*Journal of Sustainability Science and Management*
*Volume 18 Number 9, September 2023: 171-183*

*eISSN: 2672-7226*
*© UMT Press*

# PREDICTION OF WATER QUALITY FOR THE SELANGOR RIVERS USING DATA MINING APPROACH

NURAIN IBRAHIM[1,2]*, HEZLIN ARYANI ABD RAHMAN[1], AHMAD ARIFUDDIN AZRAN[1], MUHAMMAD AIMAN MOHD FADDILLAH[1] AND MUHAMMAD ADHWA' QAYYUM MOHD QAMARUDIN[1]

*[1]School of Mathematical Sciences, College of Computing, Informatics and Mathematics, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia. [2]Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Kompleks Al-Khawarizmi, Universiti Teknologi MARA, 40450 Shah Alam, Selangor, Malaysia.*

**Corresponding author: nurain@tmsk.uitm.edu.my*

*Submitted final draft: 30 May 2023* *Accepted: 30 June 2023* http://doi.org/10.46754/jssm.2023.09.0012

**Abstract:** Few studies using the data mining approach to assess the quality of water, especially for Selangor rivers. This study assesses the water quality using data mining techniques and identified the most significant variables that affect water quality. Machine learning techniques used are Decision Tree (Gini) and Decision Tree (Entropy), Logistic Regression Enter, Backward Elimination and Forward Selection and Artificial Neural Network with 4 and 8 hidden nodes. This study revealed that Logistic Regression Enter is the best model since it is neither underfit nor overfit with the sensitivity, specificity, accuracy, mean squared error and misclassification rate values of 92.51%, 97.45%, 96.36%, 0.028 and 3.64% respectively. There are other two best models: Decision Tree (Gini) and Artificial Neural Network with 4 hidden nodes. According to the variable importance output based on Decision Tree (Gini), the most important variable effect on the water quality is Biochemical Oxygen Demand (BOD) with the highest value of 0.2284, followed by Chemical Oxygen Demand with a value of 0.1471 respectively.

Keywords: Water quality, decision tree, logistic regression, Artificial Neural Network.

## Introduction

Water is an essential component of life on the planet. A human cannot go more than two days without water. It has a molecular makeup of $H_20$, and the human body is about 75% water. Water serves a variety of roles in the human body, including lubrication, body temperature regulation, the removal of pollutants and the movement of nutrients throughout the body. Water also aids in the lubrication and cushioning of our joints (Lorenzo *et al.*, 2019). Water covers 71% of the earth's surface, with the oceans accounting for 96.5% of all water on the planet. Rivers are where we obtain the majority of our fresh water. In Malaysia, rivers account for 97% of the total water supply, according to Air Selangor.

Water quality refers to the chemical, physical, and biological qualities of water, as well as its suitability for certain uses. Water quality standards for industrial applications may differ from standards for drinking. As a result, the quality of a source of water must be analysed to ensure it is suitable for its intended use. Water quality testing is critical, especially for industrial and public health purposes (drinking water). Checking whether the water quality meets standards, rules and regulations, monitoring the efficiency of a system, and working for water quality maintenance are just a few of the reasons why water quality analysis is required before distribution. Water quality analysis entails a series of procedures, from determining the appropriate parameter to selecting the best methodology, sampling, analysis, and reporting. According to the Department of Environment (DOE, 1985) the water quality index can be grouped into three classes, clean (81-100), slightly polluted (60-80), polluted (0-59).

Water is a basic need for every life on earth. Every day we consume water to keep us hydrated or take a bath to keep ourselves refreshed. Malaysia consumes 61.6% of its total treated water for domestic use (Ab Rashid *et al.*, 2021). A disruption to treated water supply can adversely impact daily life, as seen in the Selangor water crises of 2020, when 8 statewide water cuts took place.

Water quality index and prediction of water quality are important to us in our survival species. Suggested important variables that need to be obtained for the best water quality analysis and prediction model include pH value, Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Total Suspended Solid (TSS), Total Dissolved Solids (TDS), temperature and Ammoniacal Nitrogen ($NH_3$–N) (Ahmed *et al.*, 2023). However, for the Data of Analysis (DOA), all these suggested variables should be included as well as the turbidity variable. Several studies have been done on predicting water quality and each study used different variables or parameters using univariate and multivariate statistical techniques, such as principal component analysis and cluster analysis. However, there are a limited number of studies that used the data mining approach in the assessment of water quality, especially for Selangor rivers. Hence, the main aim of this study is to assess water quality using data mining techniques and identify the most significant variables that affect water quality. This study will benefit many groups of people such as students and researchers, especially those in the Statistics and Data Mining fields, and the Department of Environment, as it will discover the best technique for determining the water quality index and the condition of Selangor's rivers. This study is in line with the United Nations' Sustainability Development Goal No. 6: Clean Water and Sanitation.

## Materials and Methods

Secondary data from the Department of the Environment was used. The data comprised water samples obtained from 18 rivers around Selangor and Kuala Lumpur from basins in the Klang Valley. It consists of data from the years 2011 until 2018. The data from the 18 rivers were in the form of 18 Microsoft Excel spreadsheet files. We merged 17 files to create 3 models for classifying water as clean or polluted, and the remaining 1 file (2018 data) was used for testing the best model obtained. Drawing on work by Zainudin (2010), the water quality results were categorised into clean and polluted. A river is categorised as 'clean' if the WQI score is between 81 to 100. Rivers with a WQI of less than 81 were categorised as 'polluted'. Dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), suspended solids (SS), ammoniacal nitrogen (AN), pH, and other external variables are all taken into consideration. Table 1 shows the water quality data description for this study.

Table 1: Water Quality data description

| Terms | Definition | Level | Role | Values |
|---|---|---|---|---|
| Water Quality (WQ Status) | Water Quality Index (WQI) is a figure that expresses the water quality in a simple form by accumulating the measurements of selected parameters. | Binary | Target | 1: Clean (WQI 81 - 100)<br>0: Polluted (WQI 0 - 80) |
| Dissolved Oxygen (DO) | A measure of the amount oxygen dissolved in the water. | Ratio | Input | > 0 mg/l |
| Biochemical Oxygen Demand (BOD) | Shows how much of the dissolved oxygen is consumed by microorganisms during the oxidation of reduced substances in water and waste. | Ratio | Input | > 0 mg/l |

| | | | | |
|---|---|---|---|---|
| Chemical Oxygen Demand (COD) | Shows the degree of organic pollution in water bodies. | Ratio | Input | > 0 mg/l |
| Turbidity | The measure of relative clarity of a liquid. It is an optical characteristic of water and is a measurement of the amount of light that is scattered by material in the water when a light is shined through the water sample. | Ratio | Input | > 0 NTU |
| Temperature | One of the major influences on biological activity and growth. Temperature governs the kind of organisms that can live in rivers. | Ratio | Input | > 0 $^{o}$C |
| Suspended Solid (SS) | The actual measure of minerals and organic particles transported in the water column. | Ratio | Input | > 0 mg/l |
| Ammoniacal Nitrogen (NH3-N) | Ammoniacal nitrogen (can form ammonia in water under certain conditions), is a highly soluble and can be readily used by vascular plants and algae for growth, but it can be toxic to aquatic life in elevated concentrations. | Ratio | Input | > 0 mg/l |
| pH | A calculation of the hydrogen ion concentration of the water as ranked on a scale of 1.0 to 14.0. The lower the pH of water, the more acidic it is. The higher the pH of water, the more basic, or alkaline, it is. pH affects many chemical and biological processes in  water and different organisms have different ranges of pH within which they flourish. | Interval | Input | 1 - 14 |
| Dissolved solid (DS) | The quantity of dissolved material in water, and it is one of the vital water quality parameters, and continuously used to determine the water quality of rivers. | Ratio | Input | 0 - 1000 mg/l |
| Water Level | The depth of a river at a specific location | Ordinal | Input | Low, normal, high. |

### *Logistic Regression*

Logistic regression assists in predicting the probability of a value falling to a certain level with a range of predictors. Logistic regression was used to identify the significant variables in classifying water quality. In this study, turbidity, dissolved solids, water level, pH, temperature, suspended solids, chemical oxygen demand, biochemical oxygen demand, dissolved oxygen and ammoniacal nitrogen will be the predictors and water quality, and act as dependent variables. Three types of variable selection in logistic regression were used in this research: Enter, forward and backward. The equation of logistic regression model is displayed in (1) (Srimaneekarn *et al.*, 2022):

$$logit\,(y) = (odds) = ln\,(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots \beta_n x_n \qquad (1)$$

The method of maximum likelihood yields values for the unknown parameters ($\beta_0, \beta_1, \beta_2, \ldots \beta_n$), which maximises the probability of obtaining the observed set of water quality data given to the model. Concepts that are related to logistic regression are odds, odds ratio and logistic curve. Odds of an event are the ratio of the probability that an event will occur to the probability that it will not occur. In its formula where odds of {Event} $= p / 1 - p$, $p$ represents the probability of the event to occur while (1−p) represents the probability that it will not occur.

In terms of model summary, -2 Log likelihood, Pseudo R Square and Hosmer and Lemeshow test were used in this study. -2 Log likelihood is a likelihood ratio test, which is just the chi-square difference between the null model (i.e., with the constant only) and the model containing the predictors, is the most used method for evaluating the overall model fit in logistic regression. Meanwhile, the aim of Pseudo R Square is to assess the predictive power of the independent variables which is impeding factors. There are several types of Pseudo R Square, including Cox & Snell R Square and Nagelkerke R Square. Besides, Homser and Lemeshow Test is a goodness of fit of the model. It computed the differences between observed and expected proportions. The formula of this test as shown in (2):

$$HL = \sum_{d-1}^{D} \frac{(O_{1d} - E_{1d})^2}{N_d \tau_d (1 - \tau_d)} \qquad (2)$$

where $D$ is the number of groups, $O_{1d}$ is the number of observed Y = 1 event, $E_{1d}$ is the number of expected Y = 1 event, $N_d$ is the total number of observations and $\tau_d$ is the estimated risk for the $d^{th}$ groups.

### *Decision Tree*

Decision tree is a supervised machine learning technique that allows us to estimate a quantitative target variable or classify observations into one category of a categorical target variable by repeatedly dividing observations into mutually exclusive groups. Two types of decision tree used in this research are entropy and Gini index.

The entropy (Information Gain) method chooses the splitting characteristic with the lowest entropy value, resulting in the highest Information Gain. To choose the Decision Tree's splitting attribute, first calculate the Information Gain for each attribute, then choose the attribute that maximises the Information Gain. The following formula shown in (3) (Li *et al.*, 2022) is used to determine the Information Gain for each attribute:

$$E = \sum_{i-1}^{k} Pi\; log_2\, Pi \qquad (3)$$

where $k$ is the number of target attribute classes and $P_i$ is the number of occurrences of class $i$ divided by the total number of instances (i.e. the probability of $i$ happening).

The Gini Index is a metric for data impurity. For each attribute in the data set, the Gini Index is computed. If there are $k$ classes of the target attribute and the chance of the $i^{th}$ class being $P_i$, the Gini Index is defined as follows in (4) (Li *et al.*, 2022):

$$Gini\ index = 1 - \sum_{i-1}^{k} Pi^2 \qquad (4)$$

### *Artificial Neural Network*

Artificial Neural Network (ANN) is one of the methods of data mining. An ANNis made up of three or more interconnected layers. Input neurons make up the first layer. These neurons send input to deeper layers, which then deliver the final output data to the final output layer. The goal of the input layer is to achieve the equivalent conveyance of input data without modifying it in any way. The data transferred from the input layer is combined linearly by the hidden layer, which also calculates it using the activation function. The output layer performs calculations on the data passed by the hidden layer and outputs the results of those calculations in a manner similar to that of the hidden layer. A neural network's operation can be loosely separated into two phases. The learning phase makes up the first phase. By learning a large number of well-known input/output training samples, the neural network continuously modifies the connection weights of neurons in each layer in order to reduce the system error signal. The working phase is the second stage. The neural network's system parameters are currently fixed, and appropriate outputs are produced using the supplied inputs. The theoretical framework of ANN is displayed in Figure 1.

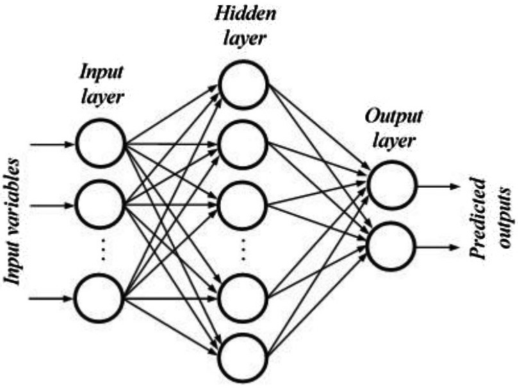

Figure 1: Theoretical Framework of ANN (McKee *et al.*, 2018)

Until now, there are no definitive rules about the number of hidden layers and the best number of neurons in the hidden layer in ANN (Arifin *et al.*, 2019). Therefore, 4 and 8 hidden layers were chosen because they are one third and two thirds of the total 12 input variables.

### *Model Assessment*

The models were assessed using sensitivity, specificity, accuracy, mean square error and misclassification rate. Sensitivity relates to the model's ability to identify positive targets. Specificity is the ability of the model to identify negative targets. The accuracy of a classifier is the probability of correctly predicting the class of an unlabelled instance and it can be estimated in several ways (Rosly *et al.*, 2015; Sweeney *et al.*, 2022). Misclassification Rate is a performance metric that indicates the percentage of incorrect predictions without distinguishing between positive and negative predictions. An estimator's Mean Squared Error (MSE) or Mean Squared Deviation (MSD) measures the average of error squares, or the average squared difference between the estimated and true values. The formulas of the assessment tools are displayed in (5) – (9):

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{5}$$

$$Specificity = \frac{TN}{(TN + FP)} \tag{6}$$

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{7}$$

$$Misclassification\ Rate = \frac{FN + FP}{(TN + TP + FP + FN)} \tag{8}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_j)^2 \tag{9}$$

where *TP* indicates true positive, *TN* indicates true negative, *FP* indicates false positive, and *FN* indicates false negative. Meanwhile, *MSE* indicates the mean squared error, *N* indicates a number of data points,  indicates observed values and  indicates predicted values.

## Result and Discussion

Seven models were built and ran to predict the WQI of 18 rivers, thus comparing between themselves to determine the best model to predict WQI in the future. Models were ran using RapidMiner software. The models consist of Decision Tree with Gini and Entropy Algorithm, Logistic Regression Enter, Forward Selection and Backward Elimination and Artificial Neural Network with 4 and 8 Hidden Nodes.

Performance parameter used is accuracy, where the highest accuracy is chosen as the best model. The accuracy value was obtained from the confusion matrix generated. Squared error and information gain were used to check whether the models underfit or overfit. The following table shows the confusion matrix of the training and testing dataset of the models. Specifically, Table 2 shows the descriptive statistics of each attribute. In terms of the response variable's distribution, the data is balanced with 3396 (45%) are cleaned and 4150 (55%) are polluted. The table consists of the number of missing values before and after imputation, the range of the values in a particular attribute.

### *Data Pre-processing*

Some of the steps involved in the data pre-processing are data profiling, data cleaning, data transformation, and data reduction. In the data transformation, 18 files containing the data of Selangor rivers were compiled. The data was imported into SPSS to check if there are any noise in the data that can interrupt the data analysis process. Several variables with missing values are found, which are DO, BOD, COD, SS, pH, NH3-N, Temperature, Turbidity, DS, and WQI (Table 3). The problem was overcome by using Replace Missing Values function in SPSS.

Table 2: Descriptive Statistics

| | **N** | | | | **Minimum** | **Maximum** |
|---|---|---|---|---|---|---|
| | **Valid** | | **Missing** | | | |
| | **Before** | **After** | **Before** | **After** | | |
| Water Level | 7546 | 7546 | 0 | 0 | - | - |
| Dissolved Oxygen (mg/l) | 7542 | 7546 | 4 | 0 | 0.2800 | 11.9800 |
| Biochemical Oxygen Demand (mg/l) | 7494 | 7546 | 52 | 0 | 1.0000 | 89.0000 |
| Chemical Oxygen Demand (mg/l) | 7483 | 7546 | 63 | 0 | 1.9000 | 317.0000 |
| Suspended Solid (mg/l) | 7212 | 7546 | 334 | 0 | .0000 | 4400.0000 |
| pH | 7542 | 7546 | 4 | 0 | 5.4700 | 10.1600 |
| Ammoniacal Nitrogen NH3-N (mg/l) | 7336 | 7546 | 210 | 0 | .0093 | 34.8390 |
| Temperature | 7546 | 7546 | 0 | 0 | 19.9800 | 34.9100 |
| Turbidity | 7492 | 7546 | 54 | 0 | .0000 | 3529.2000 |
| Dissolved Solid (mg/l) | 7249 | 7546 | 297 | 0 | 7.1000 | 27700.0000 |
| WQI | 7416 | 7546 | 130 | 0 | 13.3490 | 97.5734 |

Table 3 shows the descriptive analysis for the variable ‘Water Level’, there were three level which is ‘Low’, ‘Normal’, and ‘High’. Since there were also missing values for this variable, the missing values were replaced with water level ‘Normal’ since it is the mode, with the most frequency in the variable (Zainudin, 2010). Mode imputation was used as it is the simplest method for imputing missing values for categorical variables (Xu *et al*., 2020). Thus, the percentage of ‘Normal’ water level is 92.5% compared to 85.7% before imputation. Other than that, since ‘Low’ has a percentage of less than 5% (2.1%), thus ‘Low’ and ‘High’ water levels were merged and encoded as ‘Others’. Since our target is binary classification, which is 1 and 0 as clean and polluted respectively, the variable WQI with the values of 81-100 were altered to encoded to ‘1’ while values of 0-80 were encode to ‘0’ (Zainudin, 2010).

### *Odds Ratio in Logistic Regression Enter*

As can be seen in the previous section, Logistic Regression Enter outperforms or is comparable to other best methods, such as Decision Tree (Gini) Algorithm and Artificial Neural Network with 4 hidden nodes. Hence, further explanation on the Logistic Regression Enter outputs is in Table 7, which displays the model summary and Hosmer – Lemeshow test, and in Table 8, which displays the coefficient, odds ratio and the p-value of each variable used in building the best model among logistic regression models. In Table 4, the model Cox & Snell $R^2$ statistic indicates that 60% of the variation in the dependent variables is explained by the predictor variable and the value of Nagelkerke $R^2$ is 0.862. Hosmer-Lemeshow shows the goodness of fit of the model. The model fits well since the p-value of 0.857 is greater than the level of significance at 5%. Hence, the Logistic Regression Enter model in this study is a good fit.

### *Performance Evaluation*

### *Logistic Regression*

According to Table 5, all variables significantly affect water quality as the p-values are lower than the significance level of 0.05. Since logistic regression with Enter selection is neither underfit or overfit, Logistic Regression Enter model was chosen as the best Logistic Regression model. Consequently, the final model with significant variables is shown in (10):

$$\begin{aligned} &logit\ (y) = (odds) = In\ (\tfrac{p}{1-p}) = 4.7447 \\ &+ 1.4094[Dissolved\ Solids] - 0.6891 \\ &[Water\ Level\_Others] - 0.4970[pH] - \\ &0.1879[Temperature] - 0.0222[Suspended \\ &Solids] - 0.1121[Chemical\ Oxygen \\ &Demand] - 0.3897[Biochemical\ Oxygen \\ &Demand] + 1.5510[Dissolved\ Oxygen] - \\ &1.1726[Ammoniacal\ Nitrogen\ NH3\text{-}N] \end{aligned} \tag{10}$$

Table 3: Water Level Value’s before and after imputation

| Water Level | | |
|---|---|---|
| **Level** | **Before Imputation<br>N (%)** | **After Imputation<br>N (%)** |
| Missing Value | 510 (6.7) | - |
| High | 412 (5.4) | - |
| Low | 155 (2.1) | - |
| Others | - | 567 (7.5) |
| Normal | 6469 (85.7) | 6979 (92.5) |
| Total | 7546 (100) | 7546 (100) |

Table 4: Model Summary and Hosmer – Lemeshow Test of Logistic Regression Enter

| Model Summary | | | Hosmer – Lemeshow Test | | |
|---|---|---|---|---|---|
| -2 log likelihood | Cox & Snell R Square | Nagelkerke R Square | Chi Square | df | Sig. |
| 400.134 | 0.600 | 0.862 | 3.275 | 9 | 0.857 |

Table 5: Coefficient, Odds ratio and p-value of parameters in Logistic Regression Enter

| Attributes | Coefficient ($\beta$) | Odds Ratio $e^{\beta}$ | p-value |
|---|---|---|---|
| Intercept | 4.7447 | 114.9733 | 0.0109 |
| Turbidity | -0.0074 | 0.9926 | 0.0554 |
| Dissolved Solids | 1.4094 | 4.0935 | 0.0258 |
| Water Level_Others | -0.6891 | 0.5020 | 0.0226 |
| pH | -0.4970 | 0.6084 | 0.0122 |
| Temperature | -0.1879 | 0.8287 | 4.0371E-6 |
| Suspended Solids | -0.0222 | 0.9780 | 1.0785E-6 |
| Chemical Oxygen Demand | -0.1121 | 0.8940 | 1.0226E-7 |
| Biochemical Oxygen Demand | -0.3897 | 0.6773 | 1.3107E-9 |
| Dissolved Oxygen | 1.5510 | 4.7162 | 0.0 |
| Ammoniacal Nitrogen (NH3-N) | -1.1726 | 0.3096 | 0.0 |

Table 6 shows the interpretation of odd ratio in Logistic Regression Enter. One Unit change increase in Turbidity, Water Level (Others), pH, Temperature, Suspended Solids, Chemical Oxygen Demand, Biochemical Oxygen Demand and Ammoniacal Nitrogen will decrease the odds of having clean water. Dissolved Oxygens have the highest influence on clean water quality because 1 Unit change increase in Dissolved Solid will increase the odds of having clean water by 471.62%. Other than that, 1 Unit change increase in Dissolved Solid will increase the odds of having clean water.

Tables 7-9 summarise reports on the Sensitivity, Specificity, Accuracy Squared Error and Misclassification Rate of the models constructed. These tables also help us find whether the model is an overfit or underfit model. Overfit model occurs when the model performs well in the training data but does not perform in the testing dataset. On the other hand, underfit model occurs when the model performs poorly on the training data but performs well in the testing data. We obtained all the values from the confusion matrix from each model.

The difference of error of terms was computed by subtracting the testing and training dataset of each model. To indicate the underfit model, both of the error terms must be negative while to indicate the overfit model we compared the models that have the highest gap between each predictive model. To obtain the best predictive model, the remaining models were compared that are neither underfit nor overfit with other predictive models.

According to Table 7, both Logistic Regression with backward and forward selection are underfitted because the models perform poorly on the training dataset while performing well in the testing dataset. Also, the difference in error terms (gap) between the testing and training datasets are both negative. The gap between the mean squared error and misclassification rate for logistic regression backward selection are -0.002 and -0.06. Forward selection has a gap value for mean squared error and misclassification rate

Table 6: Interpretation of Logistic Regression Enter

| **Attributes** | **Odds Ratio** $e^{\beta}$ | **(1-$e^{\beta}$)*100** | **Interpretation** |
|---|---|---|---|
| Turbidity | 0.9926 | 0.74% | One Unit change increase in Turbidity will decrease the odds of having clean water by 0.74%. |
| Dissolved Solids | 4.0935 | - | One mg/litre increase of dissolved solids will increase the odds of having clean water by 409.35%. |
| Water Level_ Others | 0.5020 | 49.8% | Having a water level of 'Others' will decrease the odds of having clean water by 49.8% compared to having water level of 'Normal'. |
| pH | 0.6084 | 39.16% | One Unit increase in pH level will decrease the odds of having clean water by 39.16%. |
| Temperature | 0.8287 | 17.13% | One-degree Celsius increase in Temperature will decrease the odds of having clean water by 17.13%. |
| Suspended Solids | 0.9780 | 2% | One mg/litre increase in Suspended Solids will decrease the odds of having clean water by 2%. |
| Chemical Oxygen Demand | 0.8940 | 10.6% | One mg/litre increase in Chemical Oxygen Demand will decrease the odds of having clean water by 10.6%. |
| Biochemical Oxygen Demand | 0.6773 | 32.27% | One mg/litre increase in Biochemical Oxygen Demand will decrease the odds of having clean water by 32.27%. |
| Dissolved Oxygen | 4.7162 | - | One mg/litre increase of dissolved oxygen will increase the odds of having clean water by 471.62%. |
| Ammoniacal Nitrogen (NH3-N) | 0.3096 | 69.04% | One mg/litre increase in Ammoniacal Nitrogen (NH3-N) will decrease the odds of having clean water by 69.04%. |

Table 7: Model's Performance of Logistic Regression

| **Model** | **Logistic Regression Enter** | | | **Logistic Regression Backward Elimination** | | | **Logistic Regression Forward Selection** | | |
|---|---|---|---|---|---|---|---|---|---|
| Dataset | Train | Test | Gap | Train | Test | Gap | Train | Test | Gap |
| Sensitivity (%) | 91.85 | 92.51 | - | 92.30 | 92.22 | - | 79.36 | 92.51 | - |
| Specificity (%) | 97.70 | 97.45 | - | 97.68 | 97.79 | - | 96.09 | 97.70 | - |
| Accuracy (%) | 96.41 | 96.36 | - | 96.49 | 96.55 | - | 92.38 | 96.55 | - |
| Mean Squared Error | 0.028 | 0.028 | 0.00 | 0.028 | 0.026 | -0.002 | 0.077 | 0.028 | -0.049 |
| Misclassification Rate (%) | 3.59 | 3.64 | 0.05 | 3.51 | 3.45 | -0.06 | 7.62 | 3.45 | -4.17 |

of –0.049 and –4.17. The training dataset shows a value of 92.30%, 97.68%, 96.49%, 0.028 and 3.51% for sensitivity, specificity, accuracy, mean squared error and misclassification rate, respectively. On the other hand, the testing dataset shows a value of 92.22%, 97.79%, 96.55%, 0.026 and 3.45% for sensitivity, specificity, accuracy, mean squared error and misclassification rate, respectively for logistic regression backwards. For forward logistic regression, the training dataset shows a value of 79.36%, 96.09%, 92.38%, 0.077 and 7.62% for sensitivity, specificity, accuracy, mean squared error and misclassification rate, respectively. On the other hand, the testing dataset shows a value of 92.50%, 97.70%, 96.55%, 0.028 and 3.45% for sensitivity, specificity, accuracy, mean squared error and misclassification rate, respectively.

### *Decision Tree*

Meanwhile, based on Table 8, both Decision Tree (Gini) and Decision Tree (Entropy) are overfit models because the models performed well on the training dataset but did not perform that well on the testing dataset. However, the difference of error terms (gap) between testing and training dataset of Decision Tree (Entropy) is higher than the difference of the error terms of Decision Tree (Gini). The gap of mean squared error and misclassification rate for Decision Tree (Entropy) are 2.4 and 2.67, while the gap of mean squared error and misclassification rate Decision Tree (Gini) are 0.022 and 0.023. The training dataset shows a value of 97.76%, 99.55%, 99.16%, 0.007 and 0.84% for sensitivity, specificity, accuracy, mean squared error and misclassification rate respectively. On the other hand, the testing dataset shows a value of 89.52%, 98.47%, 96.49%, 0.03 and 3.51% for sensitivity, specificity, accuracy, mean squared error and misclassification rate respectively.

### *Variable Importance using Decision Tree (Gini)*

As mentioned earlier, Decision Tree (Gini) is the best model compared to Decision Tree (Entropy). Furthermore, Table 9 shows the importance of the variables in decreasing order that was used in the best model, which is Decision Tree (Gini).

Table 8: Model's Performance of Decision Tree

| **Model** | **Decision Tree (Gini)** | | | **Decision Tree (Entropy)** | | |
|---|---|---|---|---|---|---|
| Dataset | Train | Test | Gap | Train | Test | Gap |
| Sensitivity (%) | 97.16 | 91.32 | - | 97.76 | 89.52 | - |
| Specificity (%) | 99.55 | 98.13 | - | 99.55 | 98.47 | - |
| Accuracy (%) | 99.02 | 96.62 | - | 99.16 | 96.49 | - |
| Mean Squared Error | 0.009 | 0.031 | 0.022 | 0.007 | 0.03 | 0.023 |
| Misclassification Rate (%) | 0.98 | 3.38 | 2.4 | 0.84 | 3.51 | 2.67 |

Table 9: Variable Importance based on Decision Tree (Gini)

| **Attributes** | **Weights** |
|---|---|
| Biochemical Oxygen Demand | 0.2284 |
| Chemical Oxygen Demand | 0.1471 |
| Dissolve Solid | 0.1437 |
| Suspended Solid | 0.1030 |
| Turbidity | 0.0745 |
| Ammoniacal Nitrogen (NH3-N) | 0.0358 |
| Temperature | 0.0215 |
| Dissolved Oxygen | 0.0206 |
| pH | 0.0196 |
| Water Level_Others | 0.0080 |

Variable importance refers to how much a model uses it to make accurate predictions. The table below is extracted from the best model that shows the variable importance to predict water quality.

The result shows that Biochemical Oxygen Demand (BOD) and Chemical Oxygen Demand (COD) has the highest importance in predicting the water quality, with values of 0.2284 and 0.1471, respectively. Thus, these are the variable that need to be monitored closely. This is followed by, Dissolved Solid (DS) and Suspended Solid (SS) as stated in Table 9. The least important variable is Water Level Others with a value of 0.0080.

### *Artificial Neural Network*

According to Table 10, both Artificial Neural Network with 4 and 8 hidden nodes are overfit models because the model performed well on the training dataset but did not perform that well on the testing dataset. Nevertheless, the difference of error terms (gap) between testing and training dataset of Artificial Neural Network with 8 hidden nodes is higher than the difference of the error terms of Artificial Neural Network with 4 hidden nodes. The gap of mean squared error and misclassification rate for Artificial Neural Network with 8 hidden nodes are 0.003 and 0.75, and the gap of mean squared error and misclassification rate Artificial Neural Network with 4 hidden nodes are 0.003 and 0.29. The training dataset showed a value of 89.60%, 98.13%, 96.24%, 0.029 and 3.76% for sensitivity, specificity, accuracy, mean squared error and misclassification rate, respectively. On the other hand, the testing dataset showed a value 88.92%, 97.36%, 95.49%, 0.032 and 4.51% for sensitivity, specificity, accuracy, mean squared error and misclassification rate, respectively.

### *Comparison of the Models*

To find the best model, the remaining models testing accuracy for all the potential best models were compared', which are Logistic Regression Enter, Decision Tree (Gini) algorithm and Artificial Neural Network with 4 hidden nodes. Accuracy for testing dataset for Logistic Regression Enter, Decision Tree (Gini) Algorithm and Artificial Neural Network with 4 hidden nodes are 96.36%, 96.62% and 96.02%, respectively. It shows that Decision Tree (Gini) Algorithm has the highest testing accuracy compared to the other model. Therefore, it can be concluded that the best model is Decision Tree (Gini) Algorithm. However, Logistic Regression Enter and Artificial Neural Network with 4 hidden nodes are also considered the good models as their accuracy is comparable to Decision Tree (Gini). Specifically, for Logistic Regression Enter, the training dataset shows a value of 91.85%, 97.70%, 96.41%, 0.028 and 3.59% for sensitivity, specificity, accuracy, mean squared error and misclassification rate, respectively. On the other hand, the testing dataset shows a value 92.51%, 97.45%, 96.36%, 0.028 and 3.64% for sensitivity, specificity, accuracy, mean squared error and

Table 10: Model's Performance of Artificial Neural Network

| **Model** | **Artificial Neural Network (4 Hidden Nodes)** | | | **Artificial Neural Network (8 Hidden Nodes)** | | |
|---|---|---|---|---|---|---|
| Dataset | Train | Test | Gap | Train | Test | Gap |
| Sensitivity (%) | 92.00 | 91.32 | - | 89.60 | 88.92 | - |
| Specificity (%) | 97.53 | 97.36 | - | 98.13 | 97.36 | - |
| Accuracy (%) | 96.31 | 96.02 | - | 96.24 | 95.49 | - |
| Mean Squared Error | 0.028 | 0.031 | 0.003 | 0.029 | 0.032 | 0.003 |
| Misclassification Rate (%) | 3.69 | 3.98 | 0.29 | 3.76 | 4.51 | 0.75 |

misclassification rate, respectively. For Decision Tree (Gini), the training dataset shows a value of 97.16%, 99.55%, 99.02%, 0.009 and 0.98% for sensitivity, specificity, accuracy, mean squared error and misclassification rate, respectively. On the other hand, the testing dataset shows a value of 91.32%, 98.13%, 96.62%, 0.031 and 3.38% for sensitivity, specificity, accuracy, mean squared error and misclassification rate, respectively. For Neural Network with 4 hidden nodes, the training dataset shows a value of 92.00%, 97.53%, 96.31%, 0.028 and 3.69% for sensitivity, specificity, accuracy, mean squared error and misclassification rate, respectively. On the other hand, the testing dataset shows a value of 91.32%, 97.36%, 96.02%, 0.031 and 3.98% for sensitivity, specificity, accuracy, mean squared error and misclassification rate respectively.

## Conclusion

All the predictive models (Logistic Regression Enter, Logistic Regression with forward and backward elimination selection, Decision Tree with Gini and Entropy algorithm and Artificial Neural Network with 4 and 8 hidden layers) performed very well in classifying water quality since the accuracy is high (above 90%). Although Decision Tree with the Gini Algorithm is the best model for this study, any of these models can also be used to classify water quality.

Artificial Neural Network model can be considered as a model in the classification of water quality in Selangor. It is found that missing values in the water quality data is the most prominent problem in modelling water quality, thus more research in this area is required. In this study, we did not consider the time element. Therefore, we suggest that future studies include that in their research. It is hoped that this study will open more opportunities in the classification of water quality studies. It is also recommended to check the range values of the data before comparing it with the results of this study.

## Acknowledgements

The authors would like to acknowledge Institute for Big Data Analytics and Artificial Intelligence (IBDAAI), Research Nexus UiTM (ReNeU), and Research Management Centre (RMC), Universiti Teknologi MARA, for the financial support under "Pembiayaan Yuran Penerbitan Artikel (PYPA)" provided to this research project.

## References

Ab Rashid, M. F., Abd Rahman, A., & Abdul Rashid, S. M. R. (2021). Analyzing the factors and effects of water supply disruption in Penang Island, Malaysia. *Malaysian Journal of Society and Space*, *17*(3). https://doi.org/10.17576/geo-2021-1703-05

Ahmed, A. K. A., Shalaby, M., Negim, O., & Abdel-Wahed, T. (2023). Eco-friendly enhancement of secondary effluent characteristics with air and oxygen nanobubbles generated by ceramic membrane filters. *Environmental Processes*, *10*(1). https://doi.org/10.1007/s40710-023-00628-9

Arifin, F., Robbani, H., Annisa, T., & Ma'Arof, N. N. M. I. (2019). Variations in the number of layers and the number of neurons in artificial neural networks: Case study of pattern recognition. *Journal of Physics: Conference Series*, *1413*(1), 012016. IOP Publishing.

Han, S.-H., Kim, K. W., Kim, S. Y., & Youn, Y. C. (2018). Artificial neural network: Understanding the basic concepts without mathematics. Dementia and neurocognitive disorders, *17*(3), 83. https://doi.org/10.12779/dnd.2018.17.3.83 https://www.techopedia.com/definition/5967/artificial-neural-network-ann

Li, S., Xu, D., Liu, Y., Wang, R., & Zhang, J. (2022). Identification method of influencing factors of hospital catering service satisfaction based on decision tree algorithm.

*Applied Bionics and Biomechanics*, *2022*. https://doi.org/10.1155/2022/6293908

Lorenzo, I., Serra-Prat, M., & Carlos Yébenes, J. (2019). The role of water homeostasis in muscle function and frailty: A review. *Nutrients*, *11*(8), 1-15. https://doi.org/10.3390/nu11081857

McKee, C., Harmanto, D., & Whitbrook, A. (2018). A conceptual framework for combining artificial neural networks with computational aeroacoustics for design development. *Proceedings of the International Conference on Industrial Engineering and Operations Management*, *2018-March*, 741-747.

Rosly, R., Makhtar, M., Awang, M. K., Rahman, M. N. A., & Deris, M. M. (2015). The study on the accuracy of classifiers for water quality application. *International Journal of U- and e-Service, Science and Technology*, *8*(3), 145-154. https://doi.org/10.14257/ijunesst.2015.8.3.13

Srimaneekarn, N., Hayter, A., Liu, W., & Tantipoj, C. (2022). Binary response analysis using logistic regression in dentistry. *International Journal of Dentistry*, *2022*(1), 1-7. https://doi.org/10.1155/2022/5358602

Sweeney, C., Ennis, E., Mulvenna, M., Bond, R., & O'neill, S. (2022). How machine learning classification accuracy changes in a happiness dataset with different demographic groups. *Computers*, *11*(5). https://doi.org/10.3390/computers11050083

Xu, X., Xia, L., Zhang, Q., Wu, S., Wu, M., & Liu, H. (2020). The ability of different imputation methods for missing values in mental measurement questionnaires. *BMC Medical Research Methodology*, *20*(1), 1-9. https://doi.org/10.1186/s12874-020-00932-0

Zainudin, Z. (2010). *Benchmarking river water quality in Malaysia*. Jurutera. http://irep.iium.edu.my/2954/1/Feature-BenchmarkingRiverWater3pp.pdf