Journal of Hydrology 575 (2019) 148–165

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Towards a time and cost effective approach to water quality index class prediction

Jun Yung Ho[a], Haitham Abdulmohsin Afan[a,*], Amr H. El-Shafie[b], Suhana Binti Koting[a], Nuruol Syuhadaa Mohd[a], Wan Zurina Binti Jaafar[a], Lai Sai Hin[a], Marlinda Abdul Malek[c], Ali Najah Ahmed[c,d], Wan Hanna Melini Wan Mohtar[e], Amin Elshorbagy[f], Ahmed El-Shafie[a]

[a] Department of Civil Engineering, Faculty of Engineering, University Malaya, Malaysia
[b] Civil Engineering Department, El-Gazeera High Institute for Engineering, Al Moqattam, Cairo, Egypt
[c] College of Engineering, University Tenaga Nasional, Malaysia
[d] Institute for Energy Infrastructures, University Tenaga Nasional, Malaysia
[e] Department of Civil & Structural Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, Bandar Baru Bangi, Malaysia
[f] Department of Civil, Geological, & Environmental Engineering, Global Institute for Water Security (GIWS), University of Saskatchewan, Saskatchewan, Canada

ARTICLE INFO

This manuscript was handled by A. Bardossy, Editor-in-Chief, with the assistance of Purna Chandra Nayak, Associate Editor

Keywords:
River water quality
Water quality index
Prediction model
Decision tree model

ABSTRACT

The development of water quality prediction models is an important step towards better water quality management of rivers. The traditional method for computing WQI is always associated with errors due to the protracted analysis of the water quality parameters in addition to the great effort and time involved in gathering and analyzing water samples. In addition, the cost of identifying the magnitude of some of the parameters through experimental testing is very high. The water quality of rivers in Malaysia is ranked into five classes based on water quality index (WQI). WQI is function of six water quality parameters: ammoniac nitrogen (NH_3-N), biochemical oxygen demand (BOD), chemical oxygen demand (COD), dissolved oxygen (DO), pH, and suspended solids (SS). In this research, the decision tree machine learning technique is used to predict the WQI for the Klang River and its classification within a specific water quality class. Klang River is one of the most polluted rivers in Malaysia. Modeling experiments are designed to test the prediction and classification accuracy of the model based on various scenarios composed of different water quality parameters. Results show that the proposed prediction model has a promising potential to predict the class of the WQI. Moreover, the proposed model offers a more efficient process and cost-effective approach for the computation and prediction of WQI.

1. Introduction

Water managers need to manage river water quality considering that most daily water supply is sourced from rivers, especially in Malaysia. The deteriorating river water quality has an impact on river health, including fluvial ecology, which increases the risk to human health and the challenges to ensure sustainable production of drinking water. The development of water quality prediction models is an important step towards better water quality management of rivers. During the last several decades, efforts have been made to develop accurate prediction models for water quality parameters by utilizing different modelling methods (Chau, 2006; Manache and Melching, 2008; Singh et al., 2011). Researchers, among others, have given special attention to the artificial intelligence modelling methods (Maier and Dandy, 2000).

Artificial Neural Networks (ANNs) have been used with success to predict water quality parameters, such as Dissolved Oxygen (DO), Biological Oxygen Demand (BOD), and Chemical Oxygen Demand (COD) in waterways. Since DO is considered to be the most important parameter in river water quality, numerous studies were conducted to predict DO concentration based on different parameters by using various ANN models. Sarkar and Pandey (2015) studied the development of feedforward, back propagation ANNs to simulate DO concentration in the Yamuna River. The researchers found that in order for their ANN model to achieve the best performance, an optimal number of input variables have to be fed into the model. The complexity of a model increased with higher number of input variables, and the performance drastically deteriorated with the over-reduction of the model input (Afan et al., 2017; Sarkar and Pandey, 2015).

A study conducted by Csábrági et al. (2017) compared the performance of different ANNs in predicting DO concentration. The study compared the performance of multivariate linear regression (MLR) and

* Corresponding author.
E-mail address: haitham.afan@gmail.com (H.A. Afan).

https://doi.org/10.1016/j.jhydrol.2019.05.016
Received 27 October 2018; Received in revised form 19 March 2019; Accepted 5 May 2019
Available online 07 May 2019

List of symbols and abbreviations

AN/NH_3-N	Ammoniacal nitrogen
ANFIS	Adaptive neuro-fuzzy inference system
ANN	Artificial neural network
API	Air pollutant index
ASTM	American Society for Testing and Materials
BOD	Biochemical oxygen demand
BPNN	Back propagation neural network
COD	Chemical oxygen demand
DO	Dissolved oxygen
DOE	Department of Environment
FFNN	Feed forward neural network
FWQ	Fuzzy water quality
GRNN	General regression neural network
IDTL	Improved decision tree
INWQS	Interim National Water Quality Standards
km	Kilometre
km^2	Squared kilometre
m^3	Cubic metre
MANFIS	Modified adaptive neuro-fuzzy inference system
mg/l	Milligram per litre
MLD	Millions of litres per day
MLPNN	Multilayer propagation neural network
MLR	Multivariate linear regression
NO_2	Nitrogen dioxide
pH	Potential of hydrogen
PM_{10}	Particulate matter 10 μm or less in diameter
pred.	Predicted
r^2	Coefficient of determination
R^2	Correlation of coefficient
RBFNN	Radial basis function neural network
RMAE	Root mean absolute error
RMSE	Root mean square error
RoL	River of Life
SNR	Signal to noise ratio
SO_2	Sulphur dioxide
sq. miles	Squares miles
SS	Suspended solids
TDS	Total dissolved solids
WQI	Water quality index

three other ANN models; multilayer propagation neural networks (MLPNNs), radial basis networks (RBFNNs), and general regression neural networks (GRNNs). The result of the study indicated that the major drawback of MLPNNs is the need for multiple runs to avoid the possibility of a misleading outcome of a single run. The study concluded that GRNNs gave the best DO prediction in contrast to MLPNNs, RBFNNs, and MLR (Csábrági et al., 2017). These studies showed that not only are ANNs able to predict water quality parameters, but they also provide recommendations for modification and improvement of their drawbacks and prediction performance, as was shown in the GRNN.

Najah et al. (2012) conducted a study on water quality prediction by using integrated wavelet-ANFIS model. Three water quality parameters, i.e., Total Dissolved Solids (TDS), electrical conductivity, and turbidity of the Johor River, were used in the study. ANFIS was modified using a wavelet de-noising technique to reduce the complex uncertainty induced noise, which allows the model to produce superior result as compared to those obtained when using MLPNNs and ANFIS where the mean absolute error in percentage could be as low as 0.01 (Najah et al., 2012).

These studies have shown the degree to which machine learning (ML) models can be used to predict water quality parameters even though they have a complex and non-linear computational method and are stochastic in nature. In fact, the flexibility of ML models makes it possible to develop a better and more effective models to deal with the difficulties in monitoring water quality parameters. These studies however, focused on the prediction of a single water quality parameter rather than focusing on the prediction of water quality index (WQI). However, several water quality parameters have to be monitored/analysed in order to obtain the required information to estimate of WQI. As a matter of fact, the process of obtaining such predefined water quality parameters is time-consuming and very costly. In this regard, the present study will focus on developing a prediction model for WQI, which requires less water quality parameters and therefore reduce the time required to perform the analysis and minimize the cost required in order to achieve the desired WQI for rivers based on the conditions in Malaysia. The authors believe that this model could potentially be generalized for application to several river conditions worldwide.

The computation of the water quality index (WQI) in Malaysia, which involves a series of sub-index calculations, is lengthy and complicated, and is often associated with errors during the computation processes. There are complex and non-linear relationships between the WQI and the water quality parameters. Furthermore, some of the parameters require exhaustive sample collection campaign that is time-consuming, and must be conducted by skilled technicians to ensure accurate sample analysis and data representation. Even with advanced equipment and technology, high operative and management cost hinders a comprehensive spatial and temporal monitoring of river water quality. Hence, there is a need to develop a data driven model with a high capability in order to simplify the processes, reduce the errors, and reduce the need for costly and time-consuming sampling and lab analysis.

Decision tree is a popular machine learning tool and is often used to identify possible consequences such as the chance event outcomes, investment risks, decision making, and interest rate. This classifier modelling method showed an outstanding performance even when used with a complex dataset to identify its pattern behaviour (Everaert et al., 2016).

The main objective of this research is to develop a model for predicting the class of WQI which indicate the status of river water quality. The decision tree modelling has been utilized as a predictor for the WQI class. The proposed decision tree model will be evaluated under three different scenarios by utilizing the water quality data for the Klang River since it is the most polluted river in Malaysia (Othman et al., 2012).

2. Materials and method

2.1. Study area

The Klang River runs through urban and developed areas, cutting right through the middle of the Federal Territory of Kuala Lumpur while the upstream and downstream parts of the river flow through the state of Selangor. It is situated between latitudes 2°55′N and 3°25′N and longitude 101°15′E and 101°55′E. Geographically, the Klang River begins at an altitude of about 1200 m on the western slopes of Peninsula Malaysia and runs south-westwards to join the Gombak River in the centre of Kuala Lumpur. The Klang River is almost 120 km long and its drainage basin area is around 1260 km^2. The population in the Klang River basin is projected to reach 10 million by the year 2020 from the current estimated population of 7.2 million. Several Water Treatment Plants (WTPs) extract water from the tributaries within the Klang River basin, i.e., Bukit Nanas and Wangsa Maju with a design capacity of 145 MLD and 45 MLD, respectively (Mohamed et al., 2015). However, it should be noted that, despite the huge basin area, most of the treated water for Kuala Lumpur is supplied by the neighbouring state of Selangor. Fig. 1 shows the Klang River basin.

The Department of Environment (DOE) has setup numerous water

quality stations along the main stream of the Klang River, which are operated by the DID, to facilitate the monitoring and management of water resources. The six parameters (as presented in Eq. (1)) were measured, and additional assessment on conductivity, turbidity, salinity, temperature, microbial, nutrients, and heavy metals concentration are made when there is a necessity for it (Sharif et al., 2015).

The Klang River Basin lies in an area with humid tropical climate that is characterised by heavy rainfall, uniform temperature, and high humidity. In the east, at the foothill, the average annual rainfall is almost 2600 mm and the amount decreases to around 1900 mm at the coast. Peak rainfall occurs during two two-month periods, i.e. from April to May and October to November during the Southwest and Northeast Monsoons, respectively. The average monthly humidity is between 80 and 85%; the average monthly temperature in the basin ranges between 26 and 28 °C; the daily sunshine duration is between 4.5 and 7.0 h/day; and the daily evaporation ranges between 3 and 5 mm/day (El-Shafie et al., 2012).

The land use in the Klang Valley basin area is diverse, ranging from tropical forests in the headwater region to urban areas with their associated activities in the central region, and agriculture in the fringes of the basin. About 41% of the basin is agricultural area, 29% is urban, 25% is forests and swamps, and tin mines make up 5% of the total basin area. The urban area consists of recreational, residential, industrial, institutional and commercial zones.

The Klang River basin has been experiencing serious environmental degradation and flooding as a result of continuous development, industrialization, and population growth. The rapid development in the area is expected to further increase the probability of water stress condition with regard to future water supply. A report published by the DID in 2011 projected that the water demand in Kuala Lumpur Gombak, Petaling and Klang will increase to 1194 million liter per day (MLD) for the period between 2015 and 2050. The Klang River is currently critically polluted due to inefficient water quality management. Improperly treated sewage and industrial and residential discharges from the Klang River Basin flow into the river; this problem is compounded by high events of soil erosion as a result of inefficient control plan at construction sites.

2.2. Data collection

The data collected from the 15 automatic monitoring stations for the six parameters during the period between 2000 and 2010 for Klang River are demonstrated in Fig. 2. Fig. 2 shows the distribution of the six water quality parameters measured between the years 2000 and 2010. The DO parameter gives a direct assessment of river health. Fig. 2a shows that the DO data contains a significant fraction of low DO concentration, even up to < 1 mg/L, especially during the period between 2000 and 2002. The DO concentration improved substantially in 2003 with an average value of 4 mg/L. Despite the fluctuation in the value of DO concentration, which occasionally reached the desirable Class I category, the Klang River is in fact a class III river. This is supported by the average value of COD and BOD concentrations of 45 and 7 mg/L, respectively. The spikes observed in Fig. 2b (which could peak to > 50 mg/L for BOD) is believed to be due to illegal sewage discharges. Even though the Government has imposed strict penalties on those who cause pollution, the large area of the Klang River basin makes it possible for blind spot discharge locations to exist, thus allowing the polluters to avoid being caught by the authorities. High concentration of TSS was observed in the first few years of the current millennium, with the highest measured TSS of 1400 mg/L (Fig. 2d). The rapid development taking place in recent years in Kuala Lumpur, coupled with the lack of training in the proper implementation of Erosion and Sediment Control Plan (ESCP), are believed to be a major factor contributing to the high levels of sedimentation in the Klang River. The level of ammonia nitrogen has clearly exceeded the value stipulated in the guideline, with the maximum measured value being 20 times higher than the value set for Class III. The mean NH_3-N value of 4 mg/L put the Klang River in Class V category, hence posing serious ecological risk. Fig. 2e shows that the pH ranges between 5.5 and 8, which is within the acceptable range for Class II and III stipulated in the water quality standard of rivers (Table 2).

The water quality indexes during the study period fluctuate widely between the year 2000 and mid-2005. The classification of Klang River varied between Class I and Class V (with an average of Class III) throughout this period. In 2006, the state of the Klang River has improved significantly and a smaller fluctuation was observed in the water quality index. Despite of this, the Klang River is still categorized as a Class III river. The temporal data of the WQI showed that the Klang River was categorized as Class I only once (0.01 percent of the time); Class V twenty times (0.2% of the time; and class II 66 times (6.9% of the time) during the study period. During the period between the year 2000 and 2010, classifications activities were carried out each two months for the Klang River. The results showed that Klang River mostly was categorized as class III and class IV, 46% and 45% of the time, respectively.

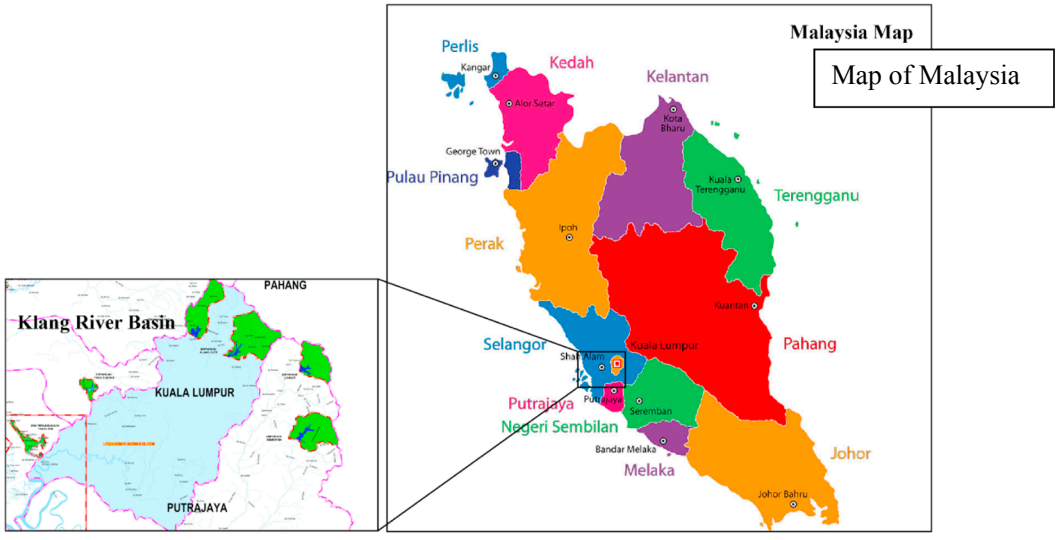

Fig. 1. Location of Klang River Basin, Malaysia.

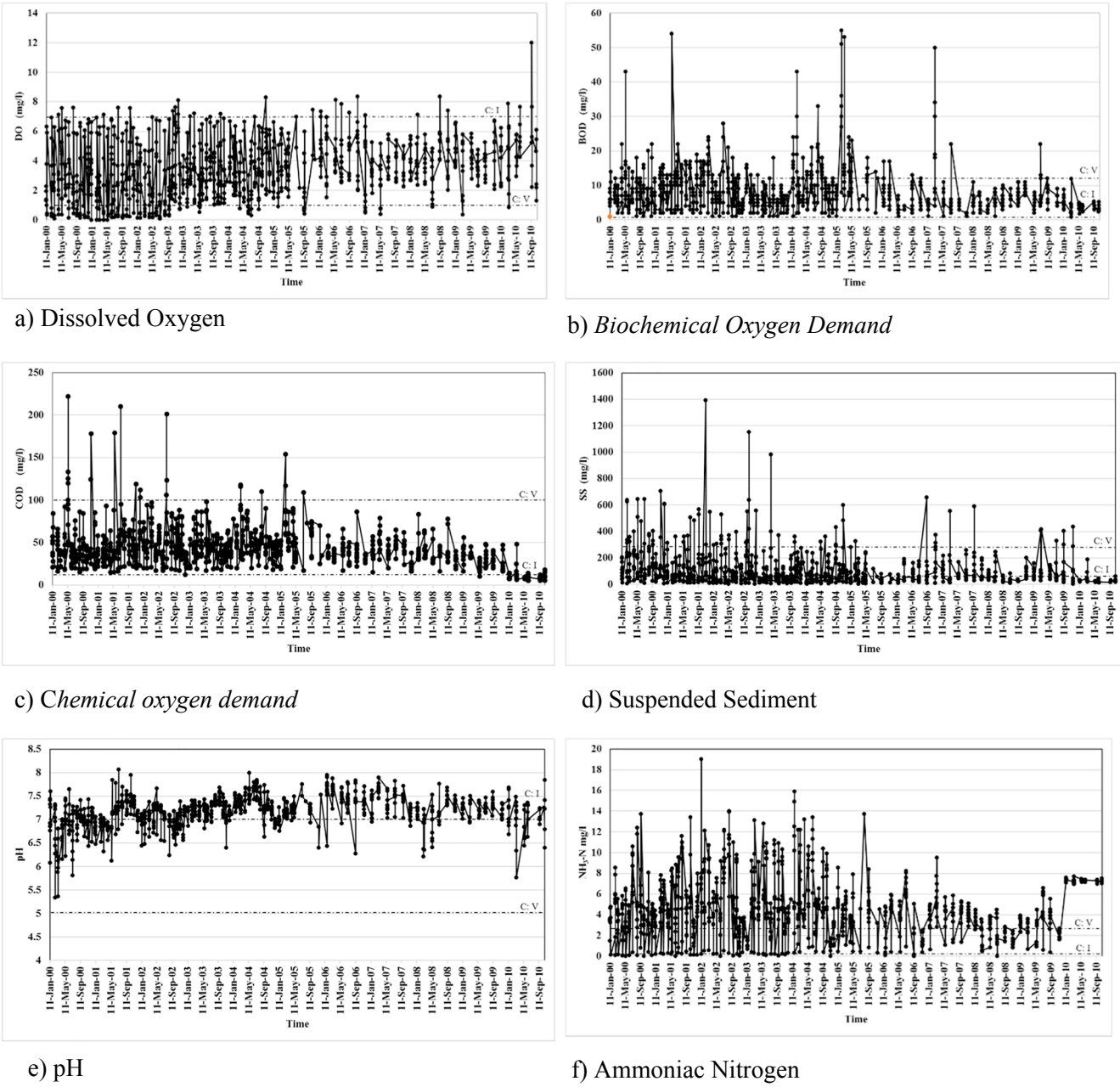

Fig. 2. 10 years distribution for a) Dissolved Oxygen, b) Biochemical Oxygen Demand, c) Chemical oxygen demand, d) Suspended Sediment, e) pH, and f) Ammoniac Nitrogen, during the period between 2000 and 2010 for Klang River.

2.3. Methodology

2.3.1. Determination of WQI

In 1978, the Department of Environment (DOE) in Malaysia introduced the water quality index (WQI) and started to monitor river water quality. The aim of this initiative was to establish a baseline monitoring system for river water quality, to detect any changes in water quality, and to identify the sources of pollution so that immediate actions could be taken to mitigate the pollution. Since then, a total of 1064 manual monitoring sites have been set up in 143 river basins in Malaysia (DOE, 2007). The formula for calculating WQI was proposed by the DOE and a panel of experts was consulted on the choice of parameters and the relative weight to be assigned to each parameter (Hameed et al., 2016).

The WQI is computed based on six water quality parameters, i.e. biological oxygen demand (BOD), chemical oxygen demand (COD), dissolved oxygen (DO), suspended solids (SS), pH, and ammoniac nitrogen (NH_3-N). The formula for computing the WQI is given by Eq. (1), and Table 1 presents the formula used to calculate the sub-indexes (SI). Water quality is ranked into Class I, II, III, IV, and V, based on the WQI and the Interim National Water Quality Standards for Malaysia (INWQS) as shown in Table 2 (DOE, 2007). Currently, there are 15 automatic monitoring stations, which continuously monitor the water quality in the Klang River.

$$WQI = 0.22\ SIDO + 0.19\ SIBOD + 0.16\ SICOD + 0.16\ SISS + 0.15\ SIAN + 0.12\ SIp \quad (1)$$

2.3.2. Decision tree modelling approach

Decision tree model is one of the most frequently used techniques in data mining. It is a popular machine learning tool and is often used to identify possible consequences, such as the chance of event outcomes,

Table 1
Sub-index calculation formula for WQI Malaysia.

Parameter	Value	Sub-index equation
DO (in % saturation)	$x \le 8$	SIDO = 0
	$x \ge 92$	SIDO = 100
	$8 < x < 92$	$SIDO = -0.395 + 0.030x^2 - 0.00020x^3$
BOD	$x \le 5$	$SIBOD = 100.4 - 4.23x$
	$x > 5$	$SIBOD = (108e^{-0.055x}) - 0.1x$
COD	$x \le 20$	$SICOD = -1.33x + 99.1$
	$x > 20$	$SICOD = (103e^{-0.0157x}) - 0.04x$
SS	$x \le 100$	$SISS = (97.5e^{-0.00676x}) + 0.05x$
	$100 < x < 1000$	$SISS = (71e^{-0.0061x}) - 0.015x$
	$x \ge 1000$	SISS = 0
NH_3-N	$x \le 0.3$	$SIAN = 100.5 - 105x$
	$0.3 < x < 4$	$SIAN = (94e^{-0.573x}) - 5(x-2)$
	$x \ge 4$	SIAN = 0
pH	$x < 5.5$	$SIpH = 17.2 - 17.2x + 5.02x^2$
	$5.5 \le x < 7$	$SIpH = -242 + 95.5x - 6.67x^2$
	$7 \le x < 8.75$	$SIpH = -181 + 82.4x - 6.05x^2$
	$x \ge 8.75$	$SIpH = 536 - 77.0x + 2.76x^2$

Table 2
DOE water quality index classification.

Parameter	Unit	Class				
		I	II	III	IV	V
NH_3-N	mg/l	< 0.1	0.1–0.3	0.3–0.9	0.9–2.7	> 2.7
BOD	mg/l	< 1	1–3	3–6	6–12	> 12
COD	mg/l	< 10	10–25	25–50	50–100	> 100
DO	mg/l	> 7	5–7	3–5	1–3	< 1
pH	–	> 7	6–7	5–6	< 5	> 5
SS	mg/l	< 25	25–50	50–150	150–300	> 300
WQI	–	< 92.7	76.5–92.7	51.9–76.5	31.0–51.9	< 31.0

investment risks, decision making, and interest rate (Azad and Moshkov, 2017; Khosravi et al., 2018). The ID3 core algorithm in the decision tree model employs a top-down, greedy search through the space of possible branches with no backtracking, and allows the handling of both categorical and numerical data. Decision tree models are able to include predictors with dependence assumptions between the predictors (Brown and Myles, 2013). Decision tree-works with a tree structure, building classification and regression models. When a dataset is fed into this model as an input layer, the system breaks down the complex dataset into small subsets while at the same time building a decision tree model by analysing those data (see Fig. 3).

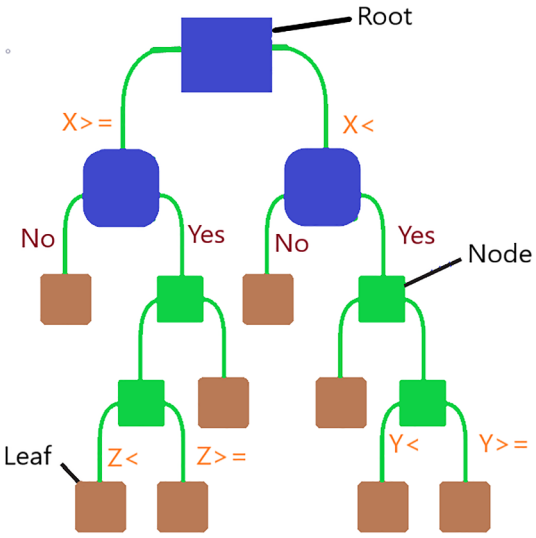

Fig. 4. Typical structure of decision tree.

The basic idea of decision trees is known as the divide-and-conquer technique. The dataset is broken down into different parts in each step, and each part is supposed to better represent one of the possible classes of the data. The result is a tree structure where each inner node represents a test for the value of an attribute and each leaf represents the decision for a particular class. A new and unknown case is then routed down the tree until it reaches one of the leaves (Brown and Myles, 2013). Fig. 4 presents the structure of a decision tree model where the decision and the leaf nodes are represented by squares and circles, respectively. Each node has two options based on the value type of the feature used at this node. For nominal features, the number of children is usually equal to the number of possible values of this feature. By using a nominal feature for a test in one of the inner nodes, the dataset at this stage is basically divided based on the different values of this feature. Hence, a nominal feature will not be tested more than once since all examples further down the tree will have the same value as this particular feature. This is different for numerical

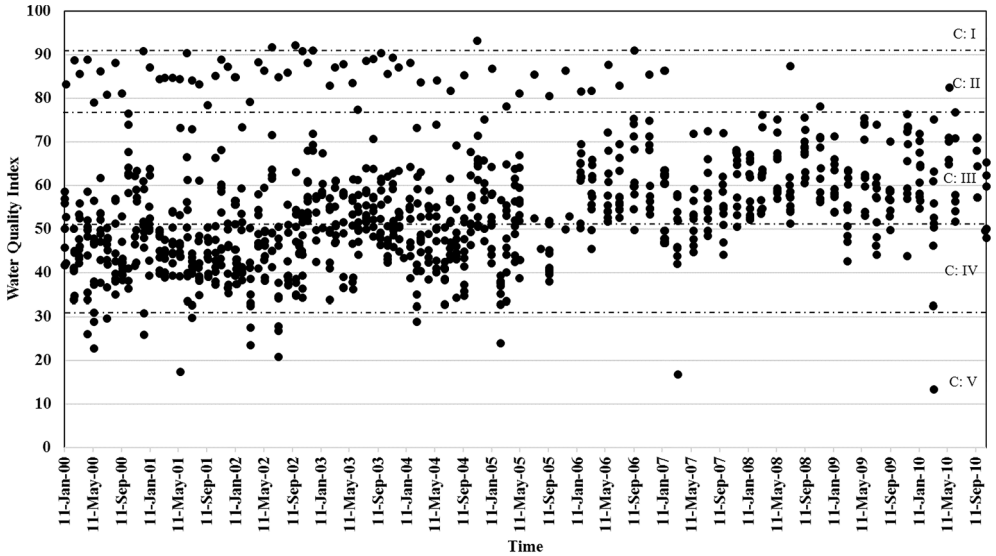

Fig. 3. Water Quality Index for Klang River during the period between years 2000 and 2010. Frequency of class classification (presented in bracket) for Class I (0.001), Class II (0.069), Class III (0.46), Class IV (0.45), and Class V (0.02).

attributes where test is performed if the attribute value is greater or less than a determined constant. The attribute could be tested several times for different constants.

A decision tree is constructed top-down in a recursive divide-and conquer mode. At the first stage, the feature for the root node is selected. Then a branch for each possible feature value is formed and the cases are split into subsets based on the possible values. Finally, these steps are repeated recursively for each branch using only cases that reach the branch. The process will stop when all instances have the same class. Among the key advantages of a decision tree model are the ease of understanding and interpreting the results and the possibility of adding new scenarios, which helps to predict an unpredicted outcome.

2.3.3. Model configuration

The predictive modelling for estimating the WQI for the Klang River is investigated based on six predictors, i.e. BOD, COD, DO, SS, pH, and NH_3-N. These input variables were normalised using the Z-transformation before being fed into the model. The normalisation subtracts the mean of the data from all values and divides them by the standard deviation. Hence, the distribution of the data has a mean of zero and a variance of one. The purpose of Z-transformation is to ensure the preservation of the original data distribution and to ensure that the modelling is not affected by outliers (Kotu and Deshpande, 2014).

In this study, data from the years 2001 to 2009 were used for training and validation (95%), whilst the data for the year 2010 (5%) were kept for testing. Fig. 5 illustrates the stages and sections of the modelling. Several parameters have to be set when using decision tree as the predictive modelling in the RapidMiner Studio (Hofmann et al., 2013). This software has several parameters to be configured, such as the criterion where it is set to “gain ratio”. Gain ratio is a splitting feature that adjusts the information gain for each feature to allow for the breadth and uniformity of the feature values. Maximum depth is the parameter used to restrict the depth of the decision tree, where this value is set to (−1) in order to not impose any bound on the depth of the tree. Another important parameter used for the pessimistic error calculation of pruning is the confidence level, which is set to be 0.25.

The present study employs the pre-pruning parameters, which represent the stopping criteria. The pre-pruning is represented by minimal gain and was set to be 0.01, minimal leaf size is two, minimal split size is two, and the number of pre-pruning alternative is two. These configurations were set to ensure that the modelling is able to maximise the analysis of parameters correlated with the WQI in order to obtain a more accurate prediction (Kotu and Deshpande, 2014). Fig. 5 and Fig. 6 show the model configuration in the RapidMiner Studio. These figures show the flow work of model inside the RapidMiner interface which is represented by two main stages training and testing.

2.3.4. Modeling scenarios

During the setting up of the modelling, it was conducted with all six water quality parameters as input variables to determine the model performance. Since the main purpose of this research is to use the decision tree to predict the WQI based on a smaller number of parameters as inputs to achieve higher effectiveness and efficiency in determining the WQI of rivers, this predictive modelling focused more on prediction accuracy based on lower number of inputs.

Each scenario was developed by reducing the number of water quality parameters (as model input parameters) to five, four, and three instead of the six parameters stipulated in the DID Manual. The successful implementation of the proposed decision tree model with a minimal number of model input variables would result in minimal cost of WQI prediction for river water quality. This could also reduce the time taken to analyse the water sample in the laboratory to determine the value of the cut-off parameters. Furthermore, these scenario analyses would also allow for the identification of the correlation between water quality parameters and the WQI classes.

In the first scenario, five parameters were used as inputs and one parameter was omitted for each case. The second scenario has four inputs, thereby providing fifteen possibilities with various configuration of water quality parameters. In the third scenario, the number of neurons was decreased to three, and thus, increasing the total number of experiments

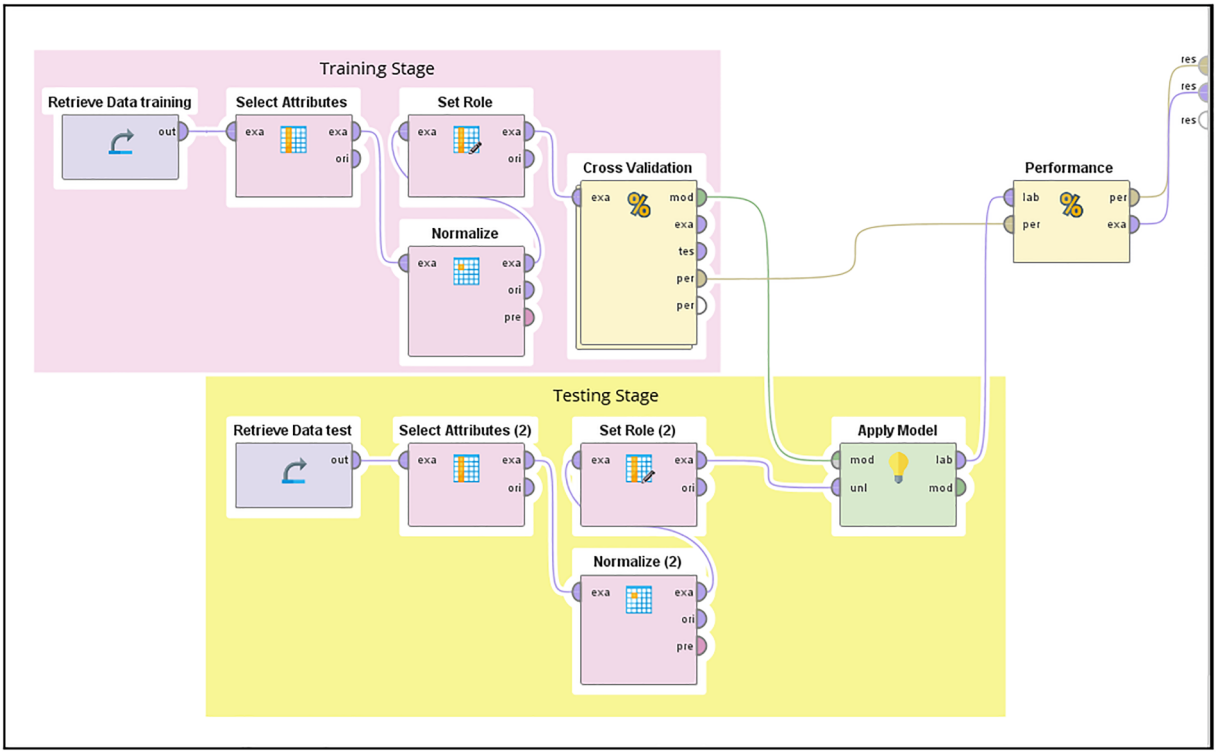

Fig. 5. Model configuration and stages in RapidMiner Studio software. The abbreviations shown are the common RapidMiner commands: exa, mod, ori, lab, tes, per, res denote examples, model, original, laboratory, test, performance and residual, respectively.

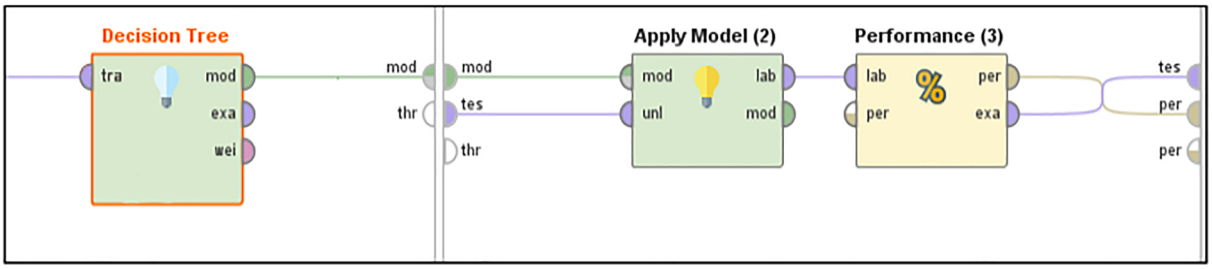

Fig. 6. Model configuration within cross validation in RapidMiner Studio. Note, refer to Fig. 5 for abbreviations.

within this scenario to 20 with more possibilities of input combinations. Tables 3–5 show all combinations of input data for the three scenarios. The monthly river water quality data for the Klang River basin between 2001 and 2010 were Z-transformed to speed up the training process and reduce the influence of outliers in the dataset. This decision tree modelling was run with infinity depth, confidence level of 0.25, and minimal gain of 0.1 to ensure that the model is able to achieve a higher degree of analysis accuracy. The optimal architecture was determined based on the prediction accuracy with a minimum benchmark of 75%.

2.3.5. Model performance criteria

The accuracy of the predictive modelling was evaluated based on three criteria in order to measure the performance of the model. The measurement of performance does not specifically look at the difference between the traditionally computed and the predicted WQI values. Instead, the investigation focused on the similarity of class between the predicted and the WQI values obtained using Eq. (1). Classical statistical measures such as, mean absolute error, maximum error and mean square error are not required for the comparison of error between the actual and the predicted WQI values. For example, the range for Class IV WQI is set between 31.0 and 51.9. If the computed value, using the WQI equations, is 50 (which indicates Class IV) and the predicted value is 32 (which falls under the Class IV), then the model is deemed to be acceptable even though the difference between 50 and 32 is rather high and has a poor prediction error in terms of the classical error measure evaluation. On the other hand, if the measured WQI is 30 (Class V) and the predicted WQI is 32 (Class IV), then the model has made an inaccurate prediction of the WQI class even though the difference between the two values is small.

In classification or class prediction, it is essential to evaluate the model performance at the overall level, but also at each class individually. Therefore, there is a need to examine all the possible cases that designate the relationship between the predicted WQI class as the model output and the actual WQI class. In fact, there are four cases that could occur and should be taken into account to measure the model performance at each class individually. The first case is a true positive, which is *correctly identifying* the WQI class. In this case, the model result is one that detects the class condition when the condition is present. The second case is the true negative test that the result is one that does not detect the condition when the condition is absent, in other words, the model output result *correctly rejects* the WQI class. It should be noted that the second case condition did not occur in this study as WQI class should fall in one of the five classes and never be in neutral condition. The third case of condition is the false positive (*incorrectly identified condition*) which is experienced when the test result is one that detects the condition when the condition is absent. Finally, the fourth case is the false negative test result is one that does not detect the condition when the condition is present, this case represents the *incorrectly rejected condition*.

Let TP denotes the number of true positives, TN the number of true negatives, FP the number of false positives, and FN the number of false negatives as shown in Fig. 7. The overall prediction accuracy for the model could be calculated as the sum of TP for all classes (the summation of all number in the yellow cells) divided by the total number of the tested data (44 in our study).

In order to carry out the model performance analysis at the class level to examine the performance of the model at each class individually, the following evaluation measures could be used. Sensitivity measures (Class recall) the ability of a test to detect the condition when the condition is present. Thus, Sensitivity = TP/(TP + FN). Predictive value positive (Class precision) is the proportion of positives that correspond to the presence of the condition. Thus, Predictive value positive = TP/(TP + FP).

The performance vector was calculated based on the confusion matrix shown in Table 6. The model performance was evaluated using three types of assessment, including prediction accuracy, class precision, and class recall. Equation 2 was used to estimate the prediction accuracy, which is defined as the ability of the classifier to select all cases that need to be selected and reject all cases that need to be rejected. For a classifier with 100% accuracy, this would imply that false negative = false positive = 0. Precision is defined as the proportion of cases found that were actually relevant based on the calculation made using Eq. (3). Finally, the recall is expressed by Eq. (4), which is defined as the proportion of the relevant cases that were actually found among all the relevant cases.

The formulae for performance criteria are expressed as follows:

$$Prediction\ accuracy_{all\ classes} = \frac{\sum_{Class=1}^{n} True\ positive}{Total\ Number\ of\ testing\ data} \times 100\%\ \text{n} = 1, 2, 3. ..5 \quad (2)$$

$$Class\ precision_{class\ n} = \frac{True\ positive_{class\ n}}{True\ positive_{class\ n} + False\ positive_{class\ n}} \times 100\% \quad (3)$$

$$Class\ recall_{class\ n} = \frac{True\ positive_{class\ n}}{True\ positive_{class\ n} + False\ negative_{class\ n}} \times 100\% \quad (4)$$

3. Results and discussion

The proposed decision tree model with the three different scenarios was developed. In order to evaluate each possible input combination for scenario, 6, 15, and 20 possible combination cases were assessed for the three scenarios, respectively. The detailed results for each scenario are presented separately in Appendix A, B, and C for first, second and third scenario, respectively, but the main findings are presented below.

Table 3
Input data with five water quality parameters, with the left-out parameter marked with "X".

5 Parameters Scenario	Water Quality Parameter					
	BOD	COD	SS	DO	pH	NH_3-N
1-1						X
1-2					X	
1-3				X		
1-4			X			
1-5		X				
1-6	X					

Table 4
Input data with four water quality parameters, with the left-out parameters marked with "X".

4 Parameters Scenario	Water Quality Parameter					
	BOD	COD	SS	DO	pH	NH_3-N
2-1					X	X
2-2			X			X
2-3				X		X
2-4		X				X
2-5	X					X
2-6			X		X	
2-7				X	X	
2-8		X			X	
2-9	X				X	
2-10			X	X		
2-11		X		X		
2-12	X			X		
2-13		X	X			
2-14	X		X			
2-15	X	X				

Table 5
Input data with three water quality parameters, with the left-out parameters marked with "X".

3 Parameters Scenario	Water Quality Parameter					
	BOD	COD	SS	DO	pH	NH_3-N
3-1				X	X	X
3-2			X		X	X
3-3		X			X	X
3-4	X				X	X
3-5			X	X		X
3-6		X		X		X
3-7	X			X		X
3-8		X	X			X
3-9	X		X			X
3-10	X	X				X
3-11			X	X	X	
3-12		X		X	X	
3-13	X			X	X	
3-14		X	X		X	
3-15	X		X		X	
3-16	X	X			X	
3-17		X	X	X		
3-18	X		X	X		
3-19	X	X		X		
3-20	X		X			

3.1. First scenario: Input data with five water quality parameters

In the first scenario, the prediction model was run based on five different water quality parameters with six different combinations as shown in Table 7. Data shows that scenario 1-1 (i.e. without NH_3-N) produced the best result with an accuracy of 84.09%. As can be seen in Table A.1, the best performance for this combination is supported by the class recall for true III, true IV, and true V (which achieved 84.85%, 100.00%, and 100.00%, respectively), and in class precision for predicted III, predicted IV and predicted V (93.99%, 61.54%, and 100.00% respectively). Table A.2 shows that although not all data were accurately predicted, the false predicted data in Test 1-1 were still close to its original class, for example, two data for true II fall in the predicted III, and five data for true III fall in the predicted IV.

In this scenario, five other tests were not able to achieve an accuracy of at least 75.00%. Interestingly, the model with these configurations was not able to accurately predict the most critical Class V. Tables A.3–A.6 show that all the data for true V fall in the predicted III, which is a wide difference between the true and the predicted values. Class V is the lowest class in the WQI and it is very crucial for this class to be correctly identified so that immediate actions can be taken to reduce the effect of the unacceptable limits for the quality of water on the environment and human use.

Based on the analysis, the influence of each water quality parameter was investigated in this predictive modelling. Result show that NH_3-N has the least effect on the predicted WQI, with notable high accuracy (84.09%) for Test 1-1. This is a somewhat positive outcome considering the high cost for experimental analysis of NH_3-N. The cost for laboratory analysis could be reduced by omitting theNH_3-N parameter from the modelling input variables.

3.2. Second scenario: Input data with four water quality parameters

The main objective for carrying out scenario II is to improve model performance by reducing the number of input variable for the model. By using only four water quality parameters, different combinations of inputs as shown in Table 8 were tested. The accuracy of this modelling was determined based on the percentage of prediction accuracy, followed by the percentage of class precision and class recall. The 15 combinations for water quality parameters are given in Appendix B. Scenario 2-1 provides the best performance with a prediction accuracy of 81.82%, with BOD, COD, SS, and DO as the four inputs. It should be noted that, besides this combination, scenarios 2-7 and 2-15 were able to achieve a prediction accuracy benchmark of 75.00%. Interestingly, the same parameters in both tests are NH_3-N and SS.

Table B.1 shows that scenario 2-1 achieved a class precision of 93.10% for predicted III, 57.14% for predicted IV, and 100.00% for predicted V. The achievement of class recalls for true III, true IV, and true V were 81.82%, 100.00%, 100.00%, respectively. In contrast to scenario 1-1, scenario 2-1 omitted pH from the inputs of the modelling, and still showed a promising result for this predictive modelling (36 out of 44), which is very close to the prediction accuracy of scenario 1–1 (37 out of 44). The results for scenarios 2-7 and 2-15 are shown in Table B.7 and Table B.15, respectively. Even though both scenarios achieved a prediction accuracy of 75.00%, they were not able to predict the most critical Class V. The results for both scenarios for true V fall in the predicted III, which is a wide deviation from its original class. Hence, both input combinations for scenarios 2-7 and 2-15 are not qualified for this predictive modelling (see Tables B.2–B.6, B.10–B.14 and C1).

The results of the analysis in the second scenario further proved that NH_3-N is the least effective parameter correlated to the prediction of WQI. Additionally, pH has been shown to be a parameter with low correlation (to a certain extent) to WQI prediction. In Malaysia, the pH of the discharge into rivers is strictly monitored and is the easiest parameter to be monitored as its measurement can be easily obtained in-situ and without further laboratory analysis. Moreover, since Malaysia is a tropical country, which receives plenty of rainfall throughout the year, the value of pH in rivers can be easily diluted and neutralised by the rainfall. Hence, the omission of pH and NH_3-N as input data for predictive modelling have no effect on the prediction of WQI class.

In addition to the least effective parameter correlated to WQI prediction, this scenario is also able to identify the most effective parameter correlated with WQI. Analysis of the lowest prediction accuracy across the 15 set of combinations show that scenario 2-12, which excludes BOD and DO from the predictive model, was able to achieve a prediction accuracy of only 25.00%. This proves that BOD and DO are important parameters in the determination of WQI.

3.3. Third scenario: Input data with three water quality parameters

In the third scenario, the number of inputs was further reduced to three parameters. A total of 20 different combinations were tested using the predictive modelling and the results are presented in Table 9. By using only three parameters to predict WQI, scenarios 3-2 and 3-20 were able to achieve a prediction accuracy of 77.27%, which is slightly higher than the benchmark of 75%. The parameters used in scenario 3-2 were BOD, COD,

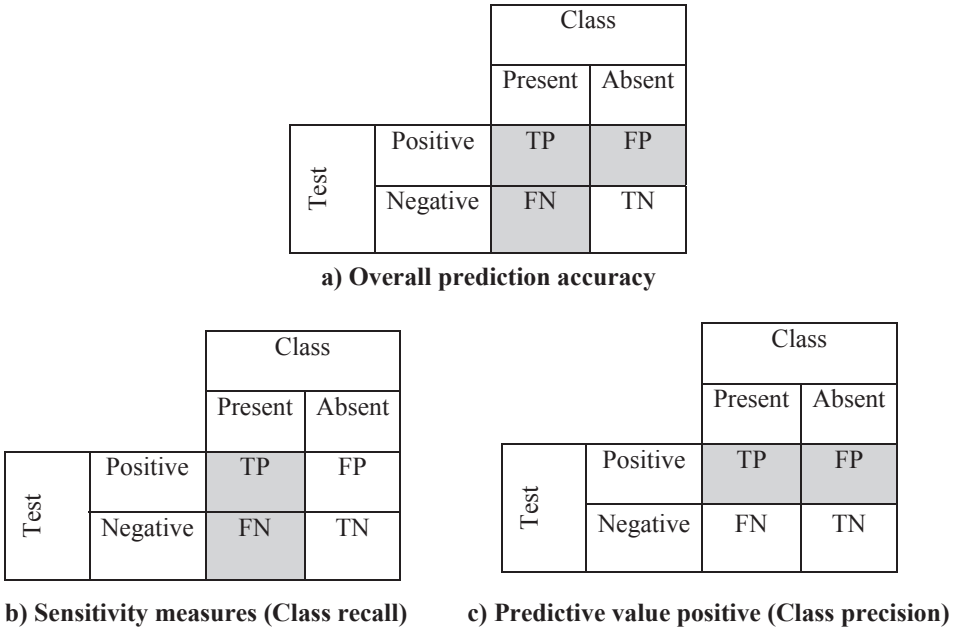

Fig. 7. Procedure for calculating a) Overall model prediction accuracy; b) Class recall and c) Class Precision.

Table 6
(2 × 2) confusion matrix.

		True condition (Actual)	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	True positive	False positive (Type 1 error)
	Predicted condition negative	False negative (Type II error)	True negative

Table 7
Prediction accuracy of each test in the first scenario.

5 Parameters Test	Water Quality Parameter						Accuracy
	BOD	COD	SS	DO	pH	NH_3-N	
1-1						X	84.09%
1-2					X		70.45%
1-3				X			68.18%
1-4			X				65.91%
1-5		X					65.91%
1-6	X						65.91%

and DO whilst those in scenario 3-20 are DO, pH, NH_3-N.

Although both tests showed good ability to predict WQI, scenario 3-2 shows a better result in contrast scenario 3-20 with regard to the accuracy of class precision and class recall. Table C.2 shows that scenario 3-2 achieved a precision of 92.59%, 50%, and 100% for the prediction of Class III, IV, and V, respectively. For class recall, the test achieved a prediction of 75.76% for true III, 100.00% for true IV, and 100.00% for true V. In addition, the false predicted classes are still close to their original classes. For example: two data points for class II fall in predicted class III and eight data points for class III fall in predicted class IV (see Tables C.3–C.19).

The result for scenario 3-20 is presented in Table C.20; it shows that two data points for true II fall in predicted III and predicted IV, four data points for true III fall in predicted IV, three data for true IV fall in predicted III, and the only data for true V fall in predicted III. The accuracy for class precision is 85.29%, 50.00% and 0.00% for predicted II, predicted IV and predicted V, respectively. The accuracy of class recall is 87.88%, 62.50%, and 0.00% for true III, true IV and true V, respectively. Results of the analysis show that even though the parameters used in scenario 3-20 were able to achieve a high accuracy, they were disqualified from being used in this predictive modelling since they were not able to accurately predict the most critical Class V. Additionally, the false predicted data show a high deviation from their true value.

The third scenario proved that NH_3-N has the lowest correlation with the prediction of WQI (based on scenario 3-2). By excluding NH_3-N, WQI can be predicted without much loss of information. In other words, BOD, COD, and DO are the most important parameters (corresponding to the relative weights given in Eq. (1)), which have a higher correlation with WQI prediction. This is proven by the result for scenario 3-19, which achieved 31.82% prediction accuracy when using SS, pH, and NH_3-N as its input data. Not only did the test achieve low prediction accuracy, most of the false predicted data deviate significantly from their true value.

The data-driven model based on decision tree procedure can be considered as a further step for achieving an adaptable water quality index prediction model. Furthermore, the utilization of such modelling procedure is not only able to accurately predict the water quality index but also able to improve the water quality monitoring program by reducing the time-consuming and costly experimental testing for each parameter, particularly NH_3-N. Additionally, the utilization of decision tree to predict water quality index could produce accurate results by allowing the use of a larger database for existing river water quality in Malaysia. The development of the proposed model in other tropical regions would allow for improvement of the present modelling system, which would then result in higher accuracy.

The ability of the proposed model to predict the class of water quality (through the calculation of WQI's class) is believed to have similar potential in predicting the indices-based river (and ecological) conditions, such as Belgium Biotic Index (BBI), Species at Risk Index (SPEAR) and German Saprobic Index (GSI) (von der Ohe et al., 2007). Interestingly, the Universal Water Quality Index (UWQI) utilises 12 parameters of water quality, intended for the abstraction of drinking water, whereby the calculated index is categorised into three classes based on EC legislation (74/440/EEC) (Boyacioglu, 2007). Based on the similarity in the indices calculation, we anticipate that the proposed model's architecture will provide a promising approach for predicting other ecological and water quality indices.

Table 8
Prediction accuracy of each test in the second scenario.

4 Parameters Test	Water Quality Parameter						Accuracy
	BOD	COD	SS	DO	pH	NH3-N	
2-1					X	X	81.82%
2-2			X			X	70.45%
2-3				X		X	50.00%
2-4		X				X	54.44%
2-5	X					X	63.64%
2-6			X		X		65.91%
2-7				X	X		75.00%
2-8		X			X		61.36%
2-9	X				X		65.91%
2-10			X	X			65.91%
2-11		X		X			59.09%
2-12	X			X			25.00%
2-13		X	X				43.18%
2-14	X		X				63.64%
2-15	X	X					75.00%

Table 9
Prediction accuracy of each test in the third scenario.

3 Parameters Test	Water Quality Parameter						Accuracy
	BOD	COD	SS	DO	pH	NH_3-N	
3-1				X	X	X	40.91%
3-2			X		X	X	77.27%
3-3		X			X	X	65.91%
3-4	X				X	X	63.64%
3-5			X	X		X	36.36%
3-6		X		X		X	50.00%
3-7	X			X		X	45.45%
3-8		X	X			X	63.64%
3-9	X		X			X	59.09%
3-10	X	X				X	72.73%
3-11			X	X	X		31.82%
3-12		X		X	X		56.82%
3-13	X			X	X		31.82%
3-14		X	X		X		61.36%
3-15	X		X		X		65.91%
3-16	X	X			X		72.73%
3-17		X	X	X			25.00%
3-18	X		X	X			31.82%
3-19	X	X		X			31.82%
3-20	X	X	X				77.27%

4. Conclusion

This research studied the use of decision tree model for water quality index prediction in a tropical environment. The monthly water quality data from the Klang River for a ten-year period (2001–2010) were utilized in this research. A decision tree algorithm was developed to predict the WQI of the Klang River by taking into account several scenarios, each of which used varying number of water quality parameters as modelling inputs. Three different scenarios were examined using the decision tree model, viz those with five, four, and three water quality parameters as the model input, with the WQI class as the target output for each scenario. In this study, the best prediction accuracy for the first scenario is 84.09% when NH_3-N was omitted from the input variables. In the second scenario, the best prediction accuracy of 81.82% was achieved when NH_3-N and pH were omitted from the input variables, and a prediction accuracy of 77.27% was achieved when NH_3-N, pH, and SS were omitted as input variables in the third scenario. The three results achieved a prediction accuracy that is higher than the benchmark of 75% prediction accuracy.

This study has proven that the number of water quality parameters in a monitoring process can be reduced. All three scenarios have shown that NH_3-N, pH, and SS have less important effect on the predicted WQI since the prediction accuracy of the model remained above the 75% benchmark when these parameters were omitted from the input variables. These findings could change the way WQI class is predicted and monitored in the future, thus allowing for better water resources management by reducing the cost and the time involved in the monitoring process.

The decision tree model is very useful in predicting WQI since it has a remarkable ability to simplify, analyse, and classify raw data to reduce its complexity and non-linearity. However, a more in-depth study needs to be carried out to further improve the prediction accuracy of the model. This predictive model could be improved by conducting an extensive study of the correlation between water quality parameters. Integration with other data pre-processing algorithm might be able to reduce the complexity of the data, hence improve the ability of the decision tree process to achieve better prediction accuracy. Although the proposed model has worthy shown appropriate prediction accuracy for WQI class for Malaysian conditions, the model in its present architecture may not directly apply to other regions without case-specific modifications.

A major outcome from the current research is that the proposed WQI's class prediction model can be of global interest, wherever the decision-makers, the regulators or other stakeholders are interested in identifying the WQI class rather than the actual value of WQ parameters. The flexibility given within the proposed prediction model's structure to identify the specifics, e.g. number of classes, variables, ranges… etc. is an essential step for model developers to adapt the model to their own case study's conditions.

Acknowledgement

The research is funded by the University of Malaya Research Grant "UMRG" (RP025A-18SUS) and supported partially by Universiti Kebangsaan Malaysia Research Grant "MI-2018-011".

Conflicts of interest

The authors declare no conflict of interest.

Appendix A:. Results for scenario I

Table A.1
Class precision and class recall accuracy for Test 1-1.

Test 1-1	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	0	0	0	0.00%
pred. III	0	2	28	0	0	93.33%
pred. IV	0	0	5	8	0	61.54%
pred. V	0	0	0	0	1	100.00%
class recall	0.00%	0.00%	84.85%	100.00%	100.00%	

Table A.2
Class precision and class recall accuracy for Test 1-2.

Test 1-2	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	2	0	0	0.00%
pred. III	0	2	25	2	1	83.33%
pred. IV	0	0	6	6	0	50.00%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	75.76%	75.00%	0.00%	

Table A.3
Class precision and class recall accuracy for Test 1-3.

Test 1-3	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	3	0	0	0.00%
pred. III	0	2	26	4	1	78.79%
pred. IV	0	0	4	4	0	50.00%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	78.79%	50.00%	0.00%	

Table A.4
Class precision and class recall accuracy for Test 1-4.

Test 1-4	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	2	0	0	0.00%
pred. III	0	2	22	1	1	84.62%
pred. IV	0	0	9	7	0	43.75%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	66.67%	87.50%	0.00%	

Table A.5
Class precision and class recall accuracy for Test 1-5.

Test 1-5	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	2	0	0	0.00%
pred. III	0	2	25	4	1	78.12%
pred. IV	0	0	6	4	0	40.00%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	75.76%	50.00%	0.00%	

Table A.6
Class precision and class recall accuracy for Test 1-6.

Test 1-6	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	2	0	0	0.00%
pred. III	0	2	24	2	1	82.76%
pred. IV	0	0	7	5	0	41.67%
pred. V	0	0	0	1	0	0.00%
class recall	0.00%	0.00%	72.73%	62.50%	0.00%	

Appendix B:. Results for scenario II

Table B.1
Class precision and class recall accuracy for Test 2-1.

Test 2-1	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	0	0	0	0.00%
pred. III	0	2	27	0	0	93.10%
pred. IV	0	0	6	8	0	57.14%
pred. V	0	0	0	0	1	100.00%
class recall	0.00%	0.00%	81.82%	100.00%	100.00%	

Table B.2
Class precision and class recall accuracy for Test 2-2.

Test 2-2	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	0	0	0	0.00%
pred. III	0	2	22	0	0	91.67%
pred. IV	0	0	11	8	0	42.11%
pred. V	0	0	0	0	1	100.00%
class recall	0.00%	0.00%	66.67%	100.00%	100.00%	

Table B.3
Class precision and class recall accuracy for Test 2-3.

Test 2-3	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	1	4	0	0	20.00%
pred. III	0	1	21	8	1	67.74%
pred. IV	0	0	8	0	0	0.00%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	50.00%	63.64%	0.00%	0.00%	

Table B.4
Class precision and class recall accuracy for Test 2-4.

Test 2-4	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	1	0	0	0.00%
pred. III	0	2	20	4	0	76.92%
pred. IV	0	0	12	4	1	23.53%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	60.61%	50.00%	0.00%	

Table B.5
Class precision and class recall accuracy for Test 2-5.

Test 2-5	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	3	0	0	0.00%
pred. III	0	2	20	0	0	90.91%
pred. IV	0	0	10	7	0	41.18%
pred. V	0	0	0	1	1	50.00%
class recall	0.00%	0.00%	60.61%	87.50%	100.00%	

Table B.6
Class precision and class recall accuracy for Test 2-6.

Test 2-6	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	2	0	0	0.00%
pred. III	0	2	22	1	1	84.62%
pred. IV	0	0	9	7	0	43.75%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	66.67%	87.50%	0.00%	

Table B.7
Class precision and class recall accuracy for Test 2-7.

Test 2-7	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	0	0	0	0.00%
pred. III	0	2	29	4	1	80.56%
pred. IV	0	0	4	4	0	50.00%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	87.88%	50.00%	0.00%	

Table B.8
Class precision and class recall accuracy for Test 2-8.

Test 2-8	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	2	0	0	0.00%
pred. III	0	2	23	4	1	76.67%
pred. IV	0	0	8	4	0	33.33%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	69.70%	50.00%	0.00%	

Table B.9
Class precision and class recall accuracy for Test 2-9.

Test 2-9	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	2	0	0	0.00%
pred. III	0	2	24	2	1	82.76%
pred. IV	0	0	7	5	0	41.67%
pred. V	0	0	0	1	0	0.00%
class recall	0.00%	0.00%	72.73%	62.50%	0.00%	

Table B.10
Class precision and class recall accuracy for Test 2-10.

Test 2-10	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	0	0	0	0.00%
pred. III	0	2	29	8	1	72.50%
pred. IV	0	0	4	0	0	0.00%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	87.88%	0.00%	0.00%	

Table B.11
Class precision and class recall accuracy for Test 2-11.

Test 2-11	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	3	0	0	0.00%
pred. III	0	2	22	4	1	75.86%
pred. IV	0	0	8	4	0	33.33%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	66.67%	50.00%	0.00%	

Table B.12
Class precision and class recall accuracy for Test 2-12.

Test 2-12	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	3	0	0	0.00%
pred. III	0	0	4	1	1	66.67%
pred. IV	0	2	26	7	0	20.00%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	12.12%	87.50%	0.00%	

Table B.13
Class precision and class recall accuracy for Test 2-13.

Test 2-13	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	2	0	0	0.00%
pred. III	0	2	15	4	1	68.18%
pred. IV	0	0	16	4	0	20.00%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	45.45%	50.00%	0.00%	

Table B.14
Class precision and class recall accuracy for Test 2-14.

Test 2-14	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	2	0	0	0.00%
pred. III	0	1	21	1	1	87.50%
pred. IV	0	1	10	7	0	38.89%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	63.64%	87.50%	0.00%	

Table B.15
Class precision and class recall accuracy for Test 2-15.

Test 2-15	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	1	0	0	0.00%
pred. III	0	1	27	2	1	87.10%
pred. IV	0	1	5	6	0	50.00%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	81.82%	75.00%	0.00%	

Appendix C:. Results for scenario III

Table C.1
Class precision and class recall accuracy for Test 3-1.

Test 3-1	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	1	4	0	1	16.67%
pred. III	0	1	14	5	0	70.00%
pred. IV	0	0	15	3	0	16.67%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	50.00%	42.42%	37.50%	0.00%	

Table C.2
Class precision and class recall accuracy for Test 3-2.

Test 3-2	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	0	0	0	0.00%
pred. III	0	2	25	0	0	92.59%
pred. IV	0	0	8	8	0	50.00%
pred. V	0	0	0	0	1	100.00%
class recall	0.00%	0.00%	75.76%	100.00%	100.00%	

Table C.3
Class precision and class recall accuracy for Test 3-3.

Test 3-3	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	1	0	0	0.00%
pred. III	0	2	24	3	0	82.76%
pred. IV	0	0	8	5	1	35.71%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	72.73%	62.50%	0.00%	

Table C.4
Class precision and class recall accuracy for Test 3-4.

Test 3-4	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	2	0	0	0.00%
pred. III	0	2	20	0	0	90.91%
pred. IV	0	0	11	7	0	38.89%
pred. V	0	0	0	1	1	50.00%
class recall	0.00%	0.00%	60.61%	87.50%	100.00%	

Table C.5
Class precision and class recall accuracy for Test 3-5.

Test 3-5	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	1	1	0	1	33.33%
pred. III	0	1	14	7	0	63.64%
pred. IV	0	0	18	1	0	5.26%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	50.00%	42.42%	12.50%	0.00%	

Table C.6
Class precision and class recall accuracy for Test 3-6.

Test 3-6	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	1	2	0	0	33.33%
pred. III	0	1	17	4	1	73.91%
pred. IV	0	0	14	4	0	22.22%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	50.00%	51.52%	50.00%	0.00%	

Table C.7
Class precision and class recall accuracy for Test 3-7.

Test 3-7	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	2	12	0	0	14.29%
pred. III	0	0	10	1	0	90.91%
pred. IV	0	0	11	7	0	38.89%
pred. V	0	0	0	0	1	100.00%
class recall	0.00%	100.00%	30.30%	87.50%	100.00%	

Table C.8
Class precision and class recall accuracy for Test 3-8.

Test 3-8	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	0	0	0	0.00%
pred. III	0	2	24	4	1	77.42%
pred. IV	0	0	9	4	0	30.77%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	72.73%	50.00%	0.00%	

Table C.9
Class precision and class recall accuracy for Test 3-9.

Test 3-9	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	1	0	0	0.00%
pred. III	0	1	17	0	0	94.44%
pred. IV	0	1	15	8	0	33.33%
pred. V	0	0	0	0	1	100.00%
class recall	0.00%	0.00%	51.52%	100.00%	100.00%	

Table C.10
Class precision and class recall accuracy for Test 3-10.

Test 3-10	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	2	0	0	0.00%
pred. III	0	2	26	2	0	86.67%
pred. IV	0	0	5	6	1	50.00%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	78.79%	75.00%	0.00%	

Table C.11
Class precision and class recall accuracy for Test 3-11.

Test 3-11	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	0	0	0	0.00%
pred. III	0	1	10	4	1	62.50%
pred. IV	0	1	23	4	0	14.29%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	30.30%	50.00%	0.00%	

Table C.12
Class precision and class recall accuracy for Test 3-12.

Test 3-12	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	1	0	0	0.00%
pred. III	0	2	21	4	1	75.00%
pred. IV	0	0	11	4	0	26.67%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	63.64%	50.00%	0.00%	

Table C.13
Class precision and class recall accuracy for Test 3-13.

Test 3-13	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	0	0	0	0.00%
pred. III	0	0	7	1	1	77.78%
pred. IV	0	2	26	7	0	20.00%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	21.21%	87.50%	0.00%	

Table C.14
Class precision and class recall accuracy for Test 3-14.

Test 3-14	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	2	0	0	0.00%
pred. III	0	2	24	5	1	75.00%
pred. IV	0	0	7	3	0	30.00%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	72.73%	37.50%	0.00%	

Table C.15
Class precision and class recall accuracy for Test 3-15.

Test 3-15	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	2	0	0	0.00%
pred. III	0	1	22	1	1	88.00%
pred. IV	0	1	9	7	0	41.18%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	66.67%	87.50%	0.00%	

Table C.16
Class precision and class recall accuracy for Test 3-16.

Test 3-16	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	1	0	0	0.00%
pred. III	0	1	27	2	1	87.10%
pred. IV	0	1	5	5	0	45.45%
pred. V	0	0	0	1	0	0.00%
class recall	0.00%	0.00%	81.82%	62.50%	0.00%	

Table C.17
Class precision and class recall accuracy for Test 3-17.

Test 3-17	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	1	0	1	0.00%
pred. III	0	1	7	4	0	58.33%
pred. IV	0	1	24	4	0	13.79%
pred. V	0	0	1	0	0	0.00%
class recall	0.00%	0.00%	21.21%	50.00%	0.00%	

Table C.18
Class precision and class recall accuracy for Test 3-18.

Test 3-18	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	0	0	0	0.00%
pred. III	0	0	7	1	1	77.78%
pred. IV	0	2	26	7	0	20.00%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	21.21%	87.50%	0.00%	

Table C.19
Class precision and class recall accuracy for Test 3-19.

Test 3-19	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	0	0	0	0.00%
pred. III	0	0	7	1	1	77.78%
pred. IV	0	2	26	7	0	20.00%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	21.21%	87.50%	0.00%	

Table C.20
Class precision and class recall accuracy for Test 3-20.

Test 3-20	true I	true II	true III	true IV	true V	class precision
pred. I	0	0	0	0	0	0.00%
pred. II	0	0	0	0	0	0.00%
pred. III	0	1	29	3	1	85.29%
pred. IV	0	1	4	5	0	50.00%
pred. V	0	0	0	0	0	0.00%
class recall	0.00%	0.00%	87.88%	62.50%	0.00%	

References

Afan, H.A., Keshtegar, B., Mohtar, W.H.M.W., El-Shafie, A., 2017. Harmonize input selection for sediment transport prediction. J. Hydrol. 552. https://doi.org/10.1016/j.jhydrol.2017.07.008.

Azad, M., Moshkov, M., 2017. Multi-stage optimization of decision and inhibitory trees for decision tables with many-valued decisions. Eur. J. Oper. Res. 263, 910–921. https://doi.org/10.1016/J.EJOR.2017.06.026.

Brown, S.D., Myles, A.J., 2013. Decision tree modeling in classification. Ref. Modul. Chem. Mol. Sci. Chem. Eng. https://doi.org/10.1016/B978-0-12-409547-2.00653-3.

Boyacioglu, H., 2007. Development of a water quality index based on a European classification scheme. Water SA 33 (1), 101–106.

Chau, K. wing, 2006. A review on integration of artificial intelligence into water quality modelling. Mar. Pollut. Bull. https://doi.org/10.1016/j.marpolbul.2006.04.003.

Csábrági, A., Molnár, S., Tanos, P., Kovács, J., 2017. Application of artificial neural networks to the forecasting of dissolved oxygen content in the Hungarian section of the river Danube. Ecol. Eng. 100, 63–72. https://doi.org/10.1016/j.ecoleng.2016.12.027.

DOE, 2007. Malaysia Environmental Quality Report 2007. Malaysia Environ. Qual. Rep. 1–86. doi: 10.1007/s13398-014-0173-7.2.

El-Shafie, A., Noureldin, A., Taha, M., Hussain, A., Mukhlisin, M., 2012. Dynamic versus static neural network model for rainfall forecasting at Klang River Basin, Malaysia. Hydrol. Earth Syst. Sci. https://doi.org/10.5194/hess-16-1151-2012.

Everaert, G., Bennetsen, E., Goethals, P.L.M., 2016. An applicability index for reliable and applicable decision trees in water quality modelling. Ecol. Inform. 32, 1–6. https://doi.org/10.1016/J.ECOINF.2015.12.004.

Hameed, M., Sharqi, S.S., Yaseen, Z.M., Afan, H.A., Hussain, A., Elshafie, A., 2016. Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia. Neural Comput. Appl., 1–13. doi: 10.1007/s00521-016-2404-7.

Hofmann, M., Klinkenberg, R., Hofmann, M., Klinkenberg, R., 2013. RapidMiner: Data Mining Use Cases and Business Analytics Applications, Zhurnal Eksperimental'noi i Teoreticheskoi Fiziki. doi: 78-1-4822-0550-3.

Khosravi, K., Pham, B.T., Chapi, K., Shirzadi, A., Shahabi, H., Revhaug, I., Prakash, I., Tien Bui, D., 2018. A comparative assessment of decision trees algorithms for flash flood susceptibility modeling at Haraz watershed, northern Iran. Sci. Total Environ. 627, 744–755. https://doi.org/10.1016/J.SCITOTENV.2018.01.266.

Kotu, V., Deshpande, B., 2014. Predictive analytics and data mining: concepts and practice with rapidminer.

Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environ. Model. Softw. https://doi.org/10.1016/S1364-8152(99)00007-9.

Manache, G., Melching, C.S., 2008. Identification of reliable regression- and correlation-based sensitivity measures for importance ranking of water-quality model parameters. Environ. Model. Softw. https://doi.org/10.1016/j.envsoft.2007.08.001.

Mohamed, I., Othman, F., Ibrahim, A.I.N., Alaa-Eldin, M.E., Yunus, R.M., 2015. Assessment of water quality parameters using multivariate analysis for Klang River basin, Malaysia. Environ. Monit. Assess. https://doi.org/10.1007/s10661-014-4182-y.

Najah, A.A., El-Shafie, A., Karim, O.A., Jaafar, O., 2012. Water quality prediction model utilizing integrated wavelet-ANFIS model with cross-validation. Neural Comput. Appl. 21, 833–841. https://doi.org/10.1007/s00521-010-0486-1.

Othman, F., M E, A.E., Mohamed, I., 2012. Trend analysis of a tropical urban river water

quality in Malaysia. J. Environ. Monit. https://doi.org/10.1039/c2em30676j.
Sarkar, A., Pandey, P., 2015. River water quality modelling using artificial neural network technique. Aquat. Procedia 4, 1070–1077. https://doi.org/10.1016/j.aqpro.2015.02.135.
Sharif, S.M., Kusin, F.M., Asha'ari, Z.H., Aris, A.Z., 2015. Characterization of water quality conditions in the Klang river basin, Malaysia using self organizing map and K-means algorithm. Procedia Environ. Sci. https://doi.org/10.1016/j.proenv.2015.10.013.
Singh, K.P., Basant, N., Gupta, S., 2011. Support vector machines in water quality management. Anal. Chim. Acta. https://doi.org/10.1016/j.aca.2011.07.027.
von der Ohe, P.C., Pruss, A., Schafer, R.B., Liess, M., de Deckere, E., Brack, W., 2007. Water quality indices across Europe–a comparison of the good ecological status of five river basins. J. Environ. Monit. 9 (9), 970–978.