Design of Predictive Model in Classifying Turbidity Using Data Mining Techniques

Jonalyn G. Ebron[*1] Andre Gabriel V. Alvarez[1], and Miguel Mababangloob[1]

[1]College of Computer and Information Science, Malayan Colleges Laguna, Pulo Diezmo, Cabuyao City, Philippines
*Corresponding author, E-mail: jgebron@mcl.edu.ph

Abstract

The study's objective was to develop a predictive model that classifies the turbidity of water using data mining techniques and aimed to help visualize data and predict the classification of the lake's water turbidity, whether it is good or bad. The parameters utilized in the study were Conductivity, Dissolved Oxygen (DO), pH, Total Suspended Solid (TSS), Total Coliform, and Temperature. Artificial Neural Network (ANN), Support Vector Machine (SVM), and k-Nearest Neighbor (KNN) are the data mining techniques used to create the models. The model's effectiveness tests for accuracy, precision, and recall. Correlation-based feature selection describes the linear relationship between different parameters and a model. The highest correlation was obtained between TSS and Turbidity among the attributes, while the temperature was the lowest. The study used three different combinations of parameters. The researchers found that the class count in the data affects the accuracy provided by the model. The less the count of one part of the binary classifier present in the data, the more likely the accuracy will be closer to one. The training of data was through the capabilities of Python. Laravel web framework used to develop the web-based application in PHP language. Furthermore, the results of high-quality development data are a foundation for meaningful insights to protect health and avoid water pollution in developing countries.

Keywords: *Data Mining, Machine Learning, Predictive modeling, Sediments, Turbidity*

1. Introduction

The importance of water quality regarding human health and food supplies has become widely recognized in recent years. To some extent, the importance of water quality in the maintenance of health and aquatic populations has gained appreciation. However, the influence of water quality on sediments, and the use of sediments as environmental factors are not widely understood, and only limited use of sediment data in most environmental quality studies (UNESCO-WHO, Publication Studies, and Reports in Hydrology, no 23) Water polluted with sediment becomes cloudy, and sediment increases the cost of treating drinking water, resulting in odor and taste problems. Anything that makes the water cloudy will increase turbidity. Turbidity is a measurement of how cloudy the water is in a lake. NTU or Nephelometric Turbidity Unit is the unit used to describe turbidity. The more turbid the water is, the lesser the quality of the water. This allowed us to classify how the water can be used based on the index.

There were previous studies that utilized machine learning for the prediction of sediments. Melesse et al. (2011) conducted a study to predict suspended sediment loads in 3 major river rivers: Mississippi, Missouri, and the Rio Grande. Melesse et al. (2011) concluded that the Artificial Neural Network (ANN) approach reduces the frequency of expensive operations for sediment measurements when hydrological data are available. Iglesias et al. (2015), Namu et al. (2017), and Al-Baidhani et al. (2017) predicted turbidity using the ANN. The most common training algorithm utilized in the ANN literature is called backpropagation. Baseri, (2010) stated in his study that the input and output layer are connected using weights adapted by backpropagation techniques that utilized a gradient search to reduce the mean square between the output pattern of the network and the desired output pattern. A study conducted by Kumar et al. (2015) used six different machine learning techniques such as artificial neural networks (ANN), radial basis function neural networks (RBFNN), least-square support vector regression (LS-SVR), multi-linear regression (MLR), and decision tree models such as Classification and Regression Tree (CART) and the M5 model tree was used to simulate and predict the daily suspended sediments using the hydro-

meteorological variables at Kopili River basin in India. The inputs for computing the turbidity can be temperature, pH, and conductivity based on the study conducted by Al-Baidhani and Alameedee (2017). However, Al-Baidhani and Alameedee (2017) also stated that parameters that are suspected to have a direct influence on the effluence feature of the water can be used, such as suspended solids, flow rate, and total coliform. Many studies used different techniques and used different parameters to yield different results. Moreover, the output of the technique varies with the parameters to be used.

However, the most of studies used the following input parameters: pH, dissolved oxygen, total coliform, temperature, and total suspended solids. They recommended exploring different combinations of input parameters for better prediction. Experimenting with using different machine-learning techniques for classification is suggested to continue the study in finding the best model to predict the turbidity in the lake. The addition of other attributes might also contribute to the performance of the machine learning techniques and might have a higher correlation with other attributes.

In the Philippines, Laguna Lake is the largest lake in the Philippines and is one of the largest in Southeast Asia. The lake uses irrigation and aquaculture, energy generation, cooling water in nearby industries, raw water for domestic use in nearby houses, a retention basin for flood control, and a bowl for treated and untreated waste (Jaraula et al., 2014). Cinco (2017) reported that Laguna Lake is dying. Domestic waste caused by residents surrounding the lake is the main contributor to pollution that hastened the death of the lake. However, several toxic pollutants affect the water quality in the west bay, making it fall below its standard (Cinco, 2018)

Laguna Lake Development Authority (LLDA) is responsible for monitoring, preservation, development, and sustainability. According to a 4-year monitoring period from 2009 to 2012, the annual average turbidities in the lake varied from 12 NTU (Nephelometric Turbidity Unit) to 55 NTU. The three tributary river stations that indicated very high annual average turbidity values of more than 100 NTU from 2009 to 2012 were Stn. 11 (San Juan River) in 2009 at 202 NTU, Stn. 24 (Morong River-Downstream) in 2011 at 127 NTU, and Stn. 1 (Marikina River) at 122 NTU in 2011 and 112 NTU in 2012. Stn. 17 NTU (Pangil River Upstream) yielded the lowest annual average turbidity among the tributary river stations at 4 NTU in 2012.

This study aims to design a predictive model that utilizes the water quality data in determining the attributes and parameters as an input in developing a web-based system for classifying sediments. The research ranked which key parameters measure the goodness of fit in the model using the attributes and parameters. The study integrates selected modeling techniques for modification, analysis, and visualization. The use of data visualization would help the data handling and analysis for the user, especially when the user is looking for trends and history over time. The system could be accessed on a web-based application so that different stations can use the web-based system for their monitoring. The study used the Artificial Neural Network, Support Vector Machine, and K-Nearest Neighbor in creating the predictive model.

The study was limited to the provided datasets from the LLDA, and the study did not focus on collecting new data. The study focused on classifying turbidity and analytical techniques to understand the relationships between water quality parameters and sediments. This study emphasizes the value of water quality sediment-related information monitoring to help make decisions about the information from the existing database, including analytical requirements, identifying anomalies, establishing references, and identifying time changes by showing trends.

2. Objectives

The general objective of this research is to provide a predictive model that utilizes the water quality data in determining the attributes and parameters used as input in designing a web-based system for classifying sediments. The study integrates data mining techniques for prediction, analysis, and visualization. Specifically, it aimed.

1) To design a predictive model using a data-mining technique that works by analyzing historical and generating a model to predict the turbidity of whether it is good or bad.

2) To develop a web-based application in the form of a dashboard that will integrate and presents graphs, predictive model, and results.
3) To test the model's accuracy, precision, and recall effectiveness.

3. Materials and Methods

The study used the developmental-descriptive research design and utilized quantitative research. The research conducted interviews and surveys of the target population. The sampling technique used was complete enumeration. Since the study was about the lake, the nine significant stations of Laguna Lake will utilize for this study. The nine influential stations are Central West Bay, East Bay, Central Bay, Northern West Bay, South Bay, San Pedro, and Sta. Rosa, Sanctuary, and Pagsanjan. The data given by the LLDA was available for all the nine significant stations from the year 2009 to 2016. The study had 432 records of data from the LLDA. The methodology in the study follows the conceptual framework of the study, which has three (3) core parts: Data (Input), Process (Data Mining Model), and Output (Predicted classification), as shown in Figure 1 below.

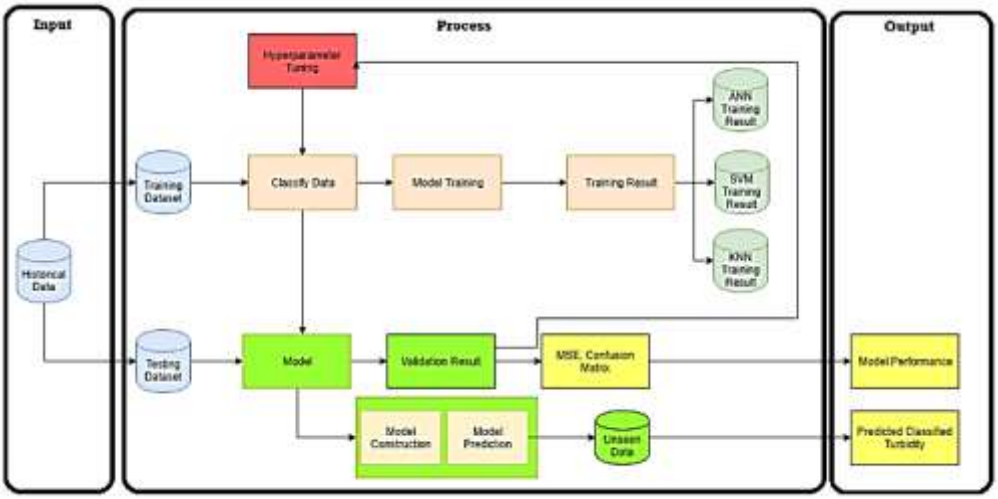

Figure 1 Conceptual Framework of the study

A. Input

A total of 432 records data from 2009 to 2016 were used for the data collection and handling with approval from L DA. The following six (6) parameters in this study are pH, DO (DO), Conductivity, Total Coliform, Temperature, and TSS, as well as the year ranging from 2009-to 2016, and months ranging from January to December from the stations, I, II, IV, V, VII, VIII, XV, XVI, and XVII.

The pre-processing of historical data includes data cleaning, transformation, and database modeling. In the data cleaning process, missing values, outliers, data standardization, and data splitting. Table 1 shows the sample datasets format with *missing values* in the raw datasets using the averaging technique to replace the missing values using Python Jupiter.

Proceedings of RSU International Research Conference (2022)

Table 1 Sample datasets recorded from 2009 to – 2016

	Year	Month	Conductivity	DO	pH	Total Coliforms	TSS	Turbidity	Station
1	2009	April	542	8.5	8.7	110	62	43	XVIII
2	2009	May	654	10.0	9.4	51	30	23	XVIII
3	2010	June	647	9.1	9.4	26	35		XVIII
4	2011	July	509	9.4	9.6	38		10	XVIII
5	2012	August	958	6.7	8.9	99	48		XVIII
6	2013	September		9.4	9.4	194	20		XVIII
7	2014	October		9.3	9.0	76	24		XVIII
…	2015	November		6.6		110	33		XVIII
432	2016	December		7.6	8.4	57	29		XVIII

The *outliers* were identified and then removed. The data's first quartile, third quartile, and interquartile range determine the outliers. The formula in equations (1) and (2) is used to find the lower and upper limits. Data would be an outlier if the data value were less than the lower limit or greater than the upper limit. The outliers detected were removed from the dataset, making the data blank

$$LowerLimit = Q_1 - 1.5 \times IQR \qquad (1)$$

$$UpperLimit = Q_2 + 1.5 \times IQR \qquad (2)$$

Data standardization of all the datasets before training all the predictive models to make the values of each feature in the dataset have zero mean and unit variance. Standardization is helpful to compare attributes that vary in units of measurement and scale where x is the original feature vector, $\bar{x}$ is the mean of that feature vector, and σ is its standard deviation in the equation (3).

$$x^s = \frac{x - \bar{x}}{\sigma} \qquad (3)$$

The study tested the correlation using Pearson's Correlation Coefficient to see the linear relations of parameters and analysis helpful in looking for patterns within datasets. Correlation delivers a measure of goodness of fit in model formulation and verification. The Pearson's Correlation Coefficient explain the covariance of the two variables divided by the product of their standard deviations (Ghose, 2018) in the equation

$$r = \frac{\sum_{i=1}^{241}(Attribute_1 - \overline{Attribute})(Turbidity_i - \overline{Turbidity})}{\sqrt{\sum_{i=1}^{241}(Attribute_1 - \overline{Attribute})^2}\sqrt{\sum_{i=1}^{241}(Turbidity_i - \overline{Turbidity})^2}} \qquad (4)$$

Attributes and turbidity are the variables used to correlate the datasets. Attributes are Conductivity, DO, TSS, Total Coliforms, Temperature, and pH. Attributes and turbidity are the means of the variables, and r is the correlation coefficient in the above equation (4).

Correlation coefficients range from -1.0 to +1.0. The nearer r is to +1 or -1 and the more closely, the two variables will be associated. As a result, if r is near zero, there is no connection between the variables. This matrix allows us to see and compare which pairs have the highest correlation and understand the commonality of data, to know whether it is linear or non-linear. The *feature selection* using f-regression and correlation was applied to the correlated data to select the attributes with a significant value concerning turbidity. Here we show that with correlation based on feature selection, removing irrelevant attributes in the dataset increased the accuracy of the models. It also reduced the time required for training the predictive models.

Furthermore, the dataset has two groups—the training and testing set to check accuracy and precisions by training and testing it. The training set is the essential component here to fit the predictive models. The testing set is to discern the model's performance on unseen data. Predicting whether the turbidity was bad or good, a model was trained depending on the input attributes selected in the test.

B. Processes

Model Specifications. The study datasets do split into two parts, training, and testing, to check accuracy and precisions training and testing. Here is the initial training data to generate multiple mini train-test splits. Use these splits to tune the model. Thus, you are giving your model a chance to be trained on one subset of data and validated and evaluated at some subset that was not used for training yet, here, need to be careful once your data is sensitive to the order/timing - if that's the case time-series split should be used, and data can't be shuffled. The best way to fight to overfit is to allow your model to learn with more data.

For training data, this study used three (3) techniques used to process the data sets, which were Artificial Neural Network (ANN), Support Vector Machine (SVM), and K-nearest neighbor (KNN). Most of the related sediments studies used ANN in predicting turbidity by combining the function, which compares the net input to the neuron or weights of the input, and the activation function that would generate the output based on the input. However, ANN requires a long training time to understand the trained function or the weights.

According to Lafdani, Nia, and Ahmadi (2012), the Support Vector Machine has been successfully used in information categorization and regression problems. SVM is a supervised learning tool well known for its discriminative power. SVM begins separating the data with a hyperplane and then extends this approach to non-linear decision boundaries using the kernel trick. SVM uses some of the specific Kernel functions, which transform the input vector as the input data from the nonlinear function in this model.

The selection of a suitable kernel function is a complicated stage and is often utilized from a standard kernel function. In general terms, SVM establishes a separated hyperplane w > x + b = 0 that classifies the objects xi correctly, while maximizing the separation boundary between both classes (2/w). For, the KNN model used the seven parameters and turbidity as inputs; the KNN model was then computed at the current dataset to produce one output. Furthermore, a supervised learning algorithm produces a function that maps new examples. Each input parameter is weighted with relative weight and generates the classification of turbidity, the sum of the weighted inputs, and the bias used as the activation function to generate output. The study used the log-sigmoid transfer function. Hence, this function generated output between 0 and 1.

The primary purpose of the activation function was to convert an input signal to a node in an ANN to an output signal. ANN without an activation function was simply a linear regression model. SVM model was used to classify the turbidity. In most cases, a binary classifier contains two possible target values, whether "good" or "bad." The KNN model used the seven parameters and turbidity as inputs; the KNN model was then computed at the current dataset to produce one output.

The performance of the ANN, SVM, and KNN models was evaluated and compared in Namu et al. (2017) and Ghose and Samantaray's (2018) study. MSE was used to measure their model by calculating the mean root squares of errors in the equation (5) below.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y_i)^2 \qquad (5)$$

Evaluation of the Predictive Models

The ANN, SVM, and KNN models' performance will be evaluated using statistical assessments of the predicted and observed outputs. The comparison used MSE as also used by Namu et al. (2017) and Ghose and Samantaray (2018) in their study; MSE computes to measure their model using the equation (5). Here, *n* was the number of data, 241, and *Y* was the vector of the predicted variable that had observed results, represented the mean, and represented the squares of errors.

Figure 5 below displays the frequency of the expected outcome and the actual value of each model. For example, if the result falls on the positive or accurate negative, the prediction is the same outcome. The prediction differed from the work if the result fell on the false positive or false negative.

A confusion matrix evaluates the modes in determining the best model. Accuracy and precision provided numerical results compared with the results from the other models. The closer the MSE was to 0, the lesser error. A 0 value of MSE is ideal but not desired. It may mean that the data was over-fitting, or the testing data was small in size.

Figure 5 is a representation of a Confusion Matrix. A Confusion Matrix is shown to display the frequency of the predicted outcome and the actual value of each model. If the result falls on the true positive or true negative, it means the prediction is the same as the outcome. If the result fell on the false positive or false negative, it meant the prediction was different from the outcome. A confusion matrix was used in the study for the evaluation of the models. Accuracy, precision, and recall provided numerical results that were compared with the result from the other models. Through this, the best model was determined.

	Predicted: Good Turbidity	Predicted: Bad Turbidity
Actual: Good Turbidity	Number of Predicted Good Turbidity	Number of Predicted Bad Turbidity
Actual: Bad Turbidity	Number of Unpredicted Good Turbidity	Number of Unpredicted Bad Turbidity

Figure 5 Confusion Matrix

The accuracy is the number of all correct predictions divided by the total number of the dataset. The closer the accuracy to one (1), the better, while the closer it is to 0, the worst. It shows how often the classifier is correct in formula (6).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

Precision is an excellent measure to determine when the costs of False positives are high by calculating the number of correct optimistic predictions divided by the number of positive predictions. The best precision is one, and the worst is 0 using the formula (7).

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

Recall calculates how many Actual Positives the model capture through labeling it as Positive or True Positive. The best is one, and the worst *is* 0. Better models have higher recall and precision. The best model will select by formula (8).

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

C. Output

A web-based application provides visualization of data and the status of the lake. It aimed to help visualize data and predict the classification of the water turbidity surrounding the lake, whether good or bad. The study used "Laravel," a web framework, to develop the web application in the PHP programming language. The system will select the metadata to visualize the statistical report and a table of all data chosen in the selected parameters—graphs designed to express the information. The operating system used in the development was Windows 10. The requirements for the PC to run the software are as follows: To use in utilizing the system, at least 4GB is the recommended RAM for running it smoothly. Windows or Linux is the recommended operating system for making and testing the software.

4. Results and Discussion

As discussed in the methodology, the study started with selecting the input variables, creating predictive models, determining the best model for the specific station, and developing the web-based system. As discussed previously in the methodology, turbidity was classified into "good" or "bad," depending on its value. Data was also pre-processed by removing outliers and using the average technique for treating the missing values. In creating the model, standardization of data and splitting before using the Artificial Neural Network (ANN), Support Vector Machine SVM, and K-Nearest Neighbor (KNN) to refine the datasets and predictive model provide accurate predictions with turbidity.

Results of Pre-Processing of Data

The data records of 432 from 2009-to 2016 were used in the pre-processing, including missing values, and reduced to 241 after pre-processing. The researcher will add a classifier to determine whether the turbidity level was "good" or "bad." Table 2 shows the statistical summary before and after pre-processing of the datasets. It shows the water quality data, including Conductivity, DO, pH, Total Coliforms, TSS, and Temperature.

TABLE 2 ANALYSIS OF VARIANCE OF THE DATASET

Parameters	Before Pre-processing			During Pre-processing		
	Mean	Standard Deviation	Variance	Mean	Standard Deviation	Variance
Conductivity	463.00	307.62	94630.35	463.00	271.70	73819.53
DO	8.00	1.08	1.16	8.00	0.97	0.95
pH	8.40	0.59	0.35	8.40	0.54	0.29
Total Coliforms	306.00	495.35	245373.10	306.00	457.97	209738.00
TSS	42.00	27.91	778.69	42.00	26.84	720.55
Temperature	29.00	2.67	7.15	29.00	2.67	7.15

The Total Coliforms had the highest standard deviation and variance, with 495.35 and 245373.1. Meanwhile, pH had the lowest standard deviation and variance, 0.59 and 0.35, respectively. The higher the standard deviation, the more the data was scattered from the main. It indicates that the values were far away from each other. The variance and the standard deviation value got lower because the data was less scattered after-treatment of the data. The mean for all parameters was constant because of the averaging technique used to treat data.

Correlation of Attributes of the Study with Turbidity

As shown in Figures 6-11, Pearson's correlation coefficient was determined to determine which key attributes affect measuring the goodness of fit in model simulation with turbidity. TSS had the highest correlation value with 0.455, followed by pH and DO, which had values of 0.24 and 0.22, respectively. Meanwhile, the temperature had the lowest correlation value at -0.038.

The results show a large gap between the TSS, with the highest correlation value with turbidity, and pH, which had the second-highest value. However, critical attributes for the predictive model, the pH, and DO were selected since only a few attributes were. As shown in Figure 6, Conductivity and Turbidity are positively correlated; however, some of the data points dispersed, depicting a weak relationship between the two attributes, which resulted in the correlation value of 0.17196.

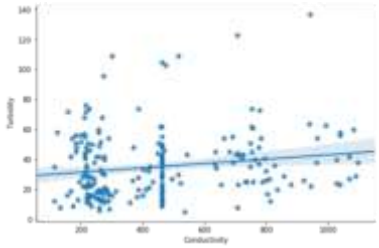

Figure 6 Correlation Between Conductivity with Turbidity

Figure 7 shows that DO and Turbidity are negatively correlated, and most of the data points in an area with the same DO values are concentrated. Some data points with extreme values did not follow the data trend. These created a moderate relationship between DO and Turbidity with a correlation of -0.216293.

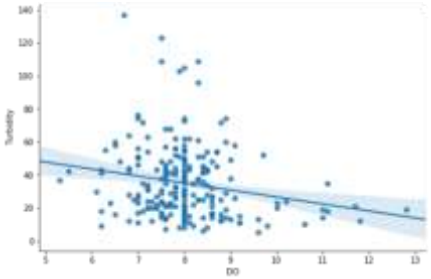

Figure 7 Correlation Between DO with Turbidity

Figure 8 shows how correlated pH is with turbidity. The two attributes are negatively correlated and moderately related to each other. -0.236069 is the correlation between pH and turbidity.

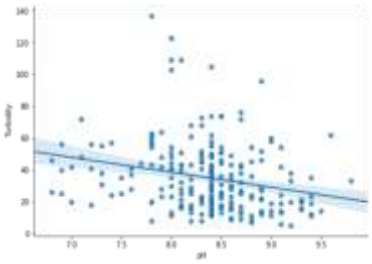

Figure 8 Correlation Between pH with turbidity

The correlation between temperature and turbidity, as represented in Figure 9, almost forms a horizontal line; this means that they have no correlation, but there is a definite pattern. The line shows that the turbidity value will be irrelevant because it will remain the same when the temperature changes. Temperature and turbidity do not correlate because the turbidity value does not show significance with the temperature value. From the looks of the data points, they are far from the line resulting in a -0.03892 correlation.

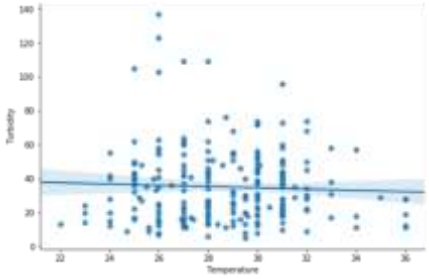

Figure 9 Correlation Between Temperature with Turbidity

As presented in Figure 10, most of the data plots scatter on the lower leftmost part of the graph. It indicates a low correlation between turbidity as the points were far from the line of best fit.

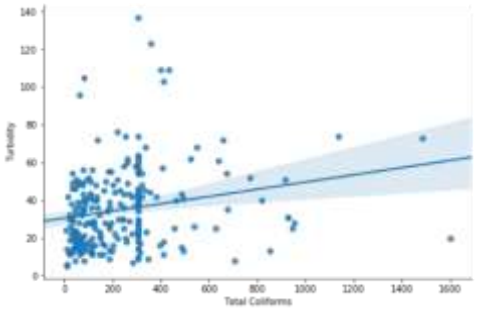

Figure 10 Correlation Between Total Coliforms with Turbidity

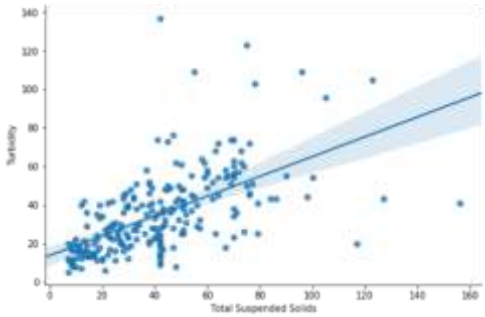

Figure 11 Correlation Between TSS with Turbidity

As shown in Figure 11, the TSS received the highest correlation with turbidity, among other attributes. The plots on the graph indicated that as the TSS goes high, so does the turbidity. A negative correlation means that the attribute's value decreases as the turbidity increases. Meanwhile, the attribute's value increases as the turbidity increases in a positive correlation. The data shows that TSS has the highest correlation among the attributes. The temperature has the lowest correlation, close to zero or no correlation.

Feature Selection Results

To further support the selection of critical attributes based on the correlation result, Table 3 shows the f-values of the attributes with turbidity using feature selection. The attributes were limited to three used in the model. However, to compare results, the study showed three possible approaches to the attributes. Our results show a significant gap between the highest score, the TSS, and the pH's second-to-the-highest score.

Therefore, one approach is to use the TSS only. Another approach includes the top three attributes, TSS, pH, and DO because the three scores were far from the next score. The last approach was to include all the attributes.

TABLE 3 F-REGRESSION OF ATTRIBUTES WITH TURBIDITY

Parameters	Score	Remarks
TSS	62.45	Selected
pH	14.11	Selected
DO	11.73	Selected
Conductivity	7.28	Removed
Total Coliforms	5.11	Removed
Temperature	0.36	Removed

This research assumes that a correlation between features can accomplish feature selection for supervised classification tasks and that such a feature selection process can be beneficial to data mining techniques. Here we show that with correlation based on feature selection, removing irrelevant attributes in the dataset increased the accuracy of the models. It also reduced the time required for training the predictive models on water quality sediment-related problems. Furthermore, the feature selector is fast and straightforward to execute. It eliminates irrelevant and redundant data and, in many cases, improves the performance of learning algorithms. The technique also produces results comparable to other research but requires less computation.

Results of Data Splitting

Before creating the predictive, the researcher divided the data into training and testing sets utilized in all predictive models. Several data splitting was done, thirty percent (30%) of the data was for testing, and seventy percent (70%) of the data was for training, showing the optimal result. The data was split into stations because each station belonged to a different environment. For Stations I, II, IV, VIII, XV, XVI, and XVII, 18 rows for training and nine (9) for testing. For stations V and XVIII, eighteen (18) rows for training and eight(8) for testing. Overall, there were one hundred sixty (162) rows of data for training; ninety (91) were for testing. The total number of data used in the research was two hundred and forty-one (241). Table 4 shows the distribution of the count of data after splitting.

TABLE 4 DISTRIBUTION OF TRAINING AND TESTING DATA

Station	Training Data	Testing Data	Summation per Station
I	18	9	27
II	18	9	27
IV	18	9	27
V	18	8	26
VIII	18	9	27
XV	18	9	27
XVI	18	9	27
XVII	18	9	27
XVIII	18	8	26
Total	162	79	241

Table 5 shows the distribution of the binary classifier in the dataset based on the Oklahoma standard classifier value. To determine what is “good” or “bad,” the turbidity value was checked; lakes with more

should not have more than 25 NTU, and Cool Water Aquatic Community/Trout Fisheries should not have more than 10 NTU, and other surface waters with 50 NTU.

TABLE 5 DISTRIBUTION OF THE BINARY CLASSIFIER IN THE DATASET

Station	Good	Bad
I	9	18
II	13	14
IV	14	13
V	3	23
VIII	10	17
XV	13	14
XVI	7	20
XVII	15	12
XVIII	12	14

Model Evaluation Using Accuracy

We can see that ANN is one of the techniques for the prediction used in the study. After simulating tests to get a sound output for the model, using 11000 max iterations and 0.008 learning rate was optimal to get the least error values—a trial and error method in determining the max iterations and learning rate. For the KNN predictive model of the study, k = 4 is for the neighbor of the KNN. The square root of the data counts to set the value of *k*—the linear kernel SVC model used for the predictive simulation of the SVM technique.

The three-attribute combination with one attribute contains DO, pH, and TSS attributes. In comparison, the six-attribute combination contains all the attributes in the study. Our results demonstrate the three-attribute combination because of the results of the featured selection. Our results demonstrate a three-attribute combination and a one-attribute combination.

Based on Table 6 below, the stations with the highest accuracy for ANN are VIII, XVI, and XVII, which is 0.78. For the SVM model, the station with the highest accuracy was station V, at 0.88, and for KNN stations V and VIII, which was 0.78. The station with the lowest accuracy for the ANN model was station I, 0.22. Meanwhile, stations I, IV, and XV, which were 0.44, received the lowest accuracy for the SVM model, and for the KNN model, stations I and XVII were 0.44.

TABLE 6 PERFORMANCE OF THE SIX-ATTRIBUTE COMBINATION MODEL USING ANN, SVM, AND KNN

	Precision			Recall			Accuracy		
Station	ANN	SVM	KNN	ANN	SVM	KNN	ANN	SVM	KNN
I	0.67	0.80	1.00	0.50	0.71	0.38	0.22	0.44	0.44
II	0.60	0.60	0.60	0.60	0.60	0.60	0.56	0.56	0.56
IV	0.25	0.25	0.43	0.33	0.33	1.00	0.44	0.44	0.56
V	0.86	0.88	0.83	0.86	1.00	0.83	0.75	0.88	0.78
VIII	0.83	0.71	0.75	0.83	0.83	1.00	0.78	0.67	0.78
XV	0.67	0.60	0.71	0.33	0.50	0.83	0.44	0.44	0.67
XVI	0.75	0.67	0.67	1.00	1.00	1.00	0.78	0.67	0.67
XVII	0.50	0.50	0.28	1.00	1.00	1.00	0.78	0.78	0.44
XVIII	0.33	0.33	0.40	0.50	0.50	1.00	0.63	0.63	0.63

Proceedings of RSU International Research Conference (2022)

Table 6 shows that the station with the highest accuracy of predicted values for all the three predictive models of the study is station V, which is 0.88 for ANN and 1.00 for SVM and KNN. The station with the lowest accuracy for the ANN model is station VIII, which is 0.44; for the SVM model, stations II and IV got the lowest accuracy, 0.44 also, and the KNN model has the lowest accuracy for station IV, which is 0.44.

TABLE 7 PERFORMANCE OF THE THREE-ATTRIBUTE COMBINATION MODELS USING ANN, SVM, AND KNN

	Precision			Recall			Accuracy		
Station	ANN	SVM	KNN	ANN	SVM	KNN	ANN	SVM	KNN
I	0.83	0.86	0.88	0.71	0.86	1.00	0.67	0.78	0.89
II	0.67	0.67	0.80	0.67	0.33	0.67	0.56	0.44	0.67
IV	1.00	0.00	0.50	0.60	0.00	0.20	0.78	0.44	0.44
V	1.00	1.00	1.00	0.88	1.00	1.00	0.88	1.00	1.00
VIII	1.00	1.00	1.00	0.38	0.50	0.50	0.44	0.56	0.56
XV	0.40	0.75	0.67	0.50	0.75	1.00	0.44	0.78	0.78
XVI	0.86	0.88	0.75	0.86	1.00	0.86	0.78	0.89	0.67
XVII	0.50	0.75	0.80	0.60	0.60	0.80	0.44	0.67	0.78
XVIII	1.00	0.60	0.60	0.33	1.00	1.00	0.75	0.75	0.75

As presented in Table 7, the station with the highest accuracy in predicting the ANN, SVM, and KNN technique was station VIII, at 1.00 for ANN ad 0.88 for SVM and KNN. The lowest for the ANN technique was Stations I and XV, which was 0.44. Station IV got the lowest accuracy for prediction for the SVM Model, at 0.44, and for the KNN Model, stations IV and XV got the lowest, at 0.56.

TABLE 8 PERFORMANCE OF THE ONE-ATTRIBUTE MODELS USING ANN, SVM, AND KNN

	Precision			Recall			Accuracy		
Station	ANN	SVM	KNN	ANN	SVM	KNN	ANN	SVM	KNN
I	0.75	0.83	0.83	0.43	0.71	0.71	0.44	0.67	0.67
II	0.71	1.00	1.00	1.00	0.40	0.40	0.78	0.67	0.67
IV	0.75	0.00	0.67	0.60	0.00	0.40	0.67	0.44	0.56
V	0.86	0.88	0.88	0.86	1.00	1.00	0.75	0.88	0.88
VIII	1.00	0.83	1.00	1.00	1.00	0.80	1.00	0.89	0.89
XV	0.40	0.60	0.5	0.50	0.75	0.75	0.44	0.67	0.56
XVI	0.75	0.75	0.75	1.00	1.00	1.00	0.78	0.78	0.78
XVII	0.50	0.50	0.5	1.00	1.00	1.00	0.78	0.78	0.78
XVIII	1.00	1.00	1.00	0.50	0.67	0.50	0.63	0.75	0.63

Here we show that when the result of the three tables was compared, KNN for the three-attribute model produced the highest accuracy value, which is five, for stations I, V, XV, XVII, and XVIII. We found in comparing that the higher the accuracy means, the better. The accuracy result of a predictive model should not be equal to one because it signifies that the data are overfitting or too few.

The analysis showed high accuracy for station V in the turbidity dataset of Station V; 23 out of 2 as “bad,” and the data was not sufficient, which led to the prediction that the model was “bad.” For SVM and KNN models for the three-attribute produced results with an accuracy of 1. However, for ANN, the

result is not one but is significantly high, with a value of 0.88. There is a low number of classified "goods" in the data for station V.

We anticipate our model to be a starting point for more sophisticated predictive models of classifying turbidity. For example, having many attributes in the model may lessen the chance of having an accuracy of 1. Furthermore, accuracy and precision for the entire test are inconsistent. The precision for the actual test is primarily from 0.5 to 1. It shows that most of the results were relevant.

Model Evaluation Using Mean Square Error

The MSE was used to evaluate the model. For accuracy, the higher the value, the better, but for MSE, the lower the MSE, the minor error the model produced. Table 9 shows that the value in Station V was significantly low due to the lack of data for the classifier. Only three were as "good" for the 26 rows of data in station V, only three as "good," and 23 as "bad."

However, the six-attribute combination and one attribute produced low MSE for station VIII, while the MSEs are in the middle of the three-attribute combination. The results were not consistent. It can indicate that having a small dataset affects the development of the model. The results of the MSE confirm the results on the accuracy. Here we show that the computation for MSE and accuracy is correct. When the accuracy is 1, the MSE is 0, and the results in evaluating the model using the accuracy method were the same as in using the Mean Square Error method. Thus, the higher the accuracy, the lower the MSE. It indicates that the more accurate the model is, the minor error it produces.

Values in Station V were significantly low due to the lack of data for the classifier. For the 26 rows of data in station V, only three as "good" and 23 as "bad." There was only one typical pattern found for all the testing, and that was for station V, which all produced a low MSE. However, the six-attribute combination and one attribute produced low MSE for station VIII, while the MSEs are in the middle of the three-attribute combination. The results were not consistent. It can indicate that having a small dataset affects the development of the model.

TABLE 9 ANALYSIS OF MEAN SQUARED ERROR FOR PREDICTIVE MODELS PER STATION

Station	TSS			DO, TSS, pH			Conductivity, DO, pH, TSS, Total Coliform		
	ANN	SVM	KNN	ANN	SVM	KNN	ANN	SVM	KNN
I	0.56	0.33	0.33	0.33	0.22	**0.11**	0.78	0.56	0.56
II	**0.22**	0.33	0.33	0.44	0.56	0.33	0.44	0.44	0.44
IV	0.33	0.56	0.44	**0.22**	0.56	0.56	0.56	0.56	0.44
V	0.25	0.13	0.13	0.13	**0.00**	**0.00**	0.25	0.13	0.13
VIII	**0.00**	0.11	0.11	0.56	0.44	0.44	0.22	0.33	0.22
XV	0.56	0.33	0.44	0.56	**0.22**	**0.22**	0.56	0.56	0.33
XVI	0.22	0.22	0.22	0.22	**0.11**	0.33	0.22	0.33	0.33
XVII	**0.22**	**0.22**	**0.22**	0.56	0.33	**0.22**	**0.22**	**0.22**	0.56
XVIII	0.38	**0.25**	0.38	**0.25**	**0.25**	**0.25**	0.38	0.38	0.38

Note: The highest values for the station are in bold

The results of the MSE confirm the results on the accuracy. It shows that the computation for MSE and accuracy is correct. When the accuracy is 1, the MSE is 0. The accuracy method was the same as the Mean Square Error method. The higher the accuracy, the lower the MSE. It indicates that the more accurate the model is, the minor error it produces.

Performance of the Model

Table 10 presents the best model that fits the stations. After comparing all combinations, the summary table shows which machine learning technique with the attribute combination will use in a specific station. An accuracy of 1 is the highest in stations V and II. 0.75 is the lowest accuracy, which is in station

XVIII. The model that appears the most is KNN-3, with a three-attribute combination that was found best in Stations I, V, XV, XVII, and XVIII

TABLE 10 PERFORMANCE EVALUATION OF DIFFERENT ATTRIBUTES PER STATION

Station	TSS			DO, TSS, pH			Conductivity, DO, pH, TSS, Total Coliform, Temperature		
	ANN	SVM	KNN	ANN	SVM	KNN	ANN	SVM	KNN
I	0.44	0.67	0.67	0.67	0.78	**0.89**	0.22	0.44	0.44
II	**0.78**	0.67	0.67	0.56	0.44	0.67	0.56	0.56	0.56
IV	0.67	0.44	0.56	**0.78**	0.44	0.44	0.44	0.44	0.56
V	0.75	0.88	0.88	0.88	**1.00**	**1.00**	0.75	0.88	0.78
VIII	**1.00**	0.89	0.89	0.44	0.56	0.56	0.78	0.67	0.78
XV	0.44	0.67	0.56	0.44	**0.78**	**0.78**	0.44	0.44	0.67
XVI	0.78	0.78	0.78	0.78	**0.89**	0.67	0.78	0.67	0.67
XVII	**0.78**	**0.78**	**0.78**	0.44	0.67	**0.78**	**0.78**	**0.78**	0.44
XVIII	0.63	**0.75**	0.63	**0.75**	**0.75**	**0.75**	0.63	0.63	0.63

Note: The highest value for the station is in bold

The three-attribute combination produced the greatest number of highest accuracies per station, at eleven. Next to the three-attribute combination is the one-attribute, at six. Lastly is the six-attribute combination, which had two counts of highest accuracy per station. The data mining technique with the highest accuracy per station was SVM, which had seven. Meanwhile, ANN and KNN had the same number, at six. The three machine learning techniques' frequencies were not far from each other. The three-attribute combination was best for all the stations except for Stations II and VIII. It signifies that the three-attribute combination is the best combination to be used when considering the totality of the lake.

We found in the model that the most is KNN-3. KNN with a three-attribute combination was found best in Stations I, V, XV, XVII, and XVIII. It was not the model with the highest accuracy in Stations II, IV, VIII, and XVI. In Station V, two models resulted in an accuracy of one. The researchers checked the dataset and detected that three out of twenty-six were classified as “good” while the other twenty-three will classify as “bad .”It indicates that a small number of one part of the binary classifier produces an accuracy of 1, which is not ideal. Stations with a closer count of “good” and “bad” most likely produced accuracy lower than the stations with a far count of “good” and “bad.”

The model performs perfectly well on the training set during the performance model training and fails badly on the testing set. The validation set in most cases indicates that the model is overfitting. However, it does not generalize well enough and thus learns perfectly well how to make predictions on a dataset of samples, but once fed some data it has never seen before, the performance goes down. Here we show that is a bias-variance trade-off to deal with: either fitting very well some data or fitting most of the data, but with, of course, some limited quality, which might result in underfitting. Furthermore, we found that the model performs exceptionally well at the training set, with a> 90-95% accuracy, probably already facing overfitting and below 75–80% underfitting. Our results demonstrate overfitting—cross-validation results in preventive measures against overfitting. For example, setting k = 10 resulted in 10-fold cross-validation by randomly shuffling the dataset into ten sets so that both sets are equal in size during the validation process. Furthermore, it will determine the model's hyperparameters testing that resulted in the lowest test error, which helped the researcher use the water quality sediment data and gave much more information.

Evaluation of the system

The User Acceptance Testing gained was conducted to validate the web-based application. The study used a survey questionnaire. Rizalina E. Germino, an Environment Management Specialist, and two

Engineers evaluated the system to test the web application's functionality. The user is satisfied system's functionality, thus getting a score of 3.0 out of 4.0. The system's overall design received a SATISFACTORY rating, with an average score of 3.0 out of 4.0. The web-based application will design and evaluated to implement the developed models in the study. The system's overall design received a SATISFACTORY rating, with an average score of 3.0 out of 4.0. The following criteria will evaluate by the looks of the color, text, words, graphics, and content.

5. Conclusion

The study found that the TSS, DO, and pH using correlation-based feature selection were the only attributes found compelling after measuring their goodness of fit. Here we show that the 3 out of 6 parameters were the superior selected critical attributes for utilizing the predictive model. However, due to a significant gap between the highest value, TSS, followed by the pH, the study also simulated the predictive model using only TSS. The researchers included all attributes to check and compare the results for testing purposes. Our results demonstrate how the attribute combinations produced different results from one another, for the six-attribute variety did not make the accuracy of 1 or an MSE of 0. The three-attribute combination and the one attribute had an accuracy of 1 or an MSE of 0. We can see that removing the attributes imposes a better model in terms of the accuracy and error percentage.

Furthermore, the accuracy and MSE of the models will compare. Many characteristics in the model may lessen the chance of accuracy of 1. The researchers found that the class count in the data affects the accuracy provided by the model. The less the count of one part of the binary classifier present in the data, the more likely the accuracy will be closer to one. They are, furthermore, experimenting on more attributes that highly correlate to turbidity to improve the model's prediction. A small data size for training can result in forecasts with extreme values, leading to an accuracy of 1. The unbalanced frequency of the classifiers in the dataset leads to an accuracy of 1, so having a vast number of data can prevent this.

During the pre-processing, statistics determine the numerical value for relationships that are usually directly obvious from the historical data. Despite the apparent ability of statistics to overcome problems within the historical data, those valuable results are entirely dependent upon suitable sample parameters, that is, the ability of the original water quality sediment sample to truly represent the environment described and analyzed—moreover, accuracy and control of subsequent analytical measurements. Here we show that data handling techniques are the foundation and must include sampling and analytical methods, missing values, outliers, duplicates, trends, and standards.

For the predictive model, these results provided better trends and information to protect the body of water for public health and avoid water pollution. Furthermore, data mining techniques models give more details on the classification of turbidity when water quality is good or bad, which help in making decisions about when and where it will be helpful to collect water quality sediment data and how to use the information for future planning and forecasting to protect water from pollution. Standard analytical techniques are available to meet most sampling and analytical requirements for water quality studies, but the same is not valid for sediment studies. Sediment sampling and analysis require the use of different techniques and equipment. The difficulty of comparing data sets caused by distinct selection and analytical procedures may arise. Therefore, the standards and comparison of the datasets' procedures are essential in classifying turbidity and necessary for water analyses. For this reason, this research intends to present a conceptual framework of data mining techniques. A more practical and generally applicable predictive modeling that can aid in understanding the relationships between parameters and turbidity analytical methods from the existing database. The study suggests the integration of IoT by implementing sensors for real-time data gathering. Therefore, the researcher recommends having more data is highly recommended.

Proceedings of RSU International Research Conference (2022)

6. Acknowledgements

While conducting the research, we received assistance and support from others. First, we would like to thank Malayan Colleges Laguna and the College of Computer and Information Science for conducting this study. We want to acknowledge the Laguna Lake Development Authority, or LLDA, for providing assistance and responding to our requests. Finally, to God, who has guided us in everything we do.

7. References

Al-Baidhani, J. H., & Alameedee, M. A. (2017). Prediction of Water Treatment Plant Outlet Turbidity using Artificial Neural Network. *International Journal of Current Engineering and Technology*, *7*(4), 1559-1565

Cinco, M. (2017). *Pollution, squatting, and industries hasten the death of Laguna de Bay*. Philippines Daily Inquirer. Retrieved from https://newsinfo.inquirer.net/859365/pollution-squatting-industries-hasten-death-of-laguna-de-bay

Iglesias, C., Torres, J. M., Nieto, P. J., Fernández, J. R., Muñiz, C. D., Piñeiro, J. I., & Taboada, J. (2014). Turbidity Prediction in a River Basin by Using Artificial Neural Networks:A Case Study in Northern Spain *Water Resources Management*. US: Springer.

Melesse, A., S, A., McClain, M., Wang, X., & Lim, Y. (2011). Suspended sediment load prediction of river systems: An artificial neu al. *AgriculturalWater Management*, *98*(5), 855-866.

Namu, P. N., Raude, J. M., Mutua, B. M., & Wambua, R. M. (2017). Prediction of Water Turbidity using Artificial Neural Networks: A Case Study of Kiriku-Kiende Settling Basin in Embu County, Ke ya. *American Journal of Water Resour*, *5*(3), 54-62.

Ghose, D. K., & Samantaray, S. (2018). Modelling sediment concentration using back propagation neural network and regression coupled with genetic algorithm. *Procedia Computer Science*, 85-92. Retrieved from ScienceDirect: https://www.sciencedirect.com/science/article/pii/S1877050917327771

Baseri, H. (2010). Modeling of Grinding Wheel Sharpness by Using Neural Network. *International Conference on Computer Systems and Technologies* (pp. 205-210). Sofia, Bulgaria: CompSysTech'10.

Kumar, D., Pandey, A., Sharma, N., & Flügel, W.-A. (2015). Daily suspended-sediment simulation using machine learning approch. *Catena*, *138*, 77-90.

Lafdani, E. K., Nia, A. M., & Ahmadi, A. (2012, November 2012). Daily suspended sediment load prediction using artificial neural networks and support vector machines. *Journal of Hydrology*, 50–62.

UNESCO-WHO, (online). Publication Studies, and Reports in Hydrology, No 23. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000129296