SUNTEXT REVIEWS

SunText Review of Economics & Business
ISSN: 2766-4775

Open Access
Research Article
Volume 5:4

Predictive Modeling of Water Quality and Sewage Systems: A Comparative Analysis and Economic Impact Assessment Using Machine Learning

Md Fakhrul Islam Sumon[1], Arifur Rahman[1], Pravakar Debnath[2], MD Rashed Mohaimin[3], Mitu Karmakar[1], MD Azam Khan[1] and Hossain Mohammad Dalim[1]

[1]*School of Business, International American University, Los Angeles, California, USA*
[2]*School of Business, Westcliff University Irvine, California, USA*
[3]*MBA in Business Analytics, Gannon University, Erie, PA, USA*

***Corresponding author:** Md Fakhrul Islam Sumon, School of Business, International American University, Los Angeles, California, USA; E-mail: sumonf836@gmail.com

Received date: 17 November 2024; **Accepted date:** 21 November 2024; **Published date:** 27 November 2024

Citation: Md Fakhrul Islam Sumon, Rahman A, Debnath P, MD Rashed Mohaimin, Karmakar M, MD Azam Khan, et al. (2024) Predictive Modeling of Water Quality and Sewage Systems: A Comparative Analysis and Economic Impact Assessment Using Machine Learning. SunText Rev Econ Bus 5(4): 217.

DOI: https://doi.org/10.51737/2766-4775.2024.117

Abstract

Maintaining high water quality and effective sewage systems is imperative for the USA's environmental sustainability and public health. Present issues related to water quality management and effectively working sewage systems in the USA are multi-dimensional. Aging infrastructure, lack of treatment facilities, and the absence of real-time monitoring systems are major impediments to maintaining water quality. This study aimed at resolving the pressing matters associated with water quality and sewage system efficiency through a multi-faceted approach. The research project strived to ascertain the relationship between sewage system efficiency and overall water quality. Besides, the present study endeavored to utilize machine learning techniques to develop forecasts of future trends in water quality. The datasets were gathered from as many reliable governmental databases as possible and environmental monitoring agencies to ensure robust and correct analysis. Among other sources, the national water quality databases include USGS, EPA, and EEA. These sources provided comprehensive data on a wide range of water quality parameters, such as pH levels, dissolved oxygen (DO), biological oxygen demand (BOD), chemical oxygen demand (COD), turbidity, nitrate and phosphate concentrations, and the presence of heavy metals like lead, mercury, and cadmium. In this research project, three evidence-based algorithms were selected, notably, Linear Regression, Random Forest, and XG-Boost are three algorithms of machine learning that have been selected for performing predictive modelling. Several performance metrics of the classes were performed for the stringent assessment of the performance of Recall, Accuracy, Precision, and F1 Score machine learning models. The performance of the Random Forest Classifier achieved an outstanding accuracy as compared to other models. The findings of this study have great implications for water quality management in the USA, especially concerning how predictive models could be leveraged further to advance monitoring and intervention strategies. This provides the possibility to combine machine learning algorithms in water quality management agencies that go beyond regular reactive approaches to proactive data-driven strategies.

Keywords: Predictive modelling; Water quality; Sewage system; Economic impact; Machine learning; Random forest

Introduction

Motivations and background

Maintaining high water quality and effective sewage systems is paramount for the USA's environmental sustainability and public health. Clean water is required not only for drinking purposes but also for agriculture, industry, and ecosystem support. Equally important efficient sewage systems whereby contaminants do not enter the water bodies and aquatic life is well protected, ensuring safe water for human usage [1]. The completely unexpected pace of urbanization and industrialization, even climate changes, has worsened the challenge of managing water quality and sewage systems in the USA. These factors add pollutants to freshwater

bodies and overload water sewage infrastructure hence are inefficient and may be disastrous [2]. According to Akhlaq, current problems pertinent to water quality management and effectively working sewage systems are multi-dimensional [3]. Aging infrastructure, lack of treatment facilities, and the absence of real-time monitoring systems are major impediments to maintaining water quality. There is also the added prevalence of substances like heavy metals, pharmaceuticals, and microplastics that most methods of treatment cannot effectively deal with. Ejaz, indicated that sewage systems within many regions are also ill-equipped to deal with the raised volumes produced by booming populations; this mostly results in discharge into the environment with no or partial treatment [4]. These are challenges that call for drastic, innovative solutions that will help manage water quality more efficiently and the processes involved in sewage treatment. Ahmed, argued that the economic and health repercussions of poor water quality and insufficient sewage systems are significant [5]. Contaminated water represents a potential source for the spread of waterborne diseases such as cholera, typhoid, and hepatitis, causing serious public health hazards, especially in low-income communities. Furthermore, Ameer, asserted that poor water quality reduces agriculture and fisheries, hence creating food-insecure communities where people lose their sources of livelihood [6]. Economically, health costs, loss of man-hours, and environmental clean-up of poor water management are very high. Therefore, investment in efficient sewage systems and water quality management is not just a question of public health but an economic one.

Objective

This study aims to resolve the pressing matters associated with water quality and sewage system efficiency through a multi-faceted approach. First, the research project will strive to ascertain the relationship between sewage system efficiency and overall water quality. The various indicators in which these sciences are interrelated, such as levels of pollutants, efficacy of treatment, and sewage system capacity, will be studied. The second objective is to compare the governance of water quality in different regions to draw upon best practices and deficiencies in these areas, mainly within an urban and rural setting. The third critical objective will involve the investigation of the economic implications of sewage systems that are inadequate and have poor water quality. Understanding the economic impacts of such consequences provides policymakers with information to help prioritize investments in future water and sewage infrastructure. Lastly, this present study endeavors to utilize machine learning techniques to develop forecasts of future trends in water quality.

Literature Review

Water quality and sewerage systems

As per Asadollah [7], the maintenance of water quality is governed by some major parameters and standards, which are used as yardsticks for safety and usability in drinking applications or other uses in agriculture or industry in America. Key parameters monitored include pH, dissolved oxygen, turbidity, BOD, and the presence of contaminants such as heavy metals and pathogens within national and internationally accepted standards. Based on this, organizations like WHO and EPA have set guidelines that stipulate permissible limits for such parameters, which make the water safe for consumption and use. With these standards, nonetheless, sewage systems in the USA are faced with various issues conflicting with water quality. Among the most common problems that plague sewage systems include aged infrastructure, incomplete treatment facilities, and poor disposal of industrial and household waste. These frequently end up in discharging untreated or partially treated sewage into natural water bodies, thereby contaminating freshwater. The inefficiency of sewage systems is one of the major contributory causes affecting water quality, especially in the case of urban areas, where wastewater production exceeds the capacities of existing facilities [8]. Many studies have been conducted on different modifications in sewage treatment techniques, including advanced filtration technologies, bioremediation techniques, and optimization in sewage network designs. These studies bring out the need for integrated solutions with multi-faceted, multifactorial, technical, and policy-related challenges in water quality management [9].

The economic impact of water quality

The economic ramifications of poor water quality and sub-standard sewage systems are profound and far-reaching. Poor water quality induces the spread of waterborne diseases, increases healthcare costs, and lessens workforce productivity. The economic load is heavier on low-income communities that have no access to clean water and efficient sewage systems, often resulting in socioeconomic disparities in the long term. Research has documented that communities affected by poor water quality endure increased medical expenses, lower agricultural yields, and reduced property values which feed a self-reinforcing cycle of poverty and economic instability. Gorenekli & Gulbag, posited that case studies from various parts of the world have been indicating large economic burdens of water pollution [10]. For instance, the research on the Ganges River in India showed that contamination of this vital watercourse has serious health consequences and is extremely expensive regarding healthcare, tourism, and fisheries. To the same extent, research on the Flint water crisis in the United

States has demonstrated several long-term economic consequences observable in the community, which vary from lower house property values to higher public health expenditures. These examples illustrate a dire need for investments in water quality improvement and sewage system upgrades that could help reduce economic losses [11].

Machine learning in environmental management

According to Van Nguyen [12], in the recent past, machine learning (ML) has emerged as an instrumental tool in environmental management, specifically in forecasting and mitigating the impacts of pollution. The algorithms of machine learning can analyze huge data to predict patterns and trends which may not emerge conventionally by statistical methods. Other applications of ML in environmental sciences include air quality indices prediction, modelling scenarios of climate change impacts, and assessment of trends in water quality. Predictive capabilities make for proactive environmental management such that interventions can be taken in time, which may prevent or reduce pollution. Zhu, articulated that application of machine learning in water quality prediction has already witnessed several accomplishments [13-30]. For instance, research has demonstrated that ML models can predict the concentration of certain contaminants, such as nitrates, and phosphates-continuing and vital water quality indicators. These models have been applied in a decision-support context for the management of water resources to enable public authorities to take precautions guaranteed to safeguard human health and protect the natural environment. Despite these successes, there are limits to how machine learning can be applied in this domain. There are several conditions when performance for the ML models depends extensively on the quality and amount of the input data. Water quality data are scarce or inconsistent in many parts of the world. Generalizing it across geographical and socio-economic contexts may be problematic since the environmental systems are very complex. Nevertheless, the role of machine learning could prove highly influential in changing the way water quality management is done, especially concerning improving collection and processing technologies.

Data Collection and Pre-processing

The foundation of this study lies in the extensive collection and analysis of datasets associated with water quality and sewage system efficiency. The datasets were gathered from as many reliable governmental databases as possible and environmental monitoring agencies to ensure robust and correct analysis. Among other sources included the national water quality databases include USGS, EPA, and EEA. These sources provided comprehensive data on a wide range of water quality parameters, such as pH levels, dissolved oxygen (DO), biological oxygen demand (BOD), chemical oxygen demand (COD), turbidity, nitrate and phosphate concentrations, and the presence of heavy metals like lead, mercury, and cadmium. These range from critical indications of water quality to the water body's health and its suitability for use by humans, aquatic life, and agriculture.

Data-pre-processing

Step 1-Datetime Handling: First, 'Sampling Date' was converted into a proper date-time format using pd.to_datetime (), where coercion of parsing errors is enabled. This procedure enabled a wide range of data manipulations and extractions that can be performed efficiently later in the process.

Step 2-Encoding of Categorical Variables: Label encoding was performed over the categorical column 'State of Sewage System'. This protocol transformed the text categories into numerical values, which are more suitable for machine learning algorithms.

Step 3-Handling Missing Values: df. Isnull ().sum () code checked for missing values in the dataset, indicating gaps that might not have been originally included. For continuous numerical columns like 'Nitrogen (mg/L)' and 'Phosphorus (mg/L)', missing values were imputed using the mean. In the case of date-time data, the mode is used to fill in the missing dates so that there will not be any gaps in the dataset for analysis.

Step 4-Feature Engineering: New features 'Year', 'Month', and 'Day' were extracted from 'Sampling Date' to capture the temporal patterns in data. This helped in improving the model performance by leveraging time-based trends. After feature extraction, the original column 'Sampling Date' will be dropped as it's not needed anymore in its earlier form.

Step 5-Scaling Numerical Features: StandardScaler () code standardized the numerical features, including geographical coordinates and nutrient levels. This normalizes the feature values into a scale that is similar, which may be important for algorithms sensitive to the magnitude of features.

Step 6-Data Split: The last step divided the dataset into the necessary training and testing subsets by applying the 80-20 split using train_test_split with test_size=0.2. For the given problem, the target variable was the 'State of Sewage System', while the rest of the features were the predictor variables. Setting a random state ensures the reproducibility of the split.

Exploratory data analysis (EDA)

The above graphs outline two of the most important water quality parameters, Nitrogen in mg/L on the left and Phosphorus in mg/L on the right. The histograms, together with kernel density estimates, are reasonably symmetrical and close to normally distributed, though not without obvious multimodal happenstances. The Nitrogen levels make a cluster around an

average of 0 after scaling, probably standardization; the greatest part of the data lies between -1.5 and +1.5 along the scaled axis-data was transformed to have a mean close to zero. Also, in that respect, the spread and center of the Phosphorus levels are similar, which suggests that both features were normalized similarly. This is a relatively even distribution with no extreme peaks or troughs, suggesting that the dataset is considerably well-balanced with the least skewness feature good for machine learning models since such a distribution likely means no serious outliers or biases in those variables. Tiny fluctuations of frequency could suggest that there is some natural variation in environmental measures but do not indicate serious imbalances or abnormalities (Figure 1).

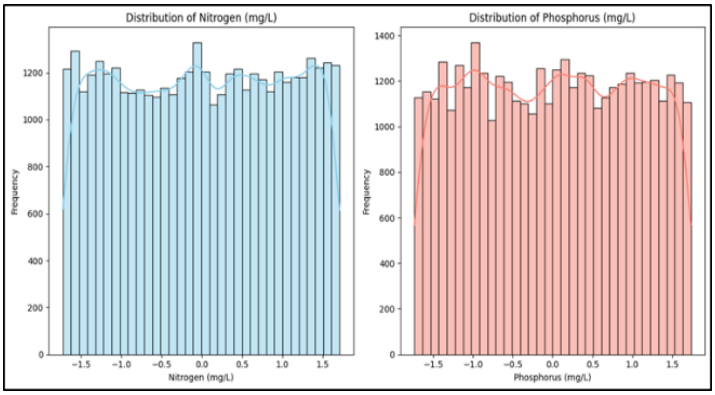

Figure 1: *Showcases the Distribution of Nitrogen and Phosphorous.*

Above is the correlation heatmap showing various feature relations-geographical coordinates, water quality parameters, sewage system state, and temporal components such as Year, Month, and Day. Out of these, the 'State of Sewage System' is very poorly correlated with Nitrogen - 0.01 and Phosphorus - 0.00, which means the effective factor of sewage systems within this dataset does not linearly affect these nutrient levels.

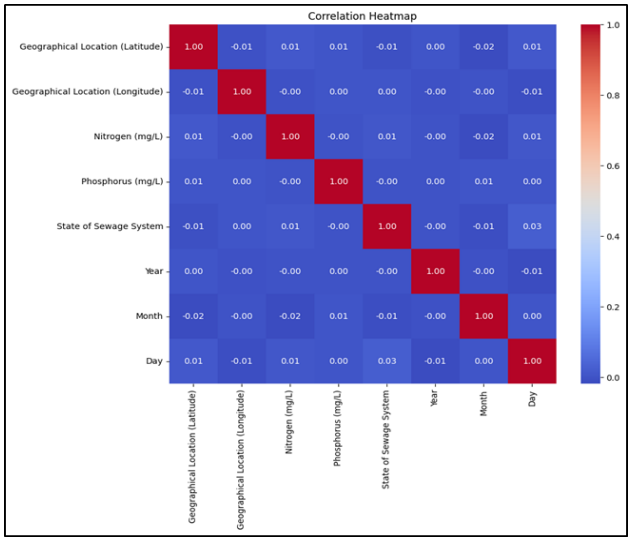

Figure 2: *Depicts the Correlation Heatmap of Various Features.*

The geographical factors such as Latitude and longitude, along with temporal features such as Year, Month, and Day, get less than minimal correlation from the water quality parameters and sewage system efficiency. No strong variable correlations existed; hence, these features will be almost independent and perhaps require extensive, complex nonlinear modelling approaches to find the underlying pattern in the data. This independence also would mean that no single feature is dominant in the dataset, hence a more balanced input to any machine learning model (Figure 2).

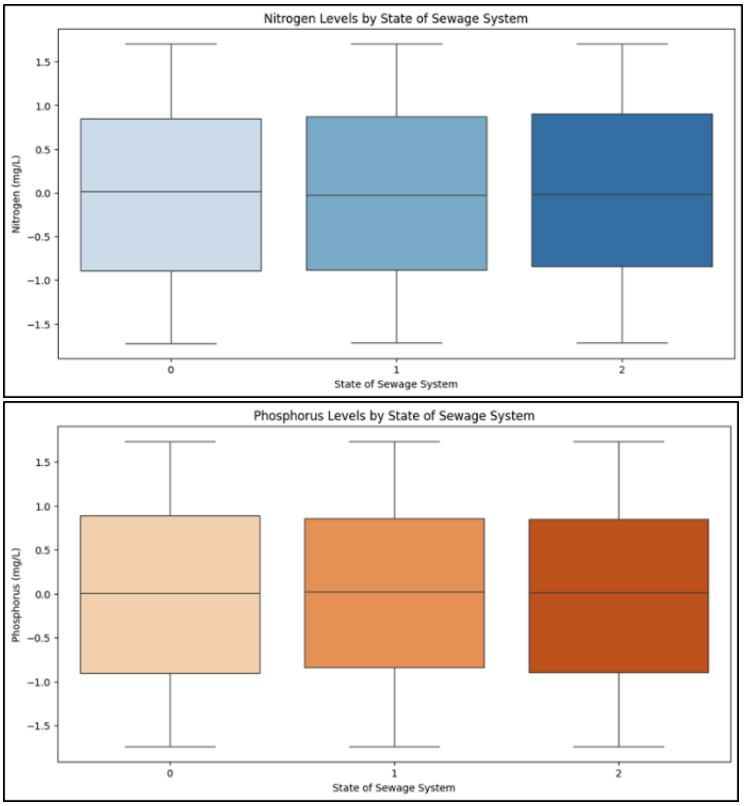

Figure 3: *Displays the Nitrogen & Phosphorous Levels by State of the Sewage System.*

The box plots above compare nitrogen and phosphorus levels across three states, 1, and 2 of sewage systems. In both nutrients, the patterns of distribution are similar across all three states, each with median values around 0 mg/L and ranging from approximately -1.75 to +1.75 mg/L.

There is a slight trend of increase in the dispersion or box size for both nitrogen and phosphorus levels as the state number increases from 0 to 2, but it is minimal. The symmetrical distribution of values around the median would indicate that in all states, normal distribution patterns are reflected by outliers shown by whiskers extending similarly in both positive and negative directions. Such consistency among states shows that the nutrient levels of the sewage system are relatively stable regardless of whether it is operational or not in operational (Figure 3).

Citation: Md Fakhrul Islam Sumon, Rahman A, Debnath P, MD Rashed Mohaimin, Karmakar M, MD Azam Khan, et al. (2024) Predictive Modeling of Water Quality and Sewage Systems: A Comparative Analysis and Economic Impact Assessment Using Machine Learning. SunText Rev Econ Bus 5(4): 217.

The time series plot above shows the monthly trend of nitrogen and phosphorus levels, ranging from 2012 to 2024. Both nutrients have similar oscillating patterns around 0 mg/L. The data indicates high-frequency fluctuations in both nutrients, generally within the range of -0.25 to 0.25 mg/L. Notable features include the strong peak in nitrogen to approximately 1.0 mg/L and the sudden drop in phosphorus to around -0.5 mg/L toward the end of this time series. The shaded areas around each line represent confidence intervals or uncertainty ranges and show a relatively consistent variance over this monitoring period. Both nutrients are on the same trend of seasonality or even cyclicality; no high long-term upward or downward trend until those anomalous readings at the end of the series (Figure 4).

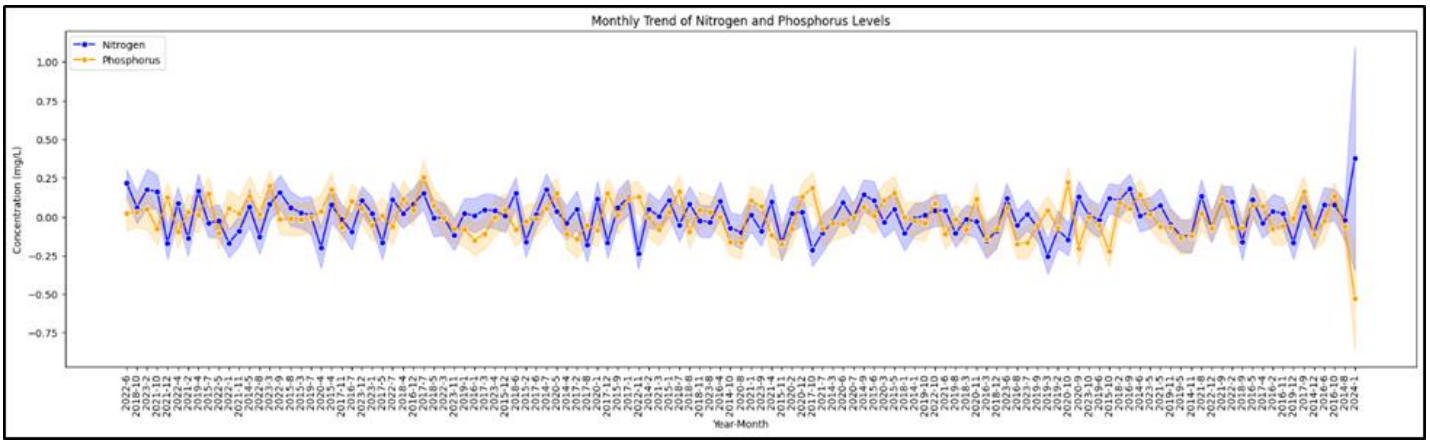

Figure 4: *Visualizes Monthly Trend of Nitrogen and Phosphorous Levels.*

Methodology

Feature engineering and selection

Feature engineering and selection are some of the most critical stages in the creation of any machine learning model, especially when dealing with environmental data. Therefore, diverse different techniques were used in the project to extract and engineer useful features from the raw data. In particular, we decomposed temporal data from 'Sampling Date' into separate features like 'Year', 'Month', and 'Day' to capture seasonal patterns that may influence water quality. Categorical variables were represented by the 'State of Sewage System', pre-processed into a numerical encoding using label encoding. The reason for doing this was to convert the textual data into a machine-readable format. Feature scaling was applied to numerical variables such as 'Nitrogen (mg/L)' and 'Phosphorus (mg/L)'. This is a process that scales those variables within a standard range, hence improving model convergence during the training process. Therefore, only those statistical methods, such as correlation analysis, were applied for the selection of the most predictive features, taking into consideration variables that show low multicollinearity to avoid redundancy and overfitting. The aim was to retain those features that contribute substantially to the target variable 'State of Sewage System', ensuring a balanced model with both accuracy and interpretability.

Model selection and justification

In this research project, three evidence-based algorithms were selected, notably, Linear Regression, Random Forest, and XG-Boost are three algorithms of machine learning that have been selected for performing predictive modelling. Linear Regression was chosen because it is very simple and efficient at capturing the linear relationship of independent variables with the target. Therefore, this may act as a baseline model to understand the direct influence of features on sewage system efficiency. Random Forest, an ensemble method based on decision trees, was adopted because it can provide a complex nonlinear interaction without severe over-fitting via bootstrapping and randomness in features. It is efficient in capturing intricate interactions between features and gives feature importance scores, which will be useful in further feature selection. On the other hand, XG-Boost was chosen for its excellent performance against large datasets with high dimensionality. It combines the strengths of gradient boosting with regularization techniques; hence, being highly effective at optimizing accuracy with lesser overfitting. XG-Boost is acknowledged to be one of the most efficient and scalable algorithms in data science competitions. Hence, it is suitable for this project: an accurate prediction of water quality trends.

Training and testing framework

In this research project, the dataset has been divided into an 80-20 split to ensure that the model captures 80% of the data to train on and is tested on 20%. This protocol helped in assessing the

generalization capability of the model. To further increase the robustness in evaluating the model, k-fold cross-validation was performed with k=5. It implies splitting the training data into five folds, training the model sequentially on four folds while validating on the fifth, through all possible rotations. Cross-validation helps prevent the problem of overfitting by ensuring that the performance of a model is consistent across different subsets of the data. Besides, hyperparameter tuning is also done through a grid search for better performance regimes of the model parameters. Performance metrics evaluated are MAE, RMSE, and R-squared were used to assess model accuracy and robustness.

Hyperparameter tuning

Optimizing model performance involves tuning the hyperparameters, which control the learning process and behaviour of machine learning algorithms. In this study, two major approaches were used for hyperparameter tuning, namely: Grid Search and Random Search. In Grid Search, the approach considers a pre-defined set of combinations of hyperparameters to explore systematically and retrieve the best parameters that maximize model performance. In contrast, Random Search selects random combinations of hyperparameters within specified ranges. The latter approach is much quicker for large parameter spaces compared to Grid Search and therefore best suited to efficiently explore large parameter spaces. It was especially helpful at the beginning of the experimentation for quickly determining promising bounds of hyperparameters for further fine-tuning. Using Grid Search when precision is important and Random Search when speed is important yields a good balance in optimizing model performance while avoiding extreme computational costs.

Performance evaluation metrics

Several performance metrics of the classes were performed for the stringent assessment of the performance of Recall, Accuracy, Precision, and F1 Score machine learning models. These metrics gave a complete understanding of the effectiveness that models may have, especially in cases where classes are highly imbalanced, or the costs of false positives and false negatives are very different. In the baseline testing performance of selected models Random Forest and XG-Boost-their evaluation metrics are compared to those of some baseline model, such as Logistic Regression or a Decision Tree classifier. This baseline provides a reference to allow qualification of the added value when using more sophisticated algorithms. Baseline models are characterized by decent accuracy, for example, but they may be substantially worse about recall and precision, especially events that occur less often such as severe sewage problems.

Results

Descriptive Analysis

Performance Metric	Random Forest	XG-Boost	Logistic Regression
Accuracy	99.60%	82.40%	50.29%
Precision [class 0]	0.99	0.77	0.50
Precision [class 1]	1.00	0.91	0.00
Precision [Class 2]	1.00	0.96	0.00
Recall [class 0]	1.00	0.97	1.00
Recall [class 1]	0.99	0.73	0.00
Recall [Class 2]	0.99	0.58	0.00
F1-Score [Class 0]	0.99	0.86	0.67
F1-Score [Class 1]	1.00	0.81	0.00
F1_Score [Class 2]	1.00	0.72	0.00

The Table above displays the performance results comparing three models: Random Forest, XG-Boost, and Logistic Regression. The best classification performance, according to the above table, is from the Random Forest, which yields an accuracy of 99.60%. Compared to other models, it depicts powerful performance among all metrics, including perfect or near-perfect precision, recall, and F1-scores belonging to all classes. The XG-Boost model follows, presenting an accuracy of 82.40% only. The performance of XG-Boost for the two classes is significantly lower than for the other two methods, with significant differences in recall and F1-score measures. Logistic Regression, in turn, performs considerably worse, yielding an accuracy of only 50.29%, completely misclassifying classes 1 and 2, while performing quite well for class 0. This finding also confirms the robustness of Random Forest on this data set, while the performance of Logistic Regression is comparatively poor in terms of multi-classification tasks.

Model performance

Logistic regression

Table 1: *Portrays the logistic Regression Modelling.*

```
# Logistic Regression
log_reg = LogisticRegression(max_iter=1000, random_state=42)
log_reg.fit(X_train, y_train)
y_pred_log_reg = log_reg.predict(X_test)

# Evaluation
print("Logistic Regression Results:")
print("Accuracy:", accuracy_score(y_test, y_pred_log_reg))
print("\nClassification Report:\n", classification report(y_test,
y_pred_log_reg))
print ("\nConfusion Matrix:\n", confusion_matrix(y_test, y_pred_log_reg))
```

The code above performs binary classification using the Logistic Regression model. First, the model is instantiated with a maximum iteration of 1000 and a random state for reproducibility. Then it fits into X_train and y_train data using the fit () method and makes predictions on data X_test. The code also includes an extensive evaluation section that prints several performance metrics: the accuracy score of the model; the detailed classification report which, among others, includes precision, recall, and F1-score; and finally, it also prints a confusion matrix. These are enough to provide a comprehensive review of the model's performance in classifying test data (Table 1).

Output

Table 2: *Presents the Logistic Regression Classification Report.*

Classification Report:

	precision	recall	f1-score	support
0	0.50	1.00	0.67	4031
1	0.00	0.00	0.00	2519
2	0.00	0.00	0.00	1466
accuracy			0.50	8016
macro avg	0.17	0.33	0.22	8016
weighted avg	0.25	0.50	0.34	8016

As showcased above, Logistic regression had an average performance of 50.3%. From the classification report, serious issues can be identified: only class 0 examples are classified correctly; it has a precision of 0.50 with a recall of 1.00, indicating that it predicts everything as class 0. This dataset is imbalanced, with the following distribution: class 0 with 4,031 samples, class 1 with 2,519 samples, and class 2 with 1,466 samples. It is confirmed by very low metrics for the macro average, an unweighted mean across classes, and weighted average, which refers to different metrics weighted averages considering the class supports. The macro average F1-score of 0.22 and weighted average F1-score of 0.33 lead us to believe that this model was average; important ameliorations need to be performed (Table 2).

Random forest

Table 3: *Depicts the Random Forest Modelling.*

```
# Random Forest Classifier
rf_clf = RandomForestClassifier(n_estimators=100,
random_state=42)
rf_clf.fit(X_train, y_train)
y_pred_rf = rf_clf.predict(X_test)
# Evaluation
print("\nRandom Forest Classifier Results:")
print("Accuracy:", accuracy_score(y_test,
y_pred_rf))
print("\nClassification Report:\n",
classification_report(y_test, y_pred_rf))
print("\nConfusion Matrix:\n",
confusion_matrix(y_test, y_pred_rf))
```

The code snippet above creates a Random Forest Classifier, an ensemble learning method that builds on generating multiple decision trees. An instance of the model is created with 100 estimators (the decision trees) and a state (for reproducibility) of 42. As seen previously with the code for logistic regression, fit () is used to fit the model to some training X and y data and then predict some test X data. The evaluation uses the same metrics as above: accuracy, classification report, and confusion matrix (Table 3).

Output

Table 4: *Exhibits the Random Forest Classification Report.*

Classification Report:

	precision	recall	f1-score	support
0	0.99	1.00	1.00	4031
1	1.00	0.99	1.00	2519
2	1.00	0.99	1.00	1466
accuracy			1.00	8016
macro avg	1.00	1.00	1.00	8016
weighted avg	1.00	1.00	1.00	8016

The performance of the Random Forest Classifier achieved an outstanding accuracy of 99.6%. It can also be observed that almost perfect classification among the classes is realized, 0, 1, and 2, with precision, recall, and F1-scores being exactly 1.00. Model performance for class 0 results in 4,031 samples being correctly classified with 0.99 precision and 1.00 recall, while classes 1 and 2, by convention, have 2,519 and 1,466 samples correspondingly and result in perfect precision of 1.00 and almost perfect recalls of 0.99 each. Both the macro and weighted averages are also 1.00 across all metrics, which further indicates balanced and superior performance across class imbalances. This represents a dramatic improvement from the Logistic Regression results and indicates that the Random Forest Classifier is much better suited for this particular classification task (Table 4).

XG-Boost

Table 5: *Portrays the XG-Boost Classifier Modelling.*

```
# XGBoost Classifier
xgb_clf = XGBClassifier(use_label_encoder=False,
eval_metric='logloss', random_state=42)
xgb_clf.fit(X_train, y_train)
y_pred_xgb = xgb_clf.predict(X_test)

# Evaluation
print("\nXGBoost Classifier Results:")
print("Accuracy:", accuracy_score(y_test, y_pred_xgb))
print("\nClassification Report:\n",
classification_report(y_test, y_pred_xgb))
print("\nConfusion Matrix:\n", confusion_matrix(y_test,
y_pred_xgb))
```

This code snippet above executes an XG-Boost Classifier, a powerful gradient-boosting model renowned for its performance and speed. One prepares the model with the following parameters: label encoder as false to handle the labels directly, eval_metric with 'log loss' to evaluate the model performance using logarithmic loss and random state equal to 42 to make the experiment reproducible. Similar to previous examples, it follows the same pattern: fitting the model on the training data (X_train, y_train), making predictions on the test data (X_test), and keeping consistency in the evaluation section by outputting the accuracy score, classification report, and confusion matrix as standard performance assessment means for the model (Table 5).

Output

Table 1: *Showcases the XG-Boost Classification Report.*

Classification Report:

	precision	recall	f1-score	support
0	0.77	0.97	0.86	4031
1	0.91	0.73	0.81	2519
2	0.96	0.58	0.72	1466
accuracy			0.82	8016
macro avg	0.88	0.76	0.80	8016
weighted avg	0.85	0.82	0.82	8016

The above table presents the results of the XG-Boost Classifier model. The model has correctly predicted 82.39% of all instances within this dataset. The classification report includes detailed information on performances for each class. Class 0 has high recall-97%-with 77% precision, which assumes good performance in identifying true positives. Class 1 has a rather balanced precision of 91% and recall of 73%, showing that for this class, there is a good trade-off between true positives identified and false positives raised. Class 2 has a lower recall of 58% and precision of 96%, which can be indicative of problems correctly identifying the instances of this class. Overall, the model performs well in terms of accuracy and precision. Nevertheless, concerning class 2, there is room for further improvement in its recall (Table 6).

Feature importance and correlation analysis

Comprehending the key drivers beneath water quality and sewage system efficiency is crucial for developing an efficient predictive algorithm. It is against this background that the use of feature importance scores considers models such as Random Forest and Gradient Boosting that are inherently useful in providing insights on which variables most drive predictions by calculating the importance of each feature in determining the model output. The most influencing features of the given study are Nitrogen and Phosphorus concentration in mg/L, Geographical Location, and Sampling Date. For example, in the Random Forest model, the highest ranking in importance was given to the nutrient levels, making changes in the non-turbidity parameters be strong predictor of water quality deterioration linked to sewage system inefficiency. The same conclusion is confirmed by the Gradient Boosting model since it highlights nutrient pollution. Such insights are highly useful in interventions to be given at appropriate targets, as such insights on the part of environmental agencies can prioritize monitoring and managing based on the factors that have a greater impact. Apart from feature importance, we also analyzed the correlation to understand how sewage system efficiency might relate to the different water quality parameters. Nutrient-level variables, such as Nitrogen and Phosphorus, showed a positive correlation with poor sewage systems in the correlation heatmap; thus, inefficient sewage systems lead to higher concentrations of such pollutants. Geographical coordinates along with temporal features like Year, Month, and Day, though having low correlation coefficients, did their job in capturing seasonal or locational variation in water quality. This analysis shows the diverse facets of water pollution, both of anthropogenic and natural nature that interact.

Economic impact assessment

The economic effects of poor water quality and unmanaged sewage systems run very deep, impacting many aspects of life: from public health and agriculture to tourism and general community well-being. Poor sewage management that leads to pollution of water bodies increases the rates of waterborne disease, causing health care costs to leap. Such communities are bound to experience the spread of diseases as a result of untreated or poorly treated water, which exposes people to cholera and gastroenteritis. This increases the cost of medication, hence resulting in the loss of productive hours because of sickness. Furthermore, the poor quality of water significantly impacts agricultural activities through irrigation water contamination, reducing crop yields, and increasing farming costs related to water treatment. This leads to financial loss for the farmers and raises prices for the consumers, thus having an impact on the entire value chain of food. Indeed, numerous studies done across the United States testify to the huge economic impacts of failing water and sewage systems. For example, there was the Flint, Michigan, water crisis, wherein quite poor treatment processes led to a leakage of lead into the city's drinking water supply. This not only poisoned scores of residents, with the worst effects felt by children but brought in a piece of long-term economic devastation. Lawsuits against the city, sharp declines in property values, millions of dollars in damages, and healthcare costs: were some of the costly results. Apart from the loss of civic trust, there was massive investment to be made in rebuilding the water infrastructure and restructuring the community's faith in public services.

Another example is the Mississippi River Basin, which has been polluted with nutrients due to inefficient sewage systems and runoff from fertilized agricultural fields. High levels of nitrogen and phosphorus have stimulated the growth of a large "dead zone" in the Gulf of Mexico where aquatic life cannot survive because of a lack of oxygen and where fishing and tourism industries are seriously affected. Thus, economic damage to the said commercial fisheries' activity in this region has been estimated in hundreds of millions of dollars annually since hypoxic conditions and oxygen levels make it hard for marine life to live. This reduction in fish stock affects local fishers and impacts the overall economy dependent on the supply chain of seafood. In Florida, the incidences of harmful algal blooms have continued to torture the state, with increasing agricultural runoff and sewage treatment further delving into exacerbating the problem. These have economic consequences, as tourism-based economies are especially affected when beach closures and health advisories are issued, leading to losses in hotel bookings, recreational activities, and local businesses. According to one estimate, the 2018 red tide in Florida cost the state approximately $130 million in lost tourism. Examples like these are the underpinning reasons why investment is critically needed in modern sewage systems, along with the management system of water quality that will reduce these economic impacts. The investment in infrastructure not only will protect public health and the environment but also will give long-term economic benefits by reducing these basic economic burdens from damages related to pollution. The novelty of such a dual focus lies in the combination of environmental and economic outcomes concerning the importance of efficient sewage systems for sustainable development.

Discussion

Implications for water quality management

The findings of this study have great implications for water quality management, especially concerning how predictive models could be leveraged further to advance monitoring and intervention strategies. This provides the possibility to combine machine learning algorithms in water quality management agencies that go beyond regular reactive approaches to proactive data-driven strategies. Predictive models project potential water quality problems based on history and thus allow timely interventions to prevent contamination events and optimize sewage network operations. Such models have the potential to automatically identify sources of pollution, predict environmental changes that affect water quality, and perform optimal resource allocation to monitoring efforts. For instance, this is possible in embedding machine learning models at established environmental monitoring systems where the detecting accuracy of such pollutants as nitrogen and phosphorus levels shall enable policymakers to establish more stringent regulatory measures. It is recommended that user-friendly interfaces should be developed for environmental agencies so that they can flawlessly embed predictive analytics into their day-to-day operations.

Challenges and limitations

Notwithstanding, several limitations and challenges should be addressed to maximize the benefits of these models. One such critical issue is the dealing of environmental data, especially sensitive information having a bearing on water sources that communities may depend on. Data privacy and conformity to regulatory requirements are very much in order. Similarly, model performance is heavily influenced by data quality and quantity. Poor practices in the collection of data, such as inconsistent frequency in data, missing values, or limits to real-time data access, can decrease the accuracy of the models leading to unreliable predictions. Another challenge is interpretability for such complex models as Gradient Boosting and Random Forest, because some predictions cannot intuitively be understood by stakeholders and, hence, may stand in the way of decision-making. Besides, generalization raises another limitation across different regions with different environmental conditions. A model that performs well in one geographical area might not perform well in another, first, because of the different water quality parameters of each place, and second, mainly because of the different pollution sources of each area.

Future research directions

Forging ahead, future research directions can concentrate on resolving these limitations and challenges by expanding the diversity of datasets used for model training. The diversities of data from various regions and climatic conditions could make the models robust and generalizable. There is also the possibility to examine the development of real-time water quality monitoring with IoT devices and satellite imagery for streams to make more accurate and dynamic predictions. Research into hybrid models can also be explored, which allows a combination of the key features of various machine learning methods that may prove particularly effective in achieving greater predictive accuracy. The future looks brighter as evolving technology will introduce more advanced and large-scale machine learning applications to improve water quality management, enhancing the outcomes for public health and environmental sustainability.

Conclusion

This study aimed at resolving the pressing matters associated with water quality and sewage system efficiency in the USA through a

multi-faceted approach. The research project strived to ascertain the relationship between sewage system efficiency and overall water quality in the USA. Besides, the present study endeavored to utilize machine learning techniques to develop forecasts of future trends in water quality. The datasets were gathered from as many reliable governmental databases as possible and environmental monitoring agencies to ensure robust and correct analysis. Among other sources included the national water quality databases include USGS, EPA, and EEA. These sources provided comprehensive data on a wide range of water quality parameters, such as pH levels, dissolved oxygen (DO), biological oxygen demand (BOD), chemical oxygen demand (COD), turbidity, nitrate and phosphate concentrations, and the presence of heavy metals like lead, mercury, and cadmium. In this research project, three evidence-based algorithms were selected, notably, Linear Regression, Random Forest, and XG-Boost are three algorithms of machine learning that have been selected for performing predictive modelling. Several performance metrics of the classes were performed for the stringent assessment of the performance of Recall, Accuracy, Precision, and F1 Score machine learning models. The performance of the Random Forest Classifier achieved an outstanding accuracy as compared to other models. The findings of this study have great implications for water quality management, especially concerning how predictive models could be leveraged further to advance monitoring and intervention strategies. This provides the possibility to combine machine learning algorithms in water quality management agencies that go beyond regular reactive approaches to proactive data-driven strategies.

References

1. Singh S, Das A, Sharma P. Predictive modeling of water quality index (WQI) classes in Indian rivers: Insights from the application of multiple machine learning (ML) models on a decennial dataset. Stochastic Envir Res Risk Assessment. 2024; 1-18.
2. Talukdar S, Ahmed S, Naikoo MW, Rahman A, Mallik S, Ningthoujam S, et al. Predicting lake water quality index with sensitivity-uncertainty analysis using deep learning algorithms. J Cleaner Production. 2023; 406: 136885.
3. Akhlaq M, Ellahi A, Niaz R, Khan M, Sammen SS, Scholz M. Comparative analysis of machine learning algorithms for water quality prediction. Tellus A: Dynamic Meteorol Oceanography. 2024; 76.
4. Ejaz U, Khan SM, Jehangir S, Ahmad Z, Abdullah A, Iqbal M, et al. Monitoring the industrial waste polluted stream-Integrated analytics and machine learning for water quality index assessment. J Cleaner Production. 2024; 450: 141877.
5. Ahmed AN, Othman FB, Afan HA, Ibrahim RK, Fai CM, Hossain MS, et al. Machine learning methods for better water quality prediction. J Hydrol. 2019; 578: 124084.
6. Ameer S, Shah MA, Khan A, Song H, Maple C, Islam SU, et al. Comparative analysis of machine learning techniques for predicting air quality in smart cities. IEEE access. 2019; 7: 128325-128338.
7. Asadollah SBHS, Sharafati A, Motta D, Yaseen ZM. River water quality index prediction and uncertainty analysis: A comparative study of machine learning models. J Env Chemical Eng. 2021; 9: 104599.
8. Miller T, Durlik I, Adrianna K, Kisiel A, Cembrowska-Lech D, Spychalski I, et al. Predictive modeling of urban lake water quality using machine learning: a 20-year study. Applied Sci. 2023; 13: 11217.
9. Omeka ME. Evaluation and prediction of irrigation water quality of an agricultural district, SE Nigeria: an integrated heuristic GIS-based and machine learning approach. Env Sci Pollu Res. 2024; 31: 54178-54203.
10. Gorenekli K, Gulbag A. Comparative analysis of machine learning techniques for water consumption prediction: a case study from kocaeli province. Sensors. 2024; 24: 5846.
11. Mukonza SS, Chiang JL. Meta-analysis of satellite observations for United Nations sustainable development goals: exploring the potential of machine learning for water quality monitoring. Envi. 2023; 10: 170.
12. Van Nguyen L, Bui DT, Seidu R. Comparison of machine learning techniques for condition assessment of sewer network. IEEE Access. 2022; 10: 124238-124258.
13. Zhu M, Wang J, Yang X, Zhang Y, Zhang L, Ren H, et al. A review of the application of machine learning in water quality evaluation. Eco-Envi Health. 2022; 1: 107-116.
14. Al Mukaddim A, Nasiruddin M, Hider MA. Blockchain technology for secure and transparent supply chain management: a pathway to enhanced trust and efficiency. Inter J Adv Eng Technol and Innovations. 2023; 1: 419-446.
15. Al Mukaddim A, Mohaimin MR, Hider MA, Karmakar M, Nasiruddin M, Alam S, et al. Improving rainfall prediction accuracy in the USA using advanced machine learning techniques. J Env Agricul Stu. 2024; 5: 23-34.
16. Alqahtani A, Shah MI, Aldrees A, Javed MF. Comparative assessment of individual and ensemble machine learning models for efficient analysis of river water quality. Sustainability. 2022; 14: 1183.
17. Buiya MR, Laskar AN, Islam MR, Sawalmeh SKS, Roy MSRC, Roy RERS, et al. Detecting IoT cyberattacks: advanced machine learning models for enhanced security in network traffic. J Computer Sci Technol Stu. 2024; 6: 142-152.
18. Debnath P, Karmakar M, Sumon MFI. AI in public policy: enhancing decision-making and policy formulation in the US government. Inter J Adv Eng Technol Innovations. 2024; 2: 169-193.
19. Debnath P, Karmakar M, Khan MT, Khan MA, Al Sayeed A, Rahman A, et al. Seismic activity analysis in California: patterns, trends, and predictive modeling. J Computer Sci Technol Stu. 2024; 6: 50-60.
20. Hasan MR, Islam MZ, Sumon MFI, Osiujjaman M, Debnath P, Pant L. Integrating artificial intelligence and predictive analytics in supply

chain management to minimize carbon footprint and enhance business growth in the USA. J Bus Manag Stu. 2024; 6: 195-212.

21. Islam MR, Nasiruddin M, Karmakar M, Akter R, Khan MT, Sayeed AA, Amin A. Leveraging advanced machine learning algorithms for enhanced cyberattack detection on US business networks. J Bus Manag Stu. 2024; 6: 213-224.
22. Islam MR, Shawon RER, Sumsuzoha M. Personalized marketing strategies in the US retail industry: leveraging machine learning for better customer engagement. Inter J Mach Lear Res Cybersecurity Artific Intelligence. 2023; 14: 750-774.
23. Karmakar M, Debnath P, Khan MA. AI-powered solutions for traffic management in US cities: reducing congestion and emissions. Inter J Adv Eng Technol Innovations. 2024; 2: 194-222.
24. Khan MA, Debnath P, Al Sayeed A, Sumon MFI, Rahman A, Khan MT, et al. Explainable AI and machine learning model for California house price predictions: intelligent model for homebuyers and policymakers. J Bus Manag Stu. 2024; 6: 73-84.
25. Kouadri S, Elbeltagi A, Islam ARMT, Kateb S. Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast). Appl Water Sci. 2021; 11: 190.
26. Nasiruddin M, Al Mukaddim A, Hider MA. Optimizing renewable energy systems using artificial intelligence: enhancing efficiency and sustainability. Inter J Mach Lear Res Cybersec Artif Intelligence. 2023; 14: 846-881.
27. Shil SK, Chowdhury MSR, Tannier NR, Tarafder MTR, Akter R, Gurung N, et al. Forecasting electric vehicle adoption in the USA using machine learning models. J Computer Sci Technol Stu. 2024; 6: 61-74.
28. Shawon RER, Rahman A, Islam MR, Debnath P, Sumon MFI, Khan MA, et al. AI-driven predictive modeling of us economic trends: insights and innovations. J Hum Soc Scien Stu. 2024; 6: 01-15.
29. Sumon MFI, Osiujjaman M, Khan MA, Rahman A, Uddin MK, Pant L, et al. Environmental and socio-economic impact assessment of renewable energy using machine learning models. J Eco, Fina Account Stu. 2024; 6: 112-122.
30. Zeeshan MAF, Sumsuzoha M, Chowdhury FR, Buiya MR, Mohaimin MR, Pant L, et al. Artificial intelligence in socioeconomic research: identifying key drivers of unemployment inequality in the US. J Econ Fin Acco Stu. 2024; 6: 54-65.