Figure 3 shows the results of cross validation of KPCA+RF and KPCA+GB model. From the experiment, we get a stable R^2 score with values higher than 0.9 for KPCA+GB and 0.85 for KPCA+RF model. These results mean that the accuracy is not dependent on how the data splitted.

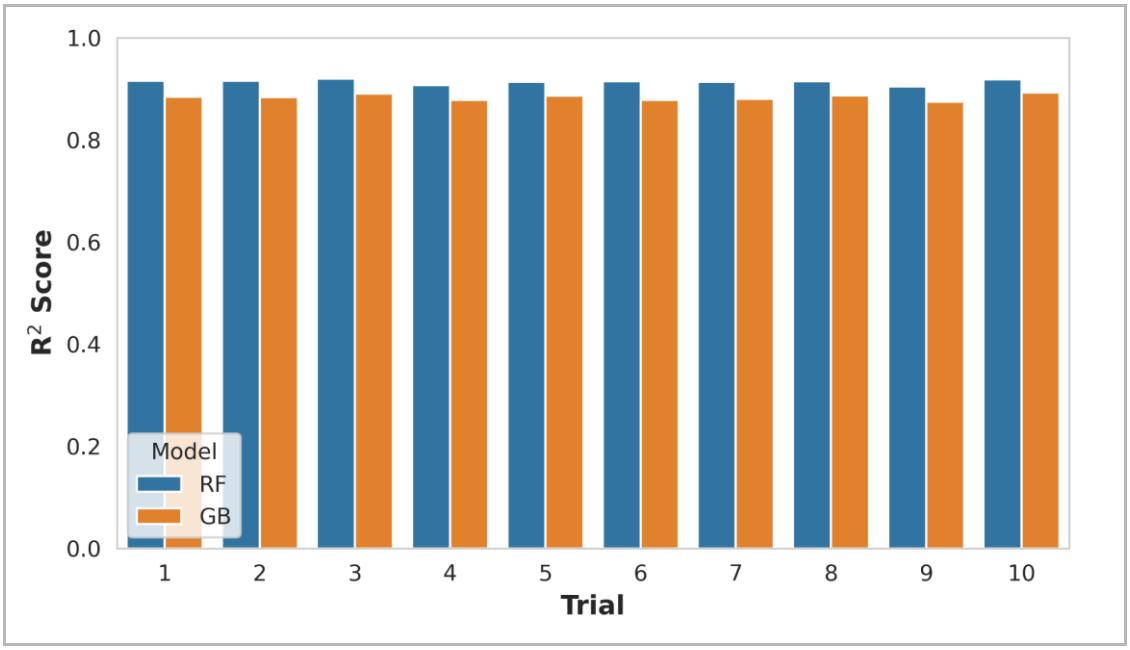

Figure 3: Results of hardness predictive model cross validation on 10 different data splits

TOM

The performance of the models in predicting TOM is illustrated in Figure 4. The KPCA+GB model has better performance in predicting TOM similar to hardness prediction. The KPCA+RF model has lower MAPE with an average value of 8.02% while the RF model has higher error with an average of 10.07%. Hyperparameter tuning was also conducted for TOM. The results were shown in Table 5 (KPCA+GB Model) and Table 6 (KPCA+RF Model).

The KPCA+RF model showed the best performance in predicting TOM after hyperparameter tuning. The R^2 value of the best model was 0.849 on the test set. On the other hand, the KPCA+GB Model had lower performance with an R^2 value of 0.790 on the test set. Table 5 and 6 showed results of hyperparameter tuning of both models for TOM Prediction with best R^2 score. We get the best R^2 score when we use configuration with 9 min samples of leaf, one degree polynomial kernel, and with the number of component 37.

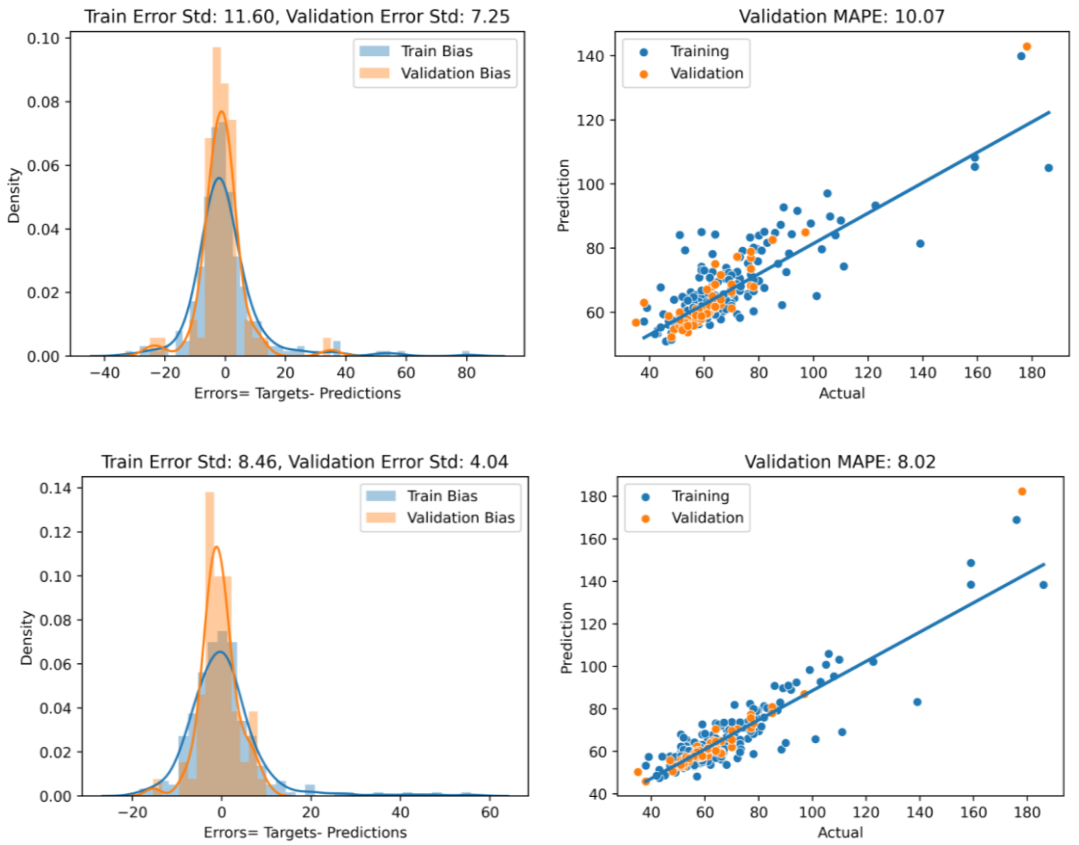

Figure 4: Model performance of (Top) Kernel PCA + GB Model and (Bottom) Kernel PCA + RF Model in Predicting Total Organic Matter

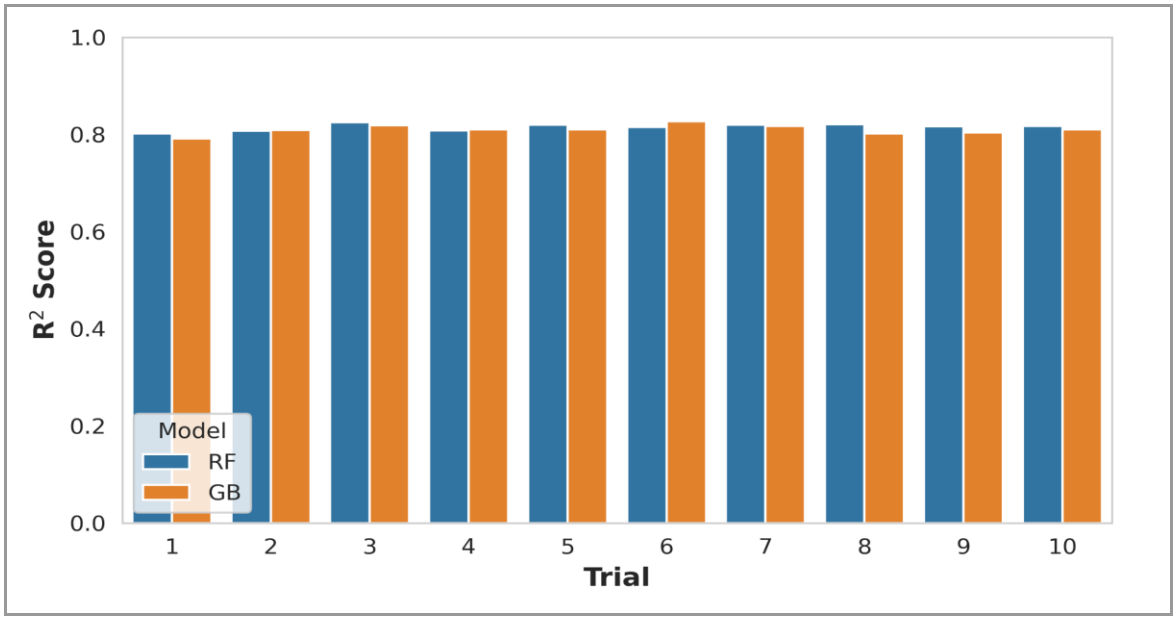

Figure 5: Results of TOM predictive model cross validation on 10 different data split

Table 5: Results of Kernelized RF Model Hyperparameter Tuning with Total Organic Matter as Predicted Variable

Kernel PCA + GB Hyperparameter			R2 Score
RF - Min Samples of Leaf	Kernel PCA - Polynomial Degree	Kernel PCA - N Component	Validation Set
9	1	37	0.790
9	3	37	0.783
9	5	37	0.777
7	1	37	0.744
7	5	37	0.732

Magnesium

Model performance in predicting magnesium illustrated in Figure 3. The figure shows that the GB model has better performance. The GB model has lower MAPE with value of 8.02.The GB model also has better precision indicated by lower error standard deviation for both during training and validation phase.

Table 6: Results of Kernelized GB Model Hyperparameter Tuning with Total Organic Matter as Predicted Variable

Kernel PCA + RF Hyperparameter			R2 Score
GB - Min Samples Of Leaf	**Kernel PCA - Polynomial Degree**	**Kernel PCA - N Component**	**Validation Set**
5	3	37	0.849
5	1	37	0.848
5	5	37	0.846
5	3	33	0.842
5	5	33	0.842

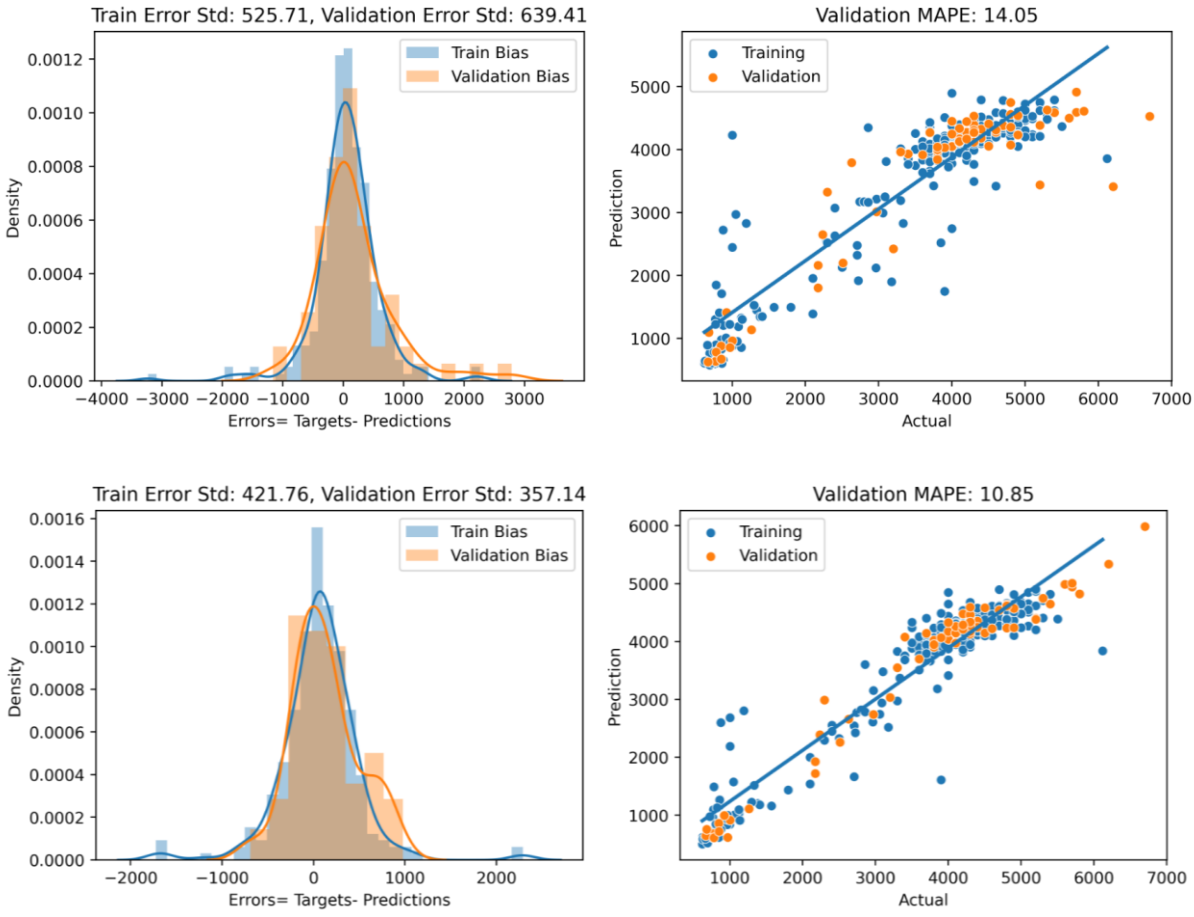

Figure 3: Model performance of (Top) Kernel PCA + RF Model and (Bottom) Kernel PCA + GB Model in Predicting Magnesium

Table 7: Results of Kernelized GB Model Hyperparameter Tuning with Magnesium as Predicted Variable

Kernel PCA + GB Hyperparameter			R2 Score	
GB - Min Samples of Leaf	**Kernel PCA - Polynomial Degree**	**Kernel PCA - N Component**	**Train Set**	**Validation Set**
5	5	33	0.979	0.928
5	3	37	0.981	0.928
5	1	37	0.981	0.928
5	1	33	0.979	0.928
9	5	37	0.977	0.928
9	3	37	0.977	0.928
5	5	37	0.981	0.928
9	1	37	0.977	0.928
5	3	33	0.979	0.928

Table 8: Results of Kernelized RF Model Hyperparameter Tuning with Magnesium as Predicted Variable

Kernel PCA + RF Hyperparameter			R2 Score	
RF - Min Samples Of Leaf	**Kernel PCA - Polynomial Degree**	**Kernel PCA - N Component**	**Train Set**	**Validation Set**
5	5	35	0.913	0.873
5	3	37	0.912	0.872
5	1	37	0.912	0.871
5	1	27	0.911	0.871
9	5	35	0.911	0.871
9	3	21	0.908	0.871
5	5	27	0.910	0.870
9	1	37	0.911	0.870
5	3	21	0.907	0.870

Table 7 and *Table 8* depicts the results of hyperparameter tuning of GB and RF Model for Magnesium prediction of models with best performance. The best performance was reached by the GB model with R2 score 0.928 during validation phase. Meanwhile, the RF model only managed to reach 0.873 of R2 score. Results of hyperparameter tuning also showed that GB models tend to have a higher degree of overfitting that was indicated by a bigger difference of performance between training phase and validation phase.

Calcium

Performance of the models in predicting calcium illustrated in Figure 4**.** The performance of the models in calcium prediction was not as high as previous water quality parameters. The GB Model has a MAPE score of 11.64% in the validation test while the RF Model only reached 15.17%. The GB model is also better in terms of precision where the GB model have lower standard deviation of error than the RF model. Similar to previous variables, the GB model had better performance in terms of accuracy and precision.

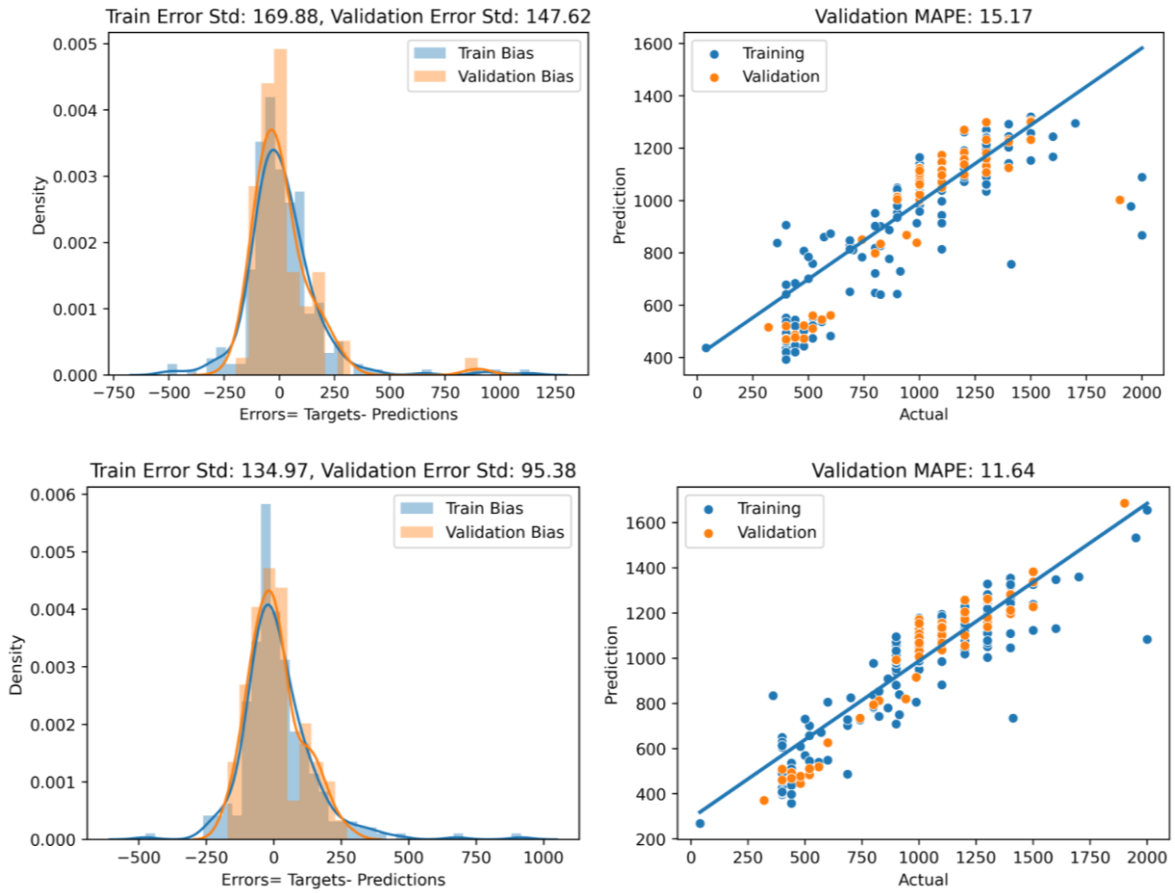

Figure 4: Model performance of (Top) Kernel PCA + RF Model and (Bottom) Kernel PCA + GB Model in Predicting Calcium

Table 9: Results of Kernelized GB Model Hyperparameter Tuning with Calcium as Predicted Variable

Kernel PCA + GB Hyperparameter			R2 Score	
GB - Min Samples of Leaf	**Kernel PCA - Polynomial Degree**	**Kernel PCA - N Component**	**Train Set**	**Validation Set**
7	1	37	0.964	0.865
7	1	33	0.960	0.865
5	1	37	0.970	0.864
7	3	33	0.960	0.864
7	5	33	0.960	0.864
5	5	33	0.965	0.863
5	1	33	0.965	0.863
5	3	33	0.965	0.863
5	5	37	0.970	0.863

Table 10: Results of Kernelized RF Model Hyperparameter Tuning with Calcium as Predicted Variable

Kernel PCA + RF Hyperparameter			R2 Score	
RF - Min Samples Of Leaf	**Kernel PCA - Polynomial Degree**	**Kernel PCA - N Component**	**Train Set**	**Validation Set**
4	1	35	0.825	0.747
4	5	37	0.823	0.746
4	3	35	0.818	0.746
4	3	37	0.824	0.744
4	1	37	0.819	0.742
4	5	35	0.818	0.740
4	5	35	0.825	0.739
4	5	27	0.817	0.738
4	3	27	0.812	0.738

Table 9 and 10 shows results of hyperparameter tuning from models with best performance. Durin hyperparameter tuning, the RF model managed to reach 0.825 in R2 score during training. Buth during the validation phase the performance dropped to 0.74. The difference between the training and validation phase indicates that there are overfitting in the model. The GB model managed to reach a higher score during training score than the RF model with R2 score up to 0.967. But similar to the RF model, the performance dropped on the tes set to 0.863.

Conclusion

The study collected data from 31 ponds that used the JALA platform. The research found that the microbiology and chemistry of water, such as total organic matter (TOM), water hardness, magnesium and calcium, can be predicted with physical properties of water and seed density of the shrimp towards pond area. Furthermore, the Gradient Boosting model combined with kernel function performed better than the Random Forest model with kernel function. The best model managed to get an R2 score higher than 0.85 during the validation test.

References

Aldhyani .H.H., Al-Yaari M., Alkahtani H., Maashi M. 2020. Water Quality Prediction Using Artificial Intelligence Algorithms. Water Quality Prediction Using Artificial Intelligence Algorithms. 2020

Ali, J., Ahmad, N.(2012). Random Forest and Decision Trees. International Journal of Computer Science Issues. 9.(3).

Ali, H., Rahman, M. M., Rico, A., Jaman, A., Basak, S. K., Islam, M. M., Khan, N., Keus, H. J., & Mohan, C. V. (2018). An assessment of health management practices and occupational health hazards in tiger shrimp (*Panaeus monodon*) and freshwater prawn (*Macrobrachium rosenbergii*) aqualculture in Bangladesh. *Veterinary and Animal Science.* 5. 10-19.

Arnaud de Myttenaere, Boris Golden, Bénédicte Le Grand, Fabrice Rossi.(2016).Mean Absolute Percentage Error for regression models. Neurocomputing. 192. 38-48

Ezukwoke K and Zareian S (2019) Kernel methods for principal component analysis (PCA): a comparative study of classical and kernel PCA. doi:10.13140/RG.2.2.17763.09760.

John Shawt-Taylor, Nello Cristianini. 2011. Kernel Methods for Pattern Analysis. Cambridge University Press,ISBN:9780511809682, 47-83

Marlina Eulis, Hartono Puji Dwi, Panjaitan Imelda. 2020. Optimal Stocking Density of Vannamei Shrimp Litopenaeus Vannamei at Low Salinity Using Spherical Tarpaulin Pond. Advances in Social Science Education and Humanities Research. 298.

Natekin, A., Knoll, A.(2013). Gradient Boosting Machines, a tutorial. Frotiers in Neurotobitcs. 7 (21).

Ostasevicius V, Paleviciute I, Paulauskaite-Taraseviciene A, Jurenas V, Eidukynas D, Kizauskiene L. Comparative Analysis of Machine Learning Methods for Predicting Robotized Incremental Metal Sheet Forming Force. Sensors (Basel). 2021 Dec 21;22(1):18. doi: 10.3390/s22010018. PMID: 35009560; PMCID: PMC8747513.

Vinod Kothari, Suman Vij, SuneshKumar Sharma, Neha Gupta. Correlation of various water quality parameters and water quality index of districts of Uttarakhand. Environmental and Sustainability Indicators. 9. 100093

Walker, P. J., & Mohan, C. V. (2009). Viral disease emergence in shrimp aquaculture: origins, impact and the effectiveness of health management strategies. *Reviews in Aquaculture*, 1, 125-154.

Xu, J.; Xu, Z.; Kuang, J.; Lin, C.; Xiao, L.; Huang, X.; Zhang, Y. 2021. An Alternative to Laboratory Testing: Random Forest-Based Water Quality Prediction Framework for Inland and Nearshore Water Bodies. Water 2021. 13. 3262. https://doi.org/10.3390/ w13223262

Zhang, X. (2011). Gaussian Distribution. In: Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_323