PREDICTIVE MODELING OF POLLUTION IN RIVER BASINS USING MACHINE LEARNING TECHNIQUE

HASLINDA BINTI ABDUL SAHAK

MCS241004_MCST 1043

PROPOSAL

UNIVERSITI TEKNOLOGI MALAYSIA

NOVEMBER 2024

SCHOOL OF COMPUTING
Faculty of Engineering

Project Proposal Form MCSD 6215
Sem: 1 Session: 2024 / 2025

SECTION A: Project Information.

Program Name:	**Masters of Data Science**
Subject Name:	**Project 1 (MCSD 6215)**
Student Name:	Haslinda binti Abdul Sahak
Metric Number:	MCS241004
Student Email & Phone:	haslinda45@graduate.utm.my & 011-7035 8229
Project Title:	Predictive Modeling of Pollution in River Basins using Machine Learning Technique
Supervisor 1:	
Supervisor 2 / Industry Advisor (if any):	

SECTION B: Project Proposal

Introduction:

Pollution in river basins is one of the most critical environmental hazards; it threatens ecosystems, human lives, and social and economic development. The traditional time-consuming monitoring schemes do not cater to the complexity of water pollution. The research intends to develop a strong machine-learning model to predict the pollution levels in the river basins of Malaysia. In this study, a historical database shall be used along with an advanced machine learning technique to convert public data into valuable insights for balanced pollution management and mitigation strategies.

Problem Background:

Water pollution is a serious problem in Malaysian river basins. Industrial wastes, agricultural runoffs, and urban sewages significantly deteriorate the quality of water. The impacts are too severe and concern environment-related problems, such as excessive plant growth, toxic algal blooms, and biodiversity losses.

Currently, water quality monitoring is slow, ineffective manual sampling and laboratory analyses hog a whole lot of resources over time, which makes it quite difficult to grasp real-time dynamics of water quality. This could restrict prompt actions against pollution events.

An immediate proactive way to curb water pollution is the prediction of pollution levels. Such accurate predictions allow decision-makers to take on-time preventive measures to reduce pollution sources and preserve water resources from harm. This would also help ensure that future generations live in a better environment and have better native water quality.

Problem Statement:

The conventional means of gauging the quality of water include manual sampling followed by laboratory analysis; these approaches consume a lot of time and are labor-intensive. Further, they may fail to capture real-time fluctuations in water quality and may result in delayed action during pollution events. Most of all, traditional means do not lay bare the whole area of pollution; because of that, sensitives, in fact, fail to understand the landscape.

Pollution in river basins can be explained mainly due to hydrological factors and meteorological conditions. Pollution can also be caused due to anthropogenic factors. Modeling and prediction require understanding such complex scenarios and their impact on water quality in terms of detecting and assessing pollution levels. Machine learning techniques hold promise in basically opening up new possibilities for analyzing big datasets for the detection of very intricate patterns. Developing accurate and reliable predictive models does, however, need proper handling of data, such as quality, feature engineering, and model selection.

It is possible that even the greatest prediction-that high-level advanced machine learning technique would address easy means of communicating with stakeholders. These tools should reach as many users as possible-from policymakers to environmental scientists and the general public. Simple visualizations and interactive dashboards would be efficient vehicles for transforming stakeholder communities from passive audiences into informed user groups that can act in their interests.

Aim of the Project:

The project intends to construct an advanced machine-learning model to predict pollution levels in the different river basins within Malaysia. This model would further help to predict future trends in water quality using historical trend data, meteorological data, and other variables that can be good for making rational pollution management and mitigation strategies. Thus, this study will bring sustainable management in the long-term Malaysia's water resources through proactive measures for ensuring water quality and in turn people's health.

Objectives of the Project:

The primary aim of this research is to establish a resilient machine-learning model for predicting the levels of pollution in Malaysian river basins for regulating timely and sound preventive measures against pollution to ensure water resource protection.

Objectives:

1. To collect historical water quality data with preprocessing before being made available at the Department of Statistics of Malaysia (DOSM) to ensure its quality and consistency.
2. To develop at least two machine learning models: Random Forest and LSTM, from pre-processed data, to achieve an initial accuracy greater than or equal to 75%.
3. To evaluate the created models according to appropriate dimensions (RMSE, MAE, R-squared) and optimize them for better accuracy.

Scopes of the Project:

This research tends to create a machine-learning application that predicts pollution levels in the river basins of Malaysia which specifically takes on the following scope:

1. Data Collection: Historical records such as water quality values, weather data, and any other associated parameter-related data would be collected from the Department of Statistics Malaysia and other reputable sources.

2. Data Preprocessing: Preprocessing of these somewhat different initial data records for missing values, outliers, and inconsistencies is referred to as data cleaning.

3. Feature Engineering: Construction of the appropriate features from the raw data, such as the temporal features, hydrological features, and socio-economic indicators.

4. Model Building and Training: Varying machine learning algorithms and their application for training appropriate models on the willing, prepared dataset, these are Random Forests, Gradient Boosting, and LSTM.

5. Model Evaluation: Performing evaluations of the model that could be based on metrics such as RMSE, MAE, R squared, or through statistical tests.

6. Prediction and Visualization: Future forecasting from the trained model on pollution levels and the drawing of results.

.

Expected Contribution of the Project:

This research is expected to contribute in the following ways:

- Improved Water Quality Monitoring: Further reference-predicting research will lead to improved tracking of pollution hot points and risk assessment for water image quality.
- Pollution Control Adoption: Research insights can direct policymakers properly on When and Where Pollution Events Occur, so they can take timely countermeasures.
- Alimentation by Development of Machine Learning Applications: This study contributes to develop machine learning techniques as in what their application will be toward environmental science applications wherein these techniques will be used as water quality prediction.

Project Requirements:

Software:	• Python (with libraries like Pandas, NumPy, Scikit-learn, TensorFlow, Keras) • R • ArcGIS
Hardware:	Personal computer with sufficient processing power and storage capacity
Technology/Technique/ Methodology:	• Machine Learning • Deep Learning • Statistical Analysis • Data Mining • Data Visualization
Algorithm:	• Random Forest • Gradient Boosting • Support Vector Regression • Long Short-Term Memory (LSTM) networks

Type of Project (Focusing on Data Science):

[√] Data Preparation and Modeling

[√] Data Analysis and Visualization

[ ] Business Intelligence and Analytics

[√] Machine Learning and Prediction

[ ] Data Science Application in Business Domain

Status of Project:

[√] New

[ ] Continued

If continued, what is the previous title?

SECTION C: Declaration

I declare that this project is proposed by:

[√] Myself

[ ] Supervisor/Industry Advisor ( )

Student Name: Haslinda binti Abdul Sahak

Signature

15/11/2024

Date

SECTION D: Supervisor Acknowledgement

The Supervisor(s) shall complete this section.

I/We agree to become the supervisor(s) for this student under aforesaid proposed title.

Name of Supervisor 1: .....

Signature **Date**

Name of Supervisor 2 (if any): .....

Signature **Date**

SECTION E: Evaluation Panel Approval

The Evaluator(s) shall complete this section.

Result:

[ ] FULL APPROVAL [ ] CONDITIONAL APPROVAL (Major)*

[ ] CONDITIONAL APPROVAL (Minor) [ ] FAIL*

***** Student has to submit new proposal form considering the evaluators' comments.

Comments:

Name of Evaluator 1: .....

Signature **Date**

Name of Evaluator 2: .....

Signature **Date**