

Pattern Extraction of Water Quality Prediction Using Machine Learning Algorithms of Water Reservoir

Jefferson L. Leros

Technological Institute of the Philippines / Graduate Programs, Manila, Philippines

Email: jeffersonlerios@yahoo.com

Mia V. Villarica

College of Computer Studies, Laguna State Polytechnic University, Sta. Cruz, Philippines

Email: mia.villarica@gmail.com

Abstract— Model prediction and pattern extraction in data mining allow data to be processed by extracting useful information stored in large databases. The study aims to implore data mining technique for pattern extraction and model prediction of water quality in water reservoir using different parameters and water quality index. A well-known machine learning algorithms such as Naive Bayes, Decision Tree, Random Forest, Gradient Boost and Deep learning algorithms were used for data analysis and interpretation. The result indicated that water quality index was mostly in fair and marginal rank that indicates of water quality was being threatened by different water pollutants.

Index Terms— data mining, water quality index, decision tree, naive bayes, deep learning, machine learning

I. INTRODUCTION

The biggest inland body of water in the Philippines and the 3rd largest in South East Asia is the Laguna De Bay. It has a surface area of 900 km², with 3.2 billion cubic meters volume of waters and an average deep of 2.5 meters. It is strategically in place at the urban development area with national and regional development in fisheries, agriculture, water supply, and the different business industry sectors. However, the latest analyses of a water sample from the Laguna Lake and rivers indicated low dissolved oxygen, high dissolved nitrogen, high level of fecal coliforms, elevated levels of phosphates and some concentrations of lead and cadmium according to the Laguna Lake Development Authority (LLDA). The degradation of the lake is mainly because of overcrowded fish pens, industrial water waste discharge, and domestic waste coming from households areas among others that are thrown in the lake. Because of these water pollutants, Laguna Lake and its elements are heavily polluted and dying [1].

Several studies were conducted to determine the environmental conditions of the lake that focused on its physical characteristics. A graphical illustration of the

lake temperature using the Environmental Fluid Code (EFDC) shows that the lake is thermally unstratified with its average depth of 3.21 meters and with mean temperature range from 25 °C to 29 °C [2]. A bio-optical model was used to monitor the water quality condition with the application of Time-series satellite imagery showing direct retrieval turbidity, chlorophyll and spatial distribution of temperature [3]. Moreover, an investigation of the water circulation using a three-dimensional (3D) hydrodynamic model was developed to analyze the numerical simulations and intensive measurements from field observations to elucidate hydrodynamic implication to eutrophic vulnerability [4].

The study aims to apply data mining techniques to extract knowledge from the dataset in evaluating and classifying the water quality based on different parameters. Data mining is a known process that converts raw data into viable information that can be used for research where data are collected and then extracted from the dataset, using machine learning technique it will be classified that could result to prediction and pattern extraction on water quality.

The paper was organized as follows: section 2 for related works, section 3 methodology, section 4 results and discussion and section 5 for the conclusion.

II. RELATED WORKS

Data mining was used to extract knowledge from the dataset in evaluating the water quality parameters using different machine learning algorithms in several scholastic articles. In the study of Aburub et.al. they investigated four well know data mining techniques such as CBA, SVMs, NB and KNN to predict the ground water sites in Jordan. The result of the analysis shows that SVMs algorithm prevailed in terms of classification, accuracy and precision evaluation measurement [5].

Other classifiers were also used to find a suitable classification model utilizing distinctive methods for predicting water quality these were Bayes Model using Naïve Bayes Algorithm, Rules Model using Conjunctive

rule Algorithm, Trees model using J48 algorithms, Lazy model using Kstar algorithms and Meta Model using Bagging Algorithms. The Lazy model using Kstar algorithm obtained the most outstanding accuracy of 86.67% [6].

Deep Learning Algorithm was used in the predictive analysis of water quality parameters in Chaskaman River, Maharashtra, India. The study assesses the different aspect of water quality in order to reduce the effect of contaminated water. The water quality parameters included in the study were dissolved oxygen (DO), Potential Hydrogen (pH) and Turbidity. These parameters were utilized to provide fairly accurate predictions on the water quality for better water management. The method employs deep learning techniques which used unsupervised learning to provide accurate results compared to supervised learning technique. The result shows that the unsupervised learning technique predicted the acceptable accuracy rate based on the data variation [7].

III. METHODOLOGY

Using the Knowledge Discovery in Database (KDD) the water quality results were utilized to evaluate and classify the best among the model and algorithms for pattern extraction and prediction. The methodology was adopted from Dr. Alexander Rieger based on the step-by-step process in KDD as illustrated in Fig 1. A software platform for data science such as RapidMiner used to provide an integrated environment for data preparation, machine learning, and deep learning.

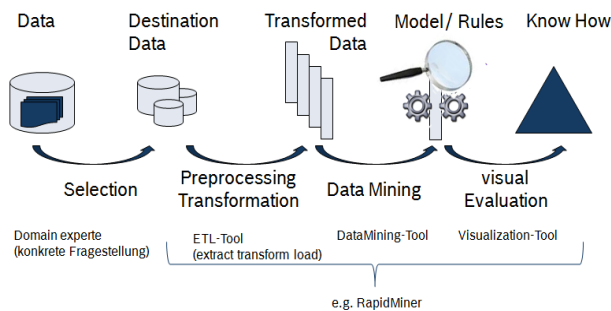


Figure 1. Knowledge discovery in databases process

A. Experimental Data

The dataset of the study was originated from the Laguna Lake Development Authority's (LLDA's) consisting of the water quality monitoring results from the nine stations namely: Central West Bay, East Bay, Central Bay, Northern West Bay, South Bay, West Bay and Central Bay of the Laguna Lake. The parameters used in the study were potential of hydrogen (pH), Ammonia (NH₃), Nitrate (NO₃-H), Inorganic Phosphate (PO₄³⁻), Biochemical Oxygen Demand (BOD), Dissolved Oxygen (DO) and Fecal Coliform (Fcoli) from the year 2015 to 2017. It also includes the station from where the water samples were collected, the month and year respectively.

To determine the Water Quality Index (WQI), the researcher utilized the Canadian Council of Ministers of the Environment (CCME), calculator. It was used to assess the water quality with a comparison to the standard set by the Department of Environment and Natural Resources (DENR). Accordingly, the model was highly applicable and acceptable in terms of water quality index evaluation. The WQI served as an important label for the data mining process for pattern extraction and prediction of its ecological conditions as shown in Table I.

TABLE I. WATER QUALITY RANKING ACCORDING TO CCME GENERAL DESCRIPTION.

Rank	WQI Range	Ecological Condition
Excellent	100-95	No Water uses that fall under certain water classification Conditions very close to natural or pristine under certain water classification
Good	94-80	Water Quality is protected but occasionally threatened or impaired Under certain water classification rarely depart from natural or desirable quality
Fair	79-65	A single water used under certain water classification may be temporarily interrupted Conditions sometimes depart from desirable quality as required under certain water classification
Marginal	64-45	Water quality under certain water classification is frequently threatened or impaired Conditions often depart from desirable quality as required under certain water classification
Poor	44-0	Water quality under certain water classification is almost always threatened or impaired Conditions usually depart from desirable quality as required under certain water classification

A. Data Transformation

The data transformation follows a series of steps that begins with data selection of water quality parameters. To avoid incorrect and missing values the process of data cleaning takes place to ensure the sanity of data integrity. A normalization technique was applied to the dataset from if it came from a large database; in this case, it was encoded into a Microsoft Excel file. Table II presents the description of the dataset used in the study.

TABLE II. DESCRIPTION OF THE DATASET

Parameters	Description
Station	The place where water samples are collected
Date	The date of water samples collection
Potential Hydrogen (pH)	Values ranging from 6.5-9.0
Ammonia (NH ₃)	mg/L (0.05)
Nitrate (NO ₃ -H)	mg/L (7)
Inorganic Phosphate (PO ₄ ³⁻)	mg/L (0.5)
Biochemical Oxygen Demand (BOD)	mg/L (1-7)
Dissolved Oxygen (DO)	mg/L (5)
Fecal Coliform (Fcoli)	mg/L (200)
Water Quality Index	the rank based on CCME (2001) Ecological Condition

B. Pattern Extraction

A model prediction framework was adopted from S.Y. Muhammad et.al to illustrate the process of pattern extraction and prediction. Fig. 2 shows the structural procedure of pattern extraction by the application of machine learning algorithms. In this study, different algorithms were applied as classifiers for model prediction and pattern extractions. These were Naïve Bayes, Generalized Linear Model, Deep Learning algorithm, Decision Tree, Random Forest, and Gradient Boosted Trees.

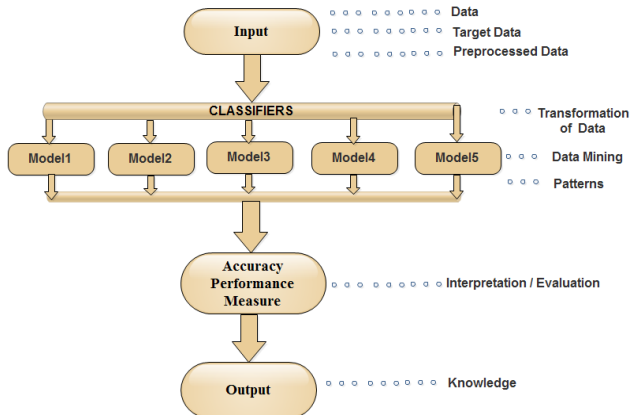
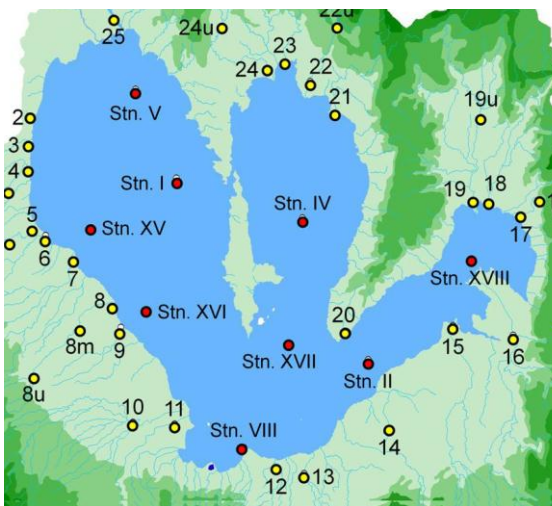


Figure 2. Model prediction framework (Muhammad et.al)

IV. RESULTS AND DISCUSSIONS

The water quality monitoring results were collected from the Nine (9) stations of the Laguna Lake. The stations were the following: Station 1 (Central West Bay), Station 2 (East Bay), Station 4 (Central Bay), Station 5 (Northern West Bay), Station 8 (South Bay), Station 15 San Pedro (West Bay), Station 16 Sta. Rosa (West Bay), Station 17 Sta. Rosa (West Bay) Station 17 Fish Sanctuary (Central Bay) and Station 18 Pagsanjan east Bay. Fig. 3 shows the geographical location of the different stations where the water samples were collected.



Source: LLDA Annual Report Quality Report on the Laguna de Bay and Its Tributary 2009-2012

Figure 3. Geographical location of the 9 stations in Laguna Lake.

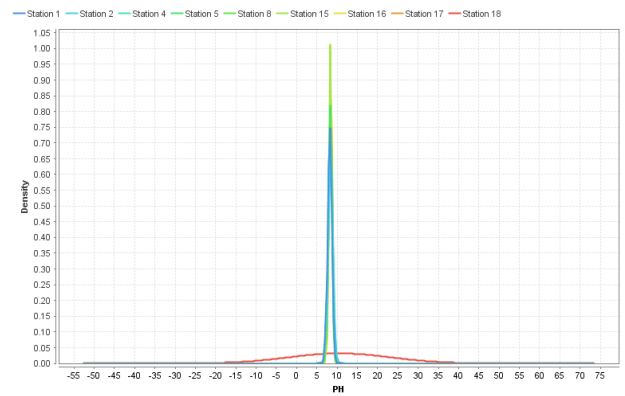


Figure 4. Potential Hydrogen Distribution

pH or potential Hydrogen indicates the acidity of the water quality alteration of pH level can also alter other chemicals in the water. The Pagsanjan East Bay or Station 18 was noted to obtain the lowest pH readings in the span of 3 years. Other stations were within the acceptable range of 6.5 to 8.5 pH.

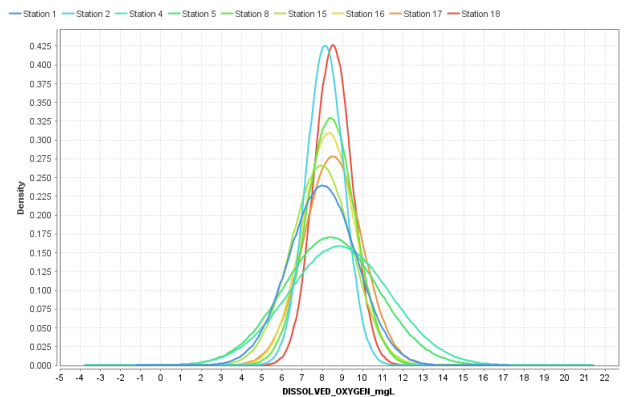


Figure 5. Dissolved Oxygen distribution

When the presence of Dissolved Oxygen (DO) is high in the water, it indicates good water quality. DO is mainly the oxygen and carbon dioxide present in both surface and ground water. The East Bay and Pagsanjan East Bay obtain the highest DO while East Bay and Northern East Bay achieved lowest readings that could potentially affect the water quality and microorganism to grow.

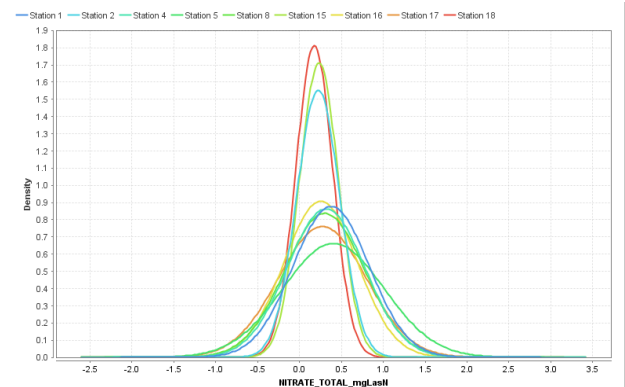


Figure 6. Nitrates Distribution

Algae used nitrates and other plants as a source of food, excess level of nitrates can create a condition for aquatic organism difficult to survive. The Pagsanjan East Bay, West Bay, and the East Bay gained the highest level of Nitrates while other stations were in the acceptable level of 7 per mg/L.

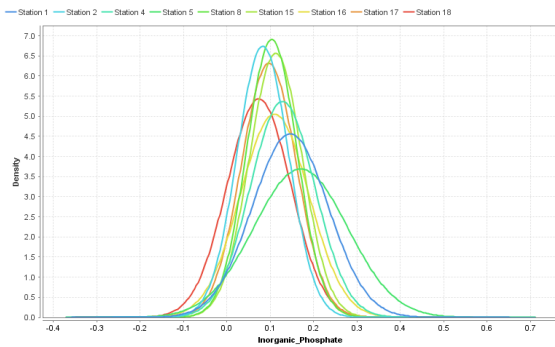


Figure 7. Inorganic phosphate distribution

In aquatic system, phosphorus can be organic phosphate or inorganic phosphate. The excess of phosphate in waterway can cause the condition of eutrophication or over-fertilization of receiving waters, algae and aquatic plants will grow wildly, choke up the waterway and use up large amounts of oxygen. The results indicated that stations Pagsanjan East Bay, San Pedro (West Bay), South Bay, Fish Sanctuary (Central Bay) and the East Bay readings inorganic phosphate were higher than the normal standard values.

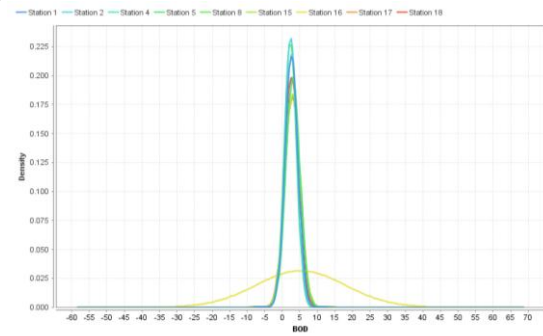


Figure 8. Biochemical oxygen demand distribution

Biochemical Oxygen Demand is defined as the amount of oxygen required to decompose the organic matter transported in surface runoff. The West Bay or the station 16 achieved the lowest reading on BOD while the other stations were within the acceptable level of 7 per mg/L.

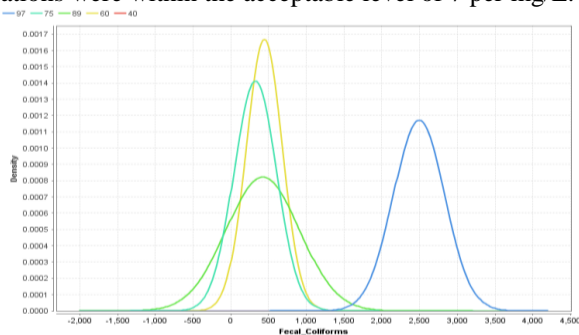


Figure 9. Fecal coliform distribution

Total coliform is a large collection of different kinds of bacteria. Fecal coliform is a type of total coliform that exist in feces. The presence of E. coli in a drinking water sample usually indicates recent fecal contamination. Figure 9 shows that based on the water quality index the results state that fecal coliform in the marginal level was higher. It could mean possible water contamination in certain areas where water samples are collected.

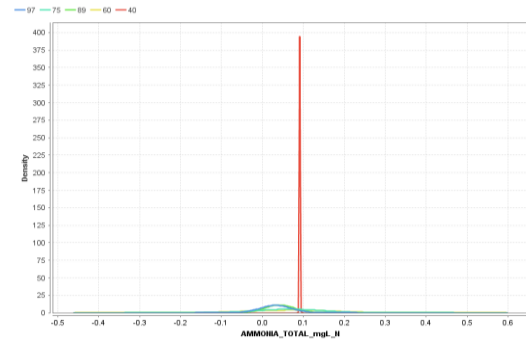


Figure 10. Ammonia distribution

In aquatic environment, ammonia is one of the several forms of nitrogen. When ammonia is present at the water at a high level, it becomes toxic that can cause potentials death of an aquatic organism. The graph shows that ammonia in poor condition is highly noticeable. However, upon thorough investigation the researcher cannot deduce vital information for the reason that the dataset on ammonia was incomplete having only of 2 years data collection 2016 and 2017.

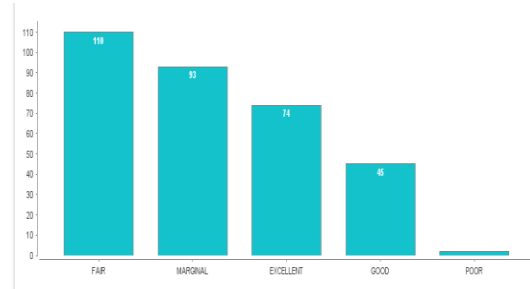


Figure 11. Summary of water quality index

Generally, the Fair rank obtained the values of 110, Marginal 93, Excellent 74, Good 45 and Poor 2. Based on the given figure, the fair rank achieved the highest rank followed by the marginal rank; it states that the ecological condition of the lake was merely in good condition because of the water pollutants. It also means that water quality was degrading in industrialized areas.

TABLE III. ALGORITHM CLASSIFIER'S ACCURACY

Algorithms	Classification Accuracy	Precision
Decision Tree	87.69%	87.7%
Naïve Bayes	72.31%	72.69%
Random Forest	78.46%	78.5%
Gradient Boost	73.8%	73.8%
Deep Learning	72.3%	72.3%

Among the classifiers, the Decision Tree Algorithm achieved the highest accuracy and precision. It was followed by Random Forest, Gradient Boost, Naïve Bayes, and Deep Learning Algorithms. Classification accuracy is the accurateness in terms of prediction while the precision is the correction of the prediction which presented in Table III.

TABLE IV. CONFUSION MATRIX

	true 97	true 75	true 89	true 60	true 40	class precision
pred. 97	73	2	0	0	0	97.33%
pred. 75	1	86	11	18	0	74.14%
pred. 89	0	8	34	2	0	77.27%
pred. 60	0	14	0	73	2	82.02%
pred. 40	0	0	0	0	0	0.00%
class recall	98.65 %	78.18 %	75.56 %	78.49 %	0.00 %	

The confusion matrix in Table IV presents the correctness of the prediction. The numeral value was the water quality index with reference to the water quality ranking in Table 1. It shows that Excellent got 97.33%, Marginal 82.02, Fair 77.27%, and Good with 74.14% class prediction which is all acceptable in prediction level. Nevertheless, the label 40 which is poor, gained nothing in the confusion in table IV. It only shows that in modeling it is necessary to have at least an almost the same number in each label. The F1 result was 67.90% which exceeds the average accuracy prediction for accuracy testing of 50% therefore; the model was classified as accepted. The developed model was shown in Table V.

TABLE V. DEVELOPED MODEL

IF	THEN
FC >2156.900	EXCELLENT
FC ≤ 1543 AND >1099	GOOD
FC ≤ 959 AND >101.500 AND IP >0.050 AND FC>483.500	MARGINAL
FC≤ 959 AND >101.500 AND IP>0.050 AND FC≤483.500 AND IP>0.180	MARGINAL
FC≤ 959 AND >101.500 AND IP>0.050 AND FC≤483.500 AND IP≤0.180 AND DO>12.400	MARGINAL
FC≤ 959 AND >101.500 AND IP>0.050 AND FC≤483.500 AND IP≤0.180 AND DO≤12.400 AND FC>430.500 AND PH>8.200	MARGINAL
FC≤ 959 AND >101.500 AND IP>0.050 AND FC≤483.500 AND IP≤0.180 AND DO≤12.400 AND FC>430.500 AND PH≤8.200 AND AMMO>0.044	FAIR
FC≤ 959 AND >101.500 AND IP>0.050 AND FC≤483.500 AND IP≤0.180 AND DO≤12.400 AND FC>430.500 AND PH≤8.200 AND AMMO≤0.044	MARGINAL
FC≤ 959 AND >101.500 AND IP>0.050 AND FC≤483.500 AND IP≤0.180 AND DO≤12.400 AND FC>430.500 AND ≤ 430.500 AND IP>0.133 AND FC>385	EXCELLENT
FC≤ 959 AND >101.500 AND IP>0.050 AND FC≤483.500 AND IP≤0.180 AND DO≤12.400 AND FC>430.500 AND ≤ 430.500 AND IP>0.133 AND FC≤385	MARGINAL
FC≤ 959 AND >101.500 AND IP>0.050 AND FC≤483.500 AND IP≤0.180 AND DO≤12.400 AND FC>430.500 AND ≤ 430.500 AND IP>0.133 AND FC≤263.500 AND NITRATE>0.055	FAIR
FC≤ 959 AND >101.500 AND IP>0.050 AND FC≤483.500 AND IP≤0.180 AND DO≤12.400 AND	MARGINAL

FC>430.500 AND ≤ 430.500 AND IP>0.133 AND FC≤263.500 AND NITRATE≤0.055 AND NITRATE>0.024	
FC≤ 959 AND >101.500 AND IP>0.050 AND FC≤483.500 AND IP≤0.180 AND DO≤12.400 AND FC>430.500 AND ≤ 430.500 AND IP>0.133 AND FC≤263.500 AND NITRATE≤0.055 AND NITRATE≤0.024	FAIR
FC≤ 959 AND >101.500 AND IP>0.050 AND FC≤483.500 AND IP≤0.180 AND DO≤12.400 AND FC>430.500 AND ≤ 430.500 AND IP≤0.133 AND PH ≤0.024	FAIR
FC≤ 959 AND >101.500 AND IP≤0.050 AND FC>855.500	MARGINAL
FC≤ 959 AND >101.500 AND IP≤0.050 AND FC>855.500 AND >317	FAIR
FC≤ 959 AND >101.500 AND IP≤0.050 AND FC>855.500 AND <317 AND PH>9.100	FAIR
FC≤ 959 AND >101.500 AND IP≤0.050 AND FC>855.500 AND <317 AND PH≤9.100	GOOD
FC≤ 959 AND >101.500 AND IP>0.154	FAIR
FC≤ 959 AND >101.500 AND IP≤0.154 AND STATION 2 AND PH>8.750 AND NITRATE>0.269	FAIR
FC≤ 959 AND >101.500 AND IP≤0.154 AND STATION 2 AND PH>8.750 AND NITRATE≤0.269	GOOD
FC≤ 959 AND >101.500 AND IP≤0.154 AND STATION 2 AND PH≤8.750	GOOD

V. CONCLUSIONS

Data mining is an interesting field to explore hidden knowledge that can be extracted from the large database. It is being widely used for data science and analysis to expedite the used of information in the specific field of specialization. The study implored data mining technique in order to analyzed the water quality monitoring results of Laguna Lake using seven parameters namely potential of hydrogen (pH), Ammonia (NH₃), Nitrate (NO₃-H), Inorganic Phosphate (PO₄³⁻), Biochemical Oxygen Demand (BOD), Dissolved Oxygen (DO) and Fecal Coliform (Fcoli) from 9 different stations. The water quality index was calculated using the Canadian Council of Ministers of the Environment (CCME) calculator with the corresponding description of the ecological condition. The study utilized (5) five models for pattern extraction and prediction. Among the classifiers, Decision Tree obtained the highest accuracy and precision. This indicates the level of confidence in terms of the model prediction.

The results showed that most of water quality index of the station belongs to fair rank a condition where the quality of water may be temporarily interrupted. The study also revealed that water samples from different parameters were in a marginal rank where water conditions were frequently threatened. This was an important information derived from the study that could serve as a basis to a more aggressive campaign on the Laguna Lake water conservation and protection

ACKNOWLEDGMENT

The researchers would like to express their heartfelt appreciation and profound gratitude to the Laguna Development Authority for providing the dataset of the study.

REFERENCES

- [1] L. de Bay. (n.d.). [Online]. Available: <http://gwhs-stg03.i.gov.ph/~s3lldagov/index.php/laguna-de-bay/>
- [2] A. M. Cunanan, J. W. "Analysis of water temperature of laguna lake using EFDC model," *International Journal Of Scientific & Technology Research*, vol. 3, no. 8, 2014.
- [3] Enrico C. Paringit, K. N., "Monitoring and assessment of Laguna de bay waters from space observations and optical modeling," *Research Gate*, 2014
- [4] Eugene C. Herrera, K. N. "Hydrodynamic investigation of a shallow tropical lake environment (Laguna Lake, Philippines) and associated implications for eutrophic vulnerability," *ASEAN Engineering Journal Part C*, vol. 4 no. 1 , EnvE Special Issue, 48
- [5] F. Aburub et.al. "Predicting groundwater areas using data mining techniques: Groundwater in Jordan as case study," *International Journal of Computer and Information Engineering*, vol. 10, no. 9 , 2016.
- [6] S. Y. Muhammad et.al, "Classification model for water quality using machine learning techniques," *International Journal of Software Engineering and Its Applications*, vol. 9, no. 6, 2015.
- [7] Archana Solanki, H. A. (2015). "Predictive analysis of water quality parameters using deep learning," *International Journal Name of Computer Applications*, vol. 125, no. 9

Jefferson L. Lerios is a graduate student from the Technological Institute of the Philippines Manila taking up Doctor of Information Technology. His research interest is in the field of information technology, data mining, and data analytics. Currently, he is an Assistant Professor II of the College of Computer Studies at Laguna State Polytechnic University – Los Baños Campus.

Dr. Mia V. Villarica earned her degree in Doctor of Information Technology at AMA Computer University, Quezon City Campus in the year 2017. She was the former Associate Dean of the College of Computer Studies Laguna State Polytechnic University Sta. Cruz, Main Campus. She attended various training and seminars in Mobile Development, Information Resource Management, and Software Engineering.