Prediction of Water Quality for Lake Monitoring System using Machine Learning Approach

HONG YUEN MUN

FACULTY OF COMPUTING AND INFORMATICS

UNIVERSITY MALAYSIA SABAH

2022

PREDICTION OF WATER QUALITY FOR LAKE MONITORING SYSTEM USING MACHINE LEARNING APPROACH

HONG YUEN MUN

THESIS SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE OF BACHELOR OF COMPUTER SCIENCE WITH HONOURS (NETWORK ENGINEERING)

FACULTY OF COMPUTING AND INFORMATICS
UNIVERSITY MALAYSIA SABAH

NAME	: HONG YUEN MUN
MATRIC NUMBER	: BI18110203
TITLE	**:** Prediction of Water Quality for Lake Monitoring System using Machine Learning Approach
DEGREE	: BACHELOR OF COMPUTER WITH HONOURS (NETWORK ENGINEERING)
VIVA'S DATE	: 24 JANUARY 2022

CERTIFIED BY;

1. **SUPERVISOR** Signature

DR. FARASHAZILLAH YAHYA

DECLARATION

I hereby declare that the material in this thesis is my own except for quotations, equations, summaries and references, which have been duly acknowledged.

17 FEBRUARY 2022

HONG YUEN MUN
BI18110203

ACKNOWLEDGEMENT

I would like to take this opportunity to express my gratitude and thanks to my supervisor Dr. Farashazillah Yahya for her support and care during the progress of my project. Dr. Farashazillah Yahya has provided guidelines and advices that were very useful to me. Dr. Farashazillah Yahya was a very responsible and dedicate lecturer. Moreover, I also want to acknowledge all the lecturers who are willing to help and guide me out. They lent me a helping hand from time to time. In addition, I would like to express sincere thanks and upmost appreciation to my family for giving me support, concern and encouragement. Besides, I also want to acknowledge my friends for their support and help. Lastly, I wish to offer my regards and blessings to all of personnel who supported me in any aspect during the completion of my project.

HONG YUEN MUN
17 FEBURARY 2022

ABSTRACT

Water quality in lakes is a critical issue due to its direct influence on public health, biological integrity of natural resources, and the economy. There are a variety of lakes, from small to big, natural and manmade reservoirs to natural lakes. Even though the lakes only consist small part of water on our planet, they play an important role in earth's biosphere, climate changes, land-use changes, and anthropogenic changes due to various urban and industrial development can lead to hydrological, chemical, and biological changes in watersheds and freshwater ecosystems resulting in altered water quality. The deteriorating quality of natural water resources like lakes is one of the dire problems and most concerning issues faced by humanity. Lack of quality lake water is most likely because lake water became contaminated due to various factors such as humans, industrial, commercial activities, and natural processes. To understand the impact of changes from upstream or surrounding watersheds and within a lake on water quality is important to people who live nearby or visit the lake and is also fundamental in providing better ecological and environmental strategies and mitigation methods to protect the freshwater ecosystems. Dissolved oxygen and other water quality will affect the growth, reproduction, and survivability of freshwater organisms. Climate variations can directly affect the temperature of an aquatic system through the surface heat exchange between the water and the surrounding atmosphere and further influence water quality characteristics. Monitoring and modeling approaches have been used by volunteers, biologists, water resources managers, engineers, and scientists to understand and further study water quality issues in the lake. Therefore, the monitoring and prediction of lake water quality will provide more in-depth and necessary information and evidence to help in managing the water quality. Monitoring data are necessary for model calibration and validation before the model can be used for scenario study, sensitive analysis, and future projection under certain changes in lakes water. In this project, a machine learning approach is proposed to assist the prediction of the lake water quality. The evaluation will then be used to predict the quality of water.

Keywords: machine learning, prediction of water quality, lake water

ABSTRAK

RAMALAN KUALITI AIR BAGI SISTEM PEMANTAUAN TASIK MENGGUNAKAN PENDEKATAN PEMBELAJARAN MESIN

Kualiti air di tasik merupakan isu kritikal kerana pengaruh langsungnya terhadap kesihatan awam, integriti biologi sumber semula jadi dan ekonomi. Terdapat pelbagai tasik, dari kecil hingga besar, takungan semula jadi dan buatan manusia kepada tasik semula jadi. Walaupun tasik hanya terdiri daripada sebahagian kecil air di planet kita, ia memainkan peranan penting dalam biosfera bumi, perubahan iklim, perubahan guna tanah, dan perubahan antropogenik akibat pelbagai pembangunan bandar dan perindustrian boleh membawa kepada hidrologi, kimia dan biologi. perubahan dalam kawasan tadahan air dan ekosistem air tawar yang mengakibatkan kualiti air berubah. Kemerosotan kualiti sumber air semula jadi seperti tasik adalah salah satu masalah yang teruk dan paling membimbangkan isu yang dihadapi oleh manusia. Kekurangan air tasik yang berkualiti berkemungkinan besar kerana air tasik menjadi tercemar disebabkan oleh pelbagai faktor seperti manusia, perindustrian, aktiviti komersial, dan proses semula jadi. Untuk memahami kesan perubahan dari hulu atau kawasan tadahan air dan di dalam tasik terhadap kualiti air adalah penting kepada orang yang tinggal berdekatan atau melawat tasik dan juga asas dalam menyediakan strategi ekologi dan alam sekitar serta kaedah mitigasi untuk melindungi ekosistem air tawar. Oksigen terlarut dan kualiti air lain akan menjejaskan pertumbuhan, pembiakan, dan kemandirian organisma air tawar. Variasi iklim secara langsung boleh mempengaruhi suhu sistem akuatik melalui pertukaran haba permukaan antara air dan atmosfera sekeliling dan seterusnya mempengaruhi ciri kualiti air. Pendekatan pemantauan dan pemodelan telah digunakan oleh sukarelawan, ahli biologi, pengurus sumber air, jurutera dan saintis untuk memahami dan mengkaji lebih lanjut isu kualiti air di tasik. Oleh itu, pemantauan dan ramalan kualiti air tasik akan memberikan maklumat dan bukti yang lebih mendalam dan perlu untuk membantu dalam menguruskan kualiti air. Data pemantauan diperlukan untuk penentukuran dan pengesahan model sebelum model boleh digunakan untuk kajian senario, analisis sensitif dan unjuran masa hadapan di bawah perubahan tertentu dalam air tasik. Dalam projek ini, pendekatan pembelajaran mesin dicadangkan untuk membantu ramalan kualiti air tasik. Penilaian kemudiannya akan digunakan untuk meramalkan kualiti air.

***Kata kunci**: pembelajaran mesin, ramalan kualiti air, air tasik,*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Introduction

This section describes the idea and motivation to develop a web-based water quality prediction system for lake monitoring system using machine learning approach. The section included are 1.1 Introduction, 1.2 Problem Background/Motivation, 1.3 Problem Statements, 1.4 Project Objectives, 1.5 Project Scope, 1.6 Organization of the report, 1.7 conclusion.

1.2 Problem Background

For all life on Earth, water is one of the most important natural resources. Water supply and efficiency have always played a key role in deciding not just where people can live, but also how happy they can be. Despite the fact that there has always been plenty of fresh water on Earth, it has not always been available when and where it is needed, nor has it always been of sufficient quality for all purposes. Water must be viewed as a limited resource with availability and suitability for usage having limits and boundaries.

The quality of most atmospheric water sources, such as rivers, lakes, and streams, is determined by strict quality requirements. Water specifications for other applications often have their own set of guidelines. Generally, lakes functioned as thermal structures, habitats for organisms and support food chain also playing an important role as nutrient providers to marine life (Hairston et al., 2014). Lake water is one of the surface water sources that play an important role as basic water resources as water is necessary for the whole life cycle of human beings and also associated activities (Asharuddin et al., 2016), (Badaii et al., 2013). Therefore, water quality is our main concern due to its impact on our human health. Water quality can be affected by many factors such as geological structure, the salinity of the water, overdraw of the groundwater, wastage of water, drainage caused by agriculture and

pollutants from chemical compounds. Due to such factors, water quality is degenerating at a rapid paced and causes various problems such as drinking water that contains unsafe levels of contaminants can cause health effects such as gastrointestinal illnesses, nervous system or reproductive effects, and chronic diseases such as cancer. Massive population growth, the industrial revolution, and the widespread usage of fertilizers and pesticides have all had a negative impact on the water quality setting. As a result, providing models for predicting the water quality is extremely useful for water quality monitoring.

Therefore, it is important to find new method of approaches to analyze and find way to predict the quality of water. Various methods and investigations are conducted to research, monitor and predict the quality of lake water. The first model in water quality monitoring was introduced by Horton in 1965 (Abdillah et al., 2013) before the discovery of newer methods by different experts. It is recommended to consider the temporal dimension for predicting the Water quality patterns to ensure the monitoring of the seasonal change of the Water quality. However, using a special variation of models together to predict the Water quality grants better results than using a single model (Theyazn et al., 2020). There are several methodologies proposed for the prediction and modeling of the Water quality. These methodologies include statistical approaches, visual modeling, analyzing algorithms, and predictive algorithms. For the sake of the determination of the correlation and relationship among different water quality parameters, multivariate statistical techniques have been employed (Farrell, 2000). The geostatistical approaches were used for transitional probability, multivariate interpolation, and regression analysis (Taskaya and Casey, 2005).

For this project, the focus will be on developing a water quality prediction system using machine learning approach.

1.3 Project Statements

One of the most critical aspects of a healthy environment is water quality. Clean water is essential for the survival of a wide range of plants and animals. Though it may appear unrelated at first, our land-based activities have an impact on the quality of our water. Pollutants, excess nutrients from fertilizers, and silt are commonly transferred into local lakes and rivers by runoff from cities and agricultural fields. Water pollution is the contamination of water bodies that occur when pollutant are

indirectly or directly discharge into water bodies without adequate treatment to remove the harmful sediment (Muyibi et al., 2008). It will have an impact on the ecosystem and human existence, and it has now become a problem. Furthermore, due to human or industrial activities, water supplies are gradually becoming polluted and unavailable. According to Rene´ P (2006), the increasing contamination of freshwater systems with thousands of industrial and natural chemical compounds is one of the key environmental problems facing humanity worldwide.

At the moment, UMS manages their lake water quality manually. So, it is difficult to measure the water quality index of the lakes daily. It will affect the safety of human and also threaten aquatic life. Therefore, it is important to be able to predict the water quality index and classify the water quality index. This project aim is to select a suitable machine learning model for predicting the quality of lake water. Several regression model will be compared to select the most suitable machine learning model for the prediction of water quality index system. A sequence of parameters captured over several years are used in selected algorithms for training and to predict the quality of lake water. Therefore, the proposed system intends to assist the lake managers. The data of the parameters collected from the past few years will be used to predict the quality of lake water.

1.4 Project Objectives

1. To compare and choose the suitable regression machine learning model based on the r-squared score and mean squared error.
2. To implement the selected regression machine learning models into a web-based water quality index prediction system.
3. To test and evaluate the usability of water quality index prediction system using the System Usability Scale (SUS).

1.5 Project Scope

Prediction of Water Quality for Lake Monitoring System using Machine Learning Approach will be created in this project in the form of a web application. Research will be conducted to select the most suitable model. The language used for machine learning approach will be Python assisted by Jupyter Notebook and visual studio code. As for the front-end framework will be managed using HTML5, CSS, PHP, and MySQL.

The function of the web application is predicting the quality of the water by inputting the parameters. Water quality data samples from Tasik Putrajaya from year 2010 to year 2015 are used for this project.

1.6 Organization of The Report

This project consisted of 6 chapters which are Chapter 1 Introduction, Chapter 2 Literature Review, Chapter 3 Methodology, Chapter 4 System Analysis and Design. Chapter 1 will discuss the introduction, problem statement and motivation, problem statements, project objectives, project scope, and organization of the report. Chapter 2 will explain, and reviews research done, fact and existing system that are related to the project. The resources are taken from internet, journals, and articles. Chapter 3 will discuss methodology for building the system. Software and hardware requirements will also be determined in this chapter. Chapter 4 will discuss system analysis and design which consisted of the flow of activities while developing this system.

1.7 Conclusion

The prediction of water quality system will be developed as a web application to assist lake monitoring system. The prediction of water quality system is developed with the approach of machine learning. Highly efficient models for prediction of water quality will be developed based on artificial neural network and deep learning algorithm. The system function is to help in predicting the water quality by inputting the required parameters. This system can be great help to assist the lake managers in UMS as they are managing their lake water quality manually.

CHAPTER 2

Preliminary Study

2.1 Introduction

This section discusses a brief introduction of machine learning and the study of existing systems that are related to utilizing machine learning algorithms to come up with item recommendations based on user preferences. This section also contains the comparison among the machine learning model and a summary of the related works. The section included are 2.1 Introduction, 2.2 Linear Regression Model, 2.3 Random Forest Regression Model, 2.4 Decision Tree Regression Model, 2.5 Support Vector Machine Regression Model, 2.6 K-nearest neighbour Regression Model, 2.7 Pros and Cons between, Linear Regression Model, Random Forest Regression Model, Decision Tree Regression Model, Support Vector Machine Regression Model, and K-nearest neighbour Regression Model, 2.8 Related Works, and 2.9 conclusion.

2.2 Linear Regression Model

Linear regression is one of the most commonly used predictive modelling techniques. It is represented by an equation $Y = a + bX + e$, where a is the intercept, b is the slope of the line and e is the error term. This equation can be used to predict the value of a target variable based on given predictor variable(s).

2.3 Random Forest Regression Model

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods. The Bayesian method uses the

knowledge of probability statistics to predict and classify datasets. The Bayesian algorithm combines prior and posterior probabilities to avoid the supervisor's bias and the overfitting phenomenon of using sample information alone.

This Naive Bayes is a type of classification algorithms based on Bayes' theorem and the assumption of the independence of characteristic conditions. Attributes are assumed to be conditionally independent of each other when the target value is given. This method greatly simplifies the complexity of the Bayesian method.

2.4 Decision Tree Regression Model

Decision Tree Regression is a machine-learning technique that may be used for both classification and regression calculations (Osei-Bryson, 2004). The DTR algorithm extracts information from a dataset and arranges them in a symbolic tree-shaped structure, with internal and terminal nodes representing splits and leaves, respectively. A tree is formed by following a set of basic guidelines. A set of rules can be formed by combining many trees, which can then be used in the prediction step. First, the training dataset is utilised to build a tree; secondly, using a binary split process, the algorithm divides the original data into two branches. This separation process is then applied to fresh growth branches, and the process is repeated until each branch is indistinguishable and the corresponding node reaches maturity. This separation procedure is then repeated on fresh growth branches until each branch is indistinguishable and the corresponding node reaches the minimum size and becomes a terminal node (Xu et al., 2005). When compared to other models, DTR has a distinct benefit in that its rules are simple to understand and follow a logical pattern expressed in the form of a tree (Mitchell, 1997). DTR, while faster than other AI models, can not give reliable answers in the presence of nonlinearity or noisy datasets, and is frequently ineffective for time series applications (Curram and Mingers, 1994; Tso and Yau, 2007).

2.5 Support Vector Machine Regression Model

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they are able to categorize new text. The SVM model was developed in 1995 by Corinna Cortes and Vapnik. It has several unique benefits in solving small samples, and nonlinear and high-

dimensional pattern recognition. It can be extended to function in the simulation of other machine learning problems. It uses the hyperplane to separate the points of the input vectors and finds the needed coefficients. The best hyperplane is the line with the largest margin, which is meant the distance between the hyperplane and the nearest input objects. The input points defined in the hyperplane are called support vectors.

2.6 K-nearest neighbour Regression Model

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

2.7 Pros and Cons between, Linear Regression Model, Random Forest Regression Model, Decision Tree Regression Model, Support Vector Machine Regression Model, and K-nearest neighbour Regression Model.

Table 1: Comparisons Between Different Machine Learning Models

Regression Machine Learning Model	Pros	Cons
Linear Regression Model	The Linear regression model is the simplest equation using which the	Linear regression makes strong assumptions that there is Predictor

	relationship between the multiple predictor variables and predicted variable can be expressed. The modeling speed of Linear regression is fast as it does not require complicated calculations and runs predictions fast when the amount of data is large. The ability of Linear regression to determine the relative influence of one or more predictor variables to the predicted value when the predictors are independent of each other is one of the key reasons of the popularity of Linear regression. The model derived using this method can express the what change in the predictor variable causes what change in the predicted or target variable.	(independent) and Predicted (dependent) variables are linearly related which may not be the case. Outliers can have a large effect on the output, as the Best Fit Line tries to minimize the MSE for the outlier points as well, resulting in a model that is not able to capture the information in the data.
Random Forest Regression Model	Robust to outliers. Works well with non-linear data. Lower risk of overfitting.	Random forests are found to be biased while dealing with categorical variables. Slow Training.

	Runs efficiently on a large dataset. Better accuracy than other classification algorithms.	Not suitable for linear methods with a lot of sparse features
Decision Tree Regression Model	Easy to interpret Handles both categorical and continuous data well. Works well on a large dataset. Not sensitive to outliers. Non-parametric in nature.	These are prone to overfitting. It can be quite large, thus making pruning necessary. It can't guarantee optimal trees. It gives low prediction accuracy for a dataset as compared to other machine learning algorithms. Calculations can become complex when there are many class variables. High Variance (Model is going to change quickly with a change in training data)
Support Vector Machine Regression Model	It works really well with a clear margin of separation. It is effective in high dimensional spaces. It is effective in cases where the number of dimensions is greater	It doesn't perform well when we have large data set because the required training time is higher. It also doesn't perform very well, when the data set has more noise i.e.

	than the number of samples. It uses a subset of training points in the decision function, so it is also memory efficient.	target classes are overlapping. SVM doesn't directly provide probability estimates, these are calculated using an expensive five-fold cross-validation.
K-Nearest Neighbour Regression Model	No assumptions about data. Useful for nonlinear data Simple algorithm. Easy to understand and interpret. High accuracy relatively but not competitive in comparison to better supervised learning models. Versatile. Can be use for classification or regression.	Computationally expensive because the algorithm stores all of the training data. High memory requirement. Stores almost all of the training data. Prediction stage might be slow. Sensitive to irrelevant features and the scale of the data.

2.8 Related Works

There are two main types for modeling and predicting water quality are available which are mechanism and non-mechanism-oriented models. The mechanism model is relatively sophisticated, it uses advanced system structure data for simulating the water quality. Therefore, it is considered as multifunctional model that can be used for any water body. One of the earliest water quality simulation model is Streeter-Phelos model which is widely used. Later, some countries have developed a variety of different models including QUAL model (Lai et al., 2011) and the WASP model (Huang et al.,2010), which have gained wide usage in mimicking the water quality of rivers. This was followed by Warren and Bach (Warren et al., 1992) who suggested