PREDICTIVE MODELING OF POLLUTION IN RIVER BASINS USING MACHINE LEARNING TECHNIQUE

HASLINDA BINTI ABDUL SAHAK
MCS241004_MCST 1043

CHAPTER 3

UNIVERSITI TEKNOLOGI MALAYSIA

DECEMBER 2024

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

It provides an exhaustive methodological framework to develop and validate predictive models to estimate pollution in river basins by using machine learning techniques. In other words, all data on the environment, water quality, and socioeconomic aspects would be harnessed to develop accurate and correctly actionable models for guiding mitigation strategies. In this way, one would not only systematically handle data but also develop specifically robust model selection so that one could use them in more realistic applications. The research process, such as how such a study begins and how the results and models are evaluated, is discussed more. In this chapter, the dataset used and its performance measurement will be introduced and illustrated.

3.2 Research Framework

The research framework for this study is structured to provide a comprehensive methodological approach to addressing the complexities of pollution prediction. There were five phases of research to develop predictive models for pollution in river basins. Phase one is planning and initial study which contributed to problem formulation and background research. A milestone of an overview of point of interest can be identified and enable an insight into the whole project. Besides that, data preparation fell into phase two in which a cleaned dataset that was ready for further analyzation was well-prepared. Furthermore, phase three is to retrieve the relevant features from the pre-processed dataset. In this phase, the underlying pattern of the dataset can be identified by machine learning model by LSTM, from pre-processed data. Additionally, random forest model will be implemented into the retrieved features to classify the data based on their similarities. In this case, a milestone for the source of river basin pollution can be illustrated. Lastly, R-squared was applied for model evaluation. The relationship between sources and river basin pollution can be visualized in this phase. Figure 3.1 illustrate the overall research framework. This will elaborate on every phase.

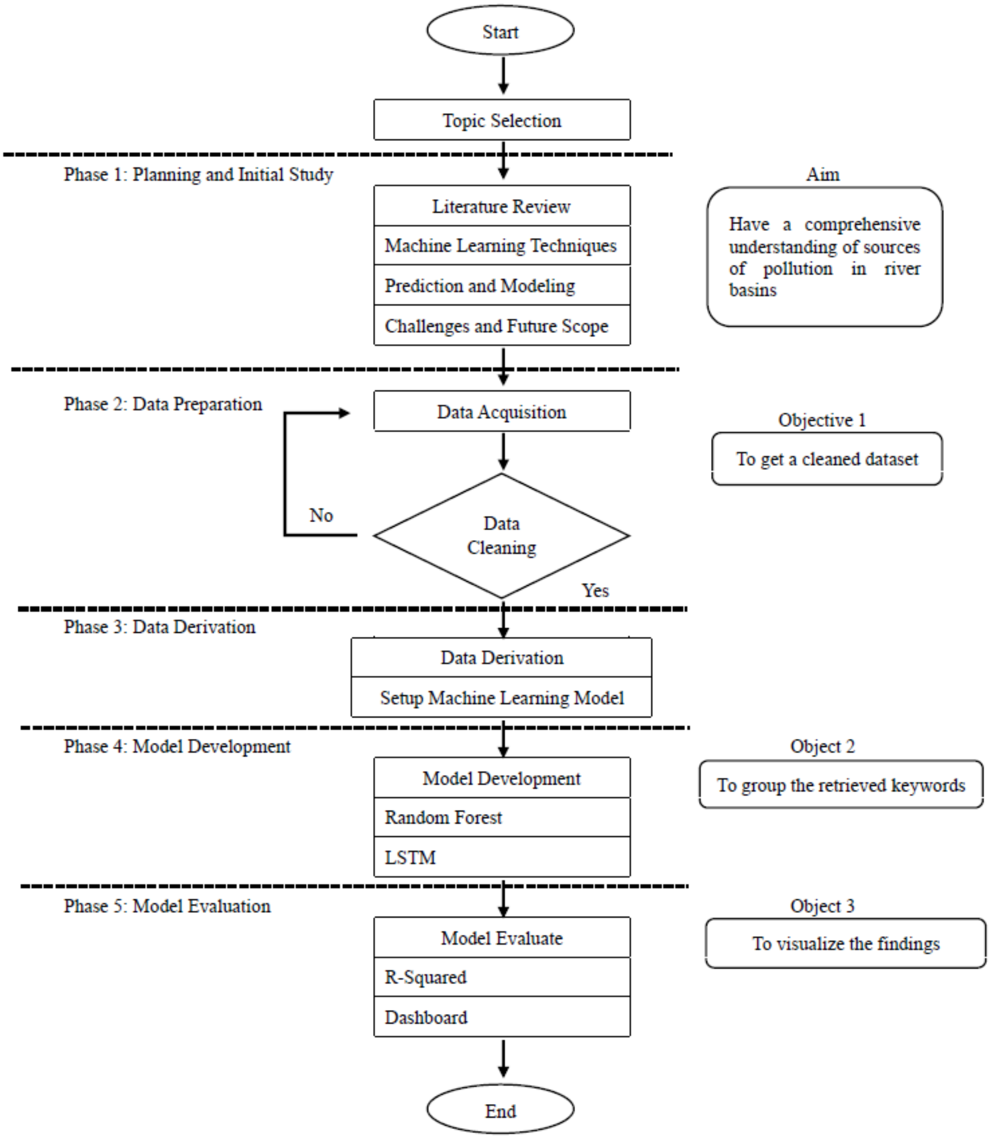
Start
Topic Selection
Phase 1: Planning and Initial Study
Literature Review
Machine Learning Techniques
Prediction and Modeling
Challenges and Future Scope
Aim
Have a comprehensive understanding of sources of pollution in river basins
Phase 2: Data Preparation
Data Acquisition
No
Data Cleaning
Yes
Objective 1
To get a cleaned dataset
Phase 3: Data Derivation
Data Derivation
Setup Machine Learning Model
Phase 4: Model Development
Model Development
Random Forest
LSTM
Object 2
To group the retrieved keywords
Phase 5: Model Evaluation
Model Evaluate
R-Squared
Dashboard
Object 3
To visualize the findings
End

Figure 3.1: Overall research framework

3.2.1 Planning and Initial Study

This phase is laying the groundwork for research by several significant strands. The major component is a detailed literature review to assess studies related to pollution of river basins, predictive modeling methods, and relevant data sources. Machine learning algorithms will also be looked into for potential suitability in time-series and spatial data to identify existing research lacunae and to guide the choice of appropriate models as well as course of action in data collection.

Then, that phase is problem definition and scope, which means the precise problem statement with regard to the research. This includes the identification of a target river basin or basins and pollutants of interest, for example specific chemical compounds or biological indicators, and also definitions in temporal and spatial scope of the study.

Research question development aligns with the SMART criteria which stand for specific, measurable, achievable, relevant, and time-bound. The questions will address the investigation's general aspects, such as the factors that have the most influence on pollution levels within the target river basins or the performance prediction of the different machine learning algorithms considered. Moreover, whether the developed models will generalize to unseen data and the practical implication of these predictive models for environmental management and policy-making are significant questions regarding this investigation. This approach gives a solid base for meaningful and powerful research.

3.2.2 Data Collection and Preparation

Identifying and using several sources of relevant data will be required for extensive coverage of the parameters concerning pollution and factors that affect them. The historical water quality monitoring data will be collected from monitoring stations with the target river basins, covering the parameters: biochemical oxygen demand (BOD), chemical oxygen demand (COD), dissolved oxygen (DO), pH, turbidity, nutrients and concentrations of heavy metals.

These include environmental data that should include factors such as rainfall patterns, temperature fluctuations, river flow rates, land use characteristics, and agricultural practices that may affect water quality. Socioeconomic data are equally important to include population density, industrial activities, capacity of wastewater treatment plants, and extent of agricultural activities within the basin. This integration of diverse approaches will result in the creation of a substantial and multi-faceted dataset that serves the analysis and predictive model for water basin pollution.

Data Preprocessing is the very first stage in preparing the data collected to be fit into machine learning algorithms. This step will include taking care of the missing values. Where appropriate techniques like mean imputation or regression imputation would have to be used to fulfill the dataset for it to be usable. Then, data normalization in itself is part of the process whereby the data set will be scaled from its initial range to a common range that helps in improving performance and convergence in machine learning.

Feature engineering is followed, in which new features are created either using existing variables or through the addition of domain knowledge to improve the prediction accuracy of models. Finally, the data sets are split into training, validation, and test sets. This is done so that systematic model building and evaluation can take place, which enhances performance assessment and reduces the chances of overfitting.

3.2.3 Data Derivation

This research data derivation is the extraction of useful information from already preprocessed data to base predictive modeling of pollution levels in the river basin. The process uses high-end language models like ChatGPT to derive insights and patterns to be incorporated into the machine learning pipeline.

The connection between the research system and ChatGPT has indeed been established using the OpenAI-provided API. Here, the acquiring of an API key and utilizing this key to interface with the data processing framework to provide requests are the 2 most important activities.

Once the API starts producing results, the next logical step is to establish prompt instructions that set the stage for the language model to operate in the specific task-oriented direction defined by the research outcomes. These instructions would be well-crafted to analyze preprocessed river basin data and extract informative insights from it. For example, prompts could instruct the model to identify pollutants present in water, determine seasonal patterns in that pollution, or even find areas with the most pollution, giving results in specified JSON formats.

The next step involves the processing of API responses that are formed in JSON according to the specification required by the output consumer for further analysis. This could be analysis in terms of grouping pollutants according to their source-for example, industrial, agricultural, or urban runoff-and how they affect river ecology. Such categorized findings allow the user to know more about pollution sources and their weight as severity.

Finally, the machine learning models use data that are derived through all of this predictive analysis. API structured outputs are integrated with pre-processed, such as meteorological and hydrological variables, into full datasets that inform up-to-the-moment inputs for training and validating predictive models on pollution predictions and environmental management.

The data derivation phase is the vital step that links raw data and works toward actionable insights in making sure that the machine learning models are fed relevant and context-rich data. Largely as a result of using LLMs, complex relationships and latent patterns in river basin data can also be revealed in a way that would otherwise be quite difficult to derive from traditional preprocessing methods alone. This would, thus, improve the predictivity of models in addition to creating a more accurate and efficient pollution prediction system.

3.2.4 Model Selection and Development

The process for selecting and developing a model consists identifying and applying appropriate machine learning algorithms to predict pollution levels on the characteristics of data and the objectives of the study. Different algorithms may be selected for a task, such as regression models such as linear regression, support vector regression (SVR), random forest regression, gradient boosting machines (GBM), extreme gradient boosting (XGBoost). Neural networks, such as multilayer perceptrons (MLP), recurrent neural networks (RNN), and convolutional neural networks (CNN), can also be employed, especially for capturing complex patterns in temporal or spatial data. Additionally, ensemble methods like bagging, boosting, and stacking are considered to leverage the strengths of multiple models for improved predictive performance.

Once the algorithms are selected, they are trained using the prepared training data, with optimization of parameters such as learning rates and regularization factors to enhance their performance. Model validation follows, where the trained models are assessed using a separate validation dataset to evaluate their accuracy and ability to generalize to unseen data. Key metrics for evaluation include R-squared (R^2), which measures the proportion of variance explained by the model, root mean squared error (RMSE), which quantifies the average magnitude of prediction errors, and mean absolute error (MAE), which calculates the average absolute difference between predicted and actual values. These metrics provide a comprehensive understanding of model performance and guide further refinements.

3.2.5 Model Evaluation and Interpretation

Model evaluation and interpretation involve assessing the performance of various machine learning models to identify the best one for predicting pollution levels. This is achieved by comparing the models using chosen evaluation metrics, such as R-squared, RMSE, and MAE, to determine which model delivers the highest accuracy and generalization ability.

Feature importance analysis is then conducted to understand the relative significance of different input features in influencing the predictions. Techniques like permutation feature importance, which involves randomly shuffling the values of a feature and observing changes in model performance, provide insights into how critical each feature is. Additionally, SHAP (SHapley Additive exPlanations) can be employed to explain the model's output by attributing contributions to individual features, offering a deeper understanding of their roles.

Model interpretability is also a key focus, as it helps uncover the relationships between input features and pollution levels. Visualization techniques, such as plotting predicted pollution levels against actual values, are used to assess model accuracy and identify patterns. Furthermore, analyzing feature interactions reveals how different variables work together to impact pollution levels, enhancing the interpretability and reliability of the model for practical applications.

3.2.6 Model Deployment and Application

Model deployment and application involve implementing the selected machine learning model for real-time prediction of pollution levels in the target river basins. This process ensures that the model is integrated into a practical system capable of generating timely and accurate predictions. A user-friendly interface is then developed to allow stakeholders to easily access model predictions and visualize results. This interface may include interactive dashboards, charts, or maps to make the outputs intuitive and actionable.

The findings from the research are communicated effectively to stakeholders, such as policymakers, environmental agencies, and local communities. This ensures that the insights derived from the model inform decision-making processes and support the development of targeted pollution mitigation strategies. By bridging the gap between complex data analysis and real-world application, this phase ensures the research has a tangible impact on environmental management efforts.

3.3 Dataset

The riven basins pollution review dataset was obtained from Department of Statistics Malaysia (DOSM) repository. It consists of 10,840,500 row of data representing the 150 river basins for three primary water pollution indicators monitored by the DOE presented in this dataset which is Biochemical Oxygen Demand (BOD), which measures the amount of oxygen that microorganisms require to decompose organic matter in the water. High BOD levels indicate the presence of a large amount of organic pollution, which can deplete oxygen in the water, harming aquatic life, Ammoniacal Nitrogen (NH_3-N), which is a measure of ammonia in the water primarily coming from agricultural runoff, sewage, and industrial discharges. High levels of ammoniacal nitrogen are toxic to fish and other aquatic organisms and can lead to the depletion of dissolved oxygen as it is broken down and Suspended Solids (SS), which are tiny particles suspended in the water, including soil, silt, organic matter, and industrial waste. High concentrations of suspended solids can reduce water clarity, block sunlight for aquatic plants, and clog the gills of fish. River basins are classified as clean, slightly polluted, or polluted based on Water Quality Index (WQI) values for each of these three indicators.

Table 3.1: Riven Basins Pollution Dataset

Date	**Basins Monitored**	**Pollution Indicator**	**Pollution Status**	**Number of Basins**	**Proportion**
2008-01-01	143	nh3n	polluted	33	23.1
2009-01-01	143	nh3n	polluted	40	28
2010-01-01	143	nh3n	polluted	42	29.4
2011-01-01	140	nh3n	polluted	35	25
2012-01-01	140	nh3n	polluted	38	27.1
2013-01-01	140	nh3n	polluted	41	29.3

3.4 Performance Measurement

R-squared (R^2) is a measure that tells us how well our machine learning model is performing in predicting pollution levels in river basins. It explains how much of the changes in pollution levels can be attributed to the factors we've included in the model, such as weather, waste from factories, or runoff from agriculture. For example, if R^2 is 0.85, it means that 85% of the variations in pollution levels are explained by the model, while the remaining 15% could be due to other factors not included in the model or random noise.

To use R^2, we first collect data on pollution levels and related factors, then train the machine learning model using this data to make predictions. We then compare the model's predictions to the actual pollution measurements to see how accurate it is. The closer the predicted values are to the actual values, the higher the R^2, will be. If the model's predictions are very close to reality, R^2 will be near 1, which indicates a good fit. If the predictions are far off R^2 will be closer to 0, showing the model is not performing well.

In this research, R^2 helps us evaluate how reliable our model is in predicting pollution levels based on the data we provide. A high R^2 indicates that the model is capturing the important relationships between the data and pollution, making it useful for decision-making and future predictions. A low R^2 suggests the model may need improvement or that additional factors should be considered.

3.5 Future Research Directions

Future research in this area can explore several advanced techniques to improve predictive modeling of pollution in river basins. One promising direction is the use of deep learning methods, such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, to capture complex temporal and spatial patterns that traditional models may miss. These techniques are particularly useful for understanding how pollution levels change over time and across different locations in the river basin. Additionally, hybrid models that combine physical approaches, like hydrological simulations, with statistical and machine learning methods could provide a more comprehensive understanding of pollution dynamics and improve prediction accuracy.

Another area for future research involves integrating diverse data sources to enhance model performance. Real-time data from Internet of Things (IoT) sensors deployed in and around the river basins, along with satellite imagery, could provide a continuous stream of data, offering more up-to-date and detailed information on environmental conditions. To address challenges like data sparsity, techniques such as transfer learning or synthetic data generation could be employed. Transfer learning allows models to leverage knowledge from related areas, while synthetic data generation can help create additional data for training when real data is limited.

Lastly, transparency in model development and outcomes is crucial for gaining trust from stakeholders. Future research should focus on improving the interpretability of models by using visualization tools and explainability techniques that make it easier to understand how predictions are made. This would help in building confidence in the models and ensure that their predictions are seen as reliable. Validating the models with independent datasets is another important step, as it would provide external verification of the model’s effectiveness and further enhance trust among stakeholders.

3.6 Conclusion

This methodology provides a robust framework for predictive modeling of river basin pollution using machine learning. By addressing each stage systematically, the research ensures scientific rigor, practical relevance, and actionable insights. Future work will focus on leveraging advanced technologies and methodologies to further improve model accuracy and applicability.