RESEARCH PAPER PROJECT

PREDICTIVE MODELLING FOR WATER QUALITY POTABILITY WITH MACHINE LEARNING

Authors: Kelsten Wuisan, Kevin Maxwell Andreas
Supervisor: Diana, S.Kom., M.T.I, Afhdal Kurniawan, S.T., M.Kom

Afiliation: School of Computer Science, Bina Nusantara University

BINUS UNIVERSITY
SCHOOL OF COMPUTER SCIENCE

ISCEE 2024 International Seminar

Abstract

Water is one of the crucial things for human life, with a total of 4 trillion cubic meters of fresh water used every year. In recent years, there has been a concern of water quality due to various reasons such as pollution and contaminants. Ensuring the sustainability of fresh water is a critical challenge with the growing population. This study presents a comprehensive approach to predictive modelling for water quality, which is capable of forecasting water quality trends, potential risks, and proactive management strategies. It aims to underscores the potential of predictive modelling as potential assets in the sustainable management of water resources, contributing to public health protection and environmental conservation by comparing each machine learning methods to find each strength and its limitations. In this paper the models used are Artificial Neural Network, Logistic Regression, Decision Tree, Support Vector Machine, Random Forest, Gradient Boost and Ensemble Method.

Introduction

Water is vital to human existence and serves a variety of purposes, including bathing, energy production, and digestive support. Concerns over water quality have increased recently as a result of pollution, population growth, and climate change. It is essential for drinking, agriculture, industry, and ecosystems to have access to clean and safe water. Even though clean water is essential, just 1% of the world's water is fit for human use, which makes it difficult for many people to get. The WHO estimates that 663 million people have difficulty accessing safe water. The Ministry of Health in Indonesia reported 314 child cases of diarrhea in 2019, most of which were brought on by tainted food or water. In 2021, Indonesia's Water Quality Index (WQI) was 53.33, which was lower than the 2020 WQI of 53.53 and below the national objective of 55.2.

Related Works

A review studies by Ariani D.A, Azmi A., Mohd. R.S., Shamila A, Salmiati and Mohd I. M. S, there are the limitations of traditional models in accurately predicting water quality due to some factors. These studies argues that AI approaches can help bridge these gaps and improve the precision of predictive models.

A study by Xian he Wang, Ying Li, Qian Qiao, Adriano Tavares and Yanchun Liang assess the predictive abilities of different machine learning models for water quality parameters using the entropy weighting method. The study examines five models: Support Vector Machine (SVM), Multilayer Perceptron (MLP), Random Forest (RF), XGBoost, and LSTM.

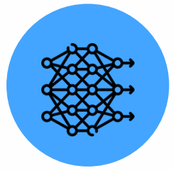
Another study by Zhao Fu, focuses on the use of artificial neural networks (ANN) and fuzzy time series forecasting (FTS) models for predicting water quality parameters in different water bodies. Zhao Fu conclude that ANN and FTS models, can be effective in predicting water quality parameters.

Dwi Hartantia and Afu Ichsan Pradanab also did a study on accuracy of Machine learning techniques which is Decision Tree, Logistic Regression, SVM, and ANN.

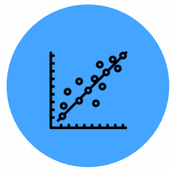
Using AI or Machine Learning models can be more efficient and more effective compared to traditional model for predicting the quality of water due to certain factors. It can cover the challenges with water quality assessment, including real-time sensor reading, historical data, weather patterns and geographical information for predicting water quality parameters. This prove that AI or Machine Learning models will be more efficient that traditional method. SVM, ANN, Logistic Regression, Decision Tree, RF are some of the best models that is suitable for predicting water quality cases.

Methodology

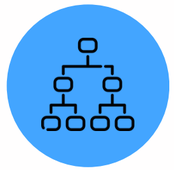
To guarantee data integrity, the procedure starts with data cleansing, which involves removing null values. After that, the features are standardized using Standard Scaler to make sure every feature contributes equally to the model. In order to determine the most pertinent characteristics influencing water quality, feature selection is carried out.
This research uses a comparison of several methods related to machine learning, such as:



Artificial Neural Network

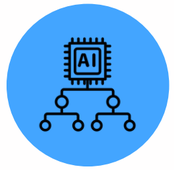


Logistic Regression



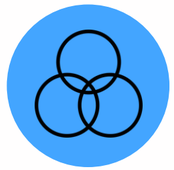
Decision Tree

Support Vector Machine



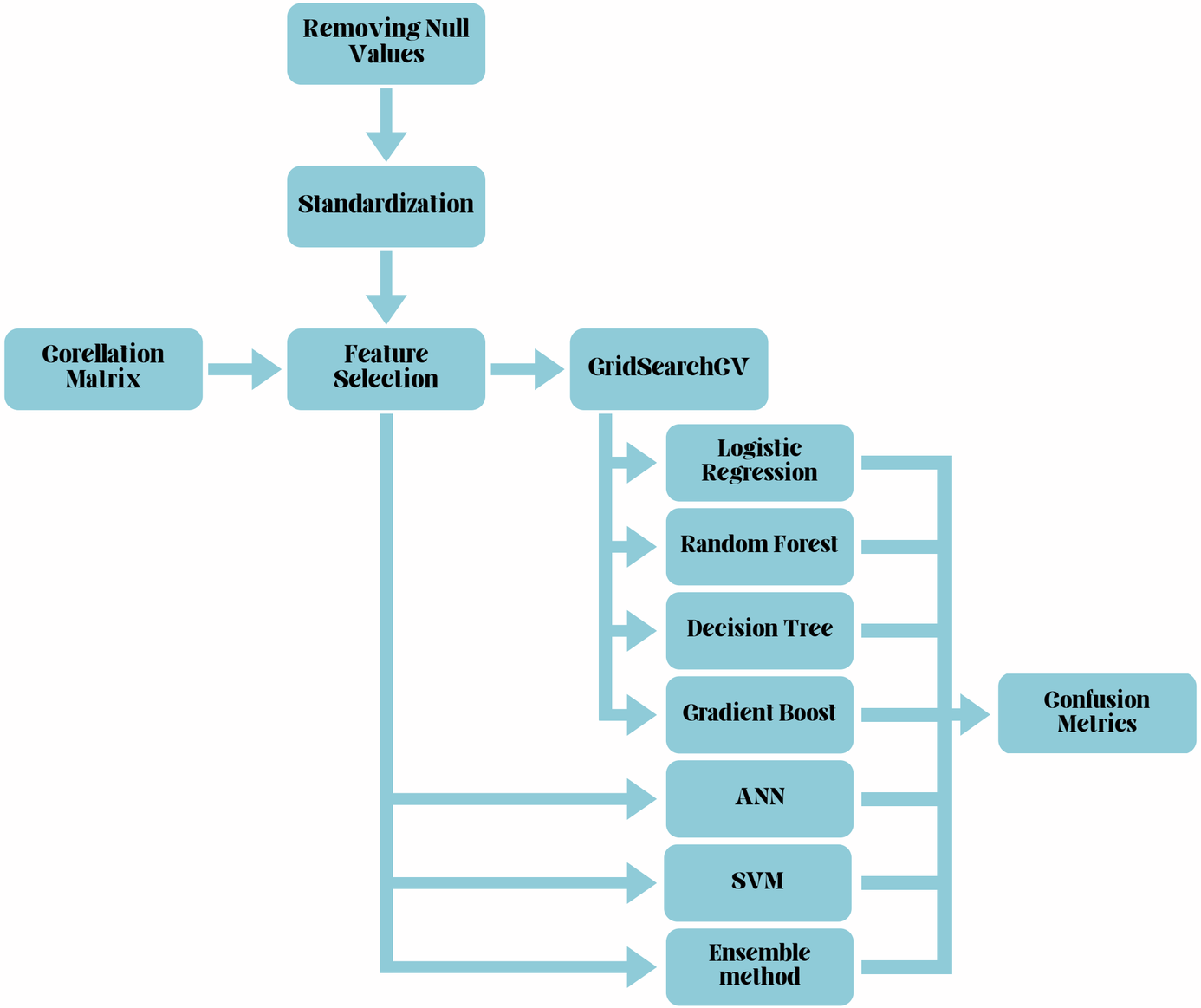
Random Forest

Gradient Boost

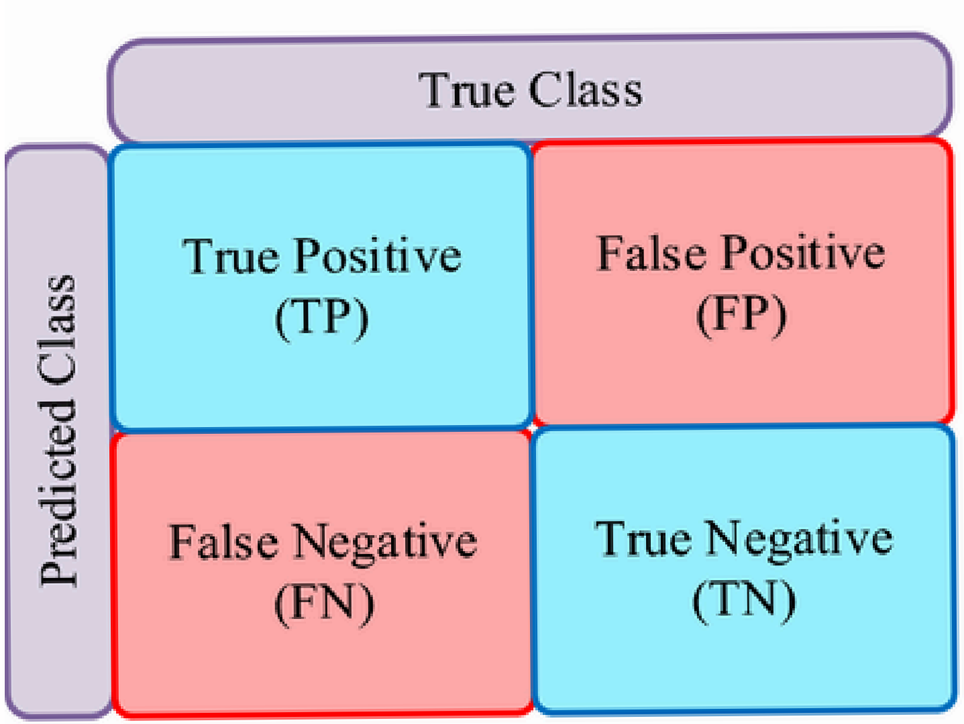


Ensemble Method

Methodology



Confusion Matrix



Function Name	Formula
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$
F1 score	$\frac{2.Precision.Recall}{Precision + Recall}$

Result and Discussion

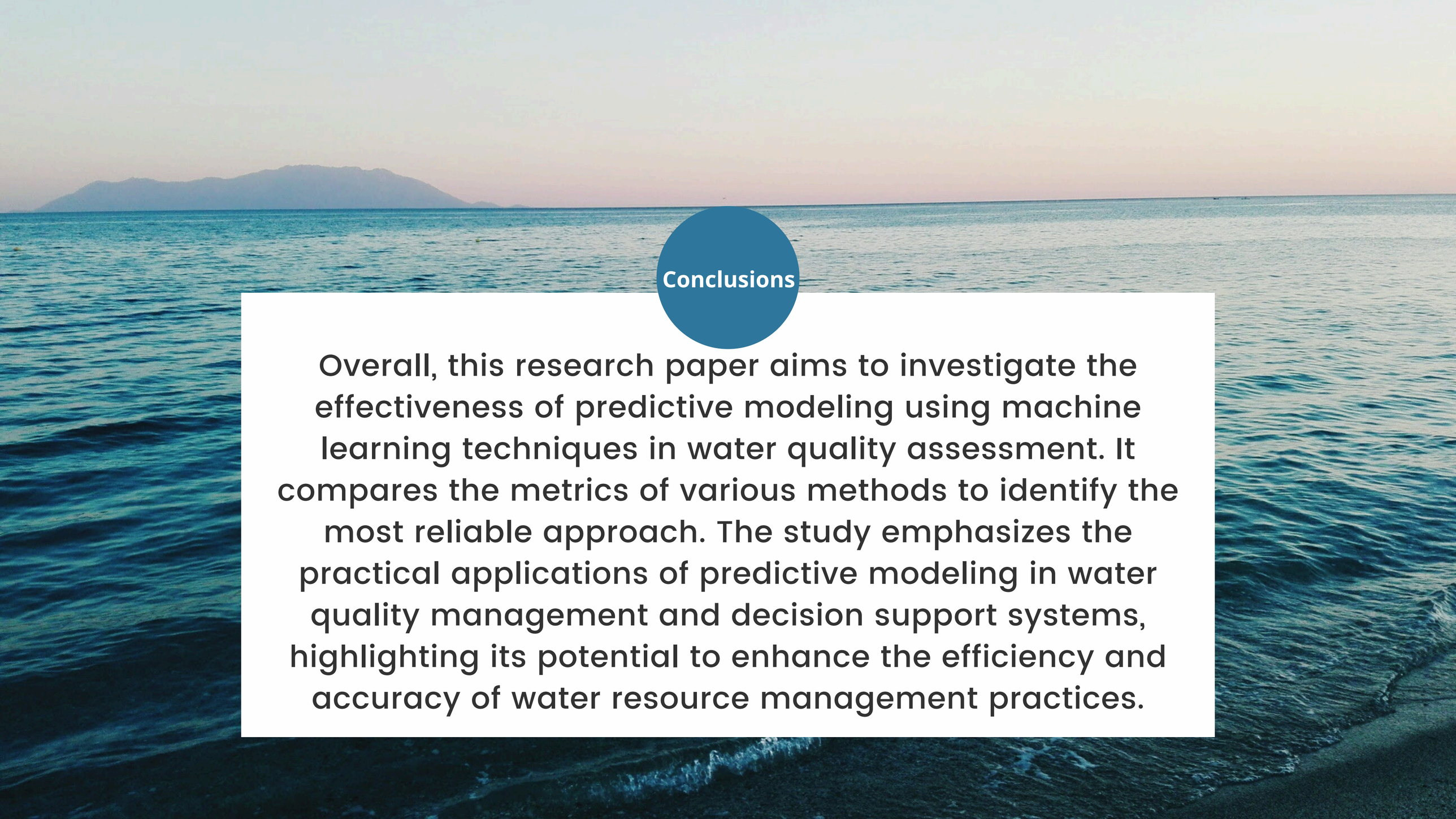
The model's exactness generally cannot be decided exclusively from the disarray lattice. Be that as it may, it appears that the demonstrate may have the next exactness (extent of genuine positives among all anticipated positives) than review (extent of genuine positives among all real positives), given the generally higher number of wrong positives compared to wrong negatives. This indicates Support Vector Machine methods has the highest Precision and Accuracy rate with 73% precision rate, 35% recall rate, 47% F1-Score and 71% accuracy. With the result of our analysis, it shows the most potential of using machine learning models for predicting water quality. The Support Vector Machine method can get the highest accuracy because this method can handle the complexity of the data such as the features it has, besides that SVM is strong in dealing with outliers which usually appear in environmental datasets. And finally, SVM's ability to find the optimal hyperplane helps in generalizing well to unseen data, crucial for accurate potability predictions.

Model Result Metrics

	Precision	Recall	F1-Score	Accuracy
Artificial Neural Network	62%	35%	48%	68%
Logistic Regression	0%	0%	0%	63%
Support Vector Machine	73%	35%	47%	71%
Decision Tree	53%	29%	37%	64.2%
Random Forest	70%	67%	68%	68%
Gradient Boost	59%	35%	44%	66.9%
Ensemble Method	68%	33%	45%	69%

CONCLUSION

RESEARCH PAPER PROJECT



Conclusions

Overall, this research paper aims to investigate the effectiveness of predictive modeling using machine learning techniques in water quality assessment. It compares the metrics of various methods to identify the most reliable approach. The study emphasizes the practical applications of predictive modeling in water quality management and decision support systems, highlighting its potential to enhance the efficiency and accuracy of water resource management practices.

Conclusions

The research contributes to the selection of the most suitable machine learning methods for addressing specific water quality challenges and achieving sustainable water resource management goals. By comparing the strengths and limitations of different models, the study helps in identifying the best approach for addressing various water quality issues. The findings from this study will improve our understanding of aquatic systems and support evidence-based decision-making for the protection and conservation of water resources globally.

Through the research, Support Vector Machine method has the highest accuracy out of 7 methods from what has been tried. From this it can be seen that SVM is proven to be superior in handling the complexity and high dimensionality of water quality parameters and is able to avoid overfitting. Therefore, SVM is an effective tool for predicting water potability, making significant contributions in water resources management and public health-related decision making.

Thank You