Water Quality Using Machine Learning

Manar Alsayed, Abrar Sebiany, Joury Alzayat, Wahbia Saleh, Haya Aldossary

Department of Artificial Intelligence, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia

2200001689@iau.edu.sa, 2200002067@iau.edu.sa, 2200002640@iau.edu.sa, 2200006267@iau.edu.sa, 2190002968@iau.edu.sa

Acknowledgement

We are blessed to be able to demonstrate our studies in the Machine Learning course by implementing this project. We are also thankful to Imam Abdulrahman bin Faisal University and the Artificial Intelligence department for this opportunity to showcase our skills. And most importantly, special thanks to our instructor and supervisor Dr. Nawaf Alharbi for aiding us throughout this process and supporting our efforts constantly. Without his input, feedback, and expertise, this project would not have been possible to complete in this trimester.

Abstract

Precise detection of water quality is crucial in sustaining and preserving the safety of the environment, wildlife, and humans. However, most tools and approaches used in water quality detection tend to be costly. Therefore, developing economic, trustworthy, and efficient tools that aid in classifying water potability is essential. Moreover, based on the literature survey outcomes, it was found that recent research papers have heavily relied on using more than two classes for the water quality classification. Nevertheless, this study aims to classify water quality by using binary classification, specifically, it will use only two classes to categorize the water as potable or non-potable focusing on the pH level attribute as well as deploying other attributes. This model will be developed by selecting the machine learning algorithm that achieves the highest accuracy and the best overall result. Furthermore, Python programming language will also be used to implement reliability and thoroughness. The proposed model shall have the ability to classify water quality using the pH level attribute, hence enabling us to make definitive decisions that impact the safety and health of generations to come.

Keywords: water quality, water potability, binary classification, pollutants, chemicals, dissolved oxygen, pH level, solids, sulfates, supervised machine learning algorithm.

Table of Contents

Table of Figures

Table of Tables

1 Introduction

In aquatic systems, water quality is an important indicator of overall health. Numerous adverse effects can result from poor water quality, both on human health and the surrounding environment. Water pollution and certain activities, such as agricultural run-off, can alter the chemistry of the water and cause irregular levels of key contaminants. These irregular levels can lead to a decrease in water quality, with consequences including algal blooms, habitat destruction, a decrease in fish populations, and increases in toxic substances that are dangerous for drinking or swimming [1]. Therefore, we decided to implement a water quality severity indicator to know the regular levels of contaminants in our waterways that are essential for identifying potential risks to health and ecology and maintaining proper water quality standards. In previous studies, many machine learning methods, such as Random Forest Models (RF), Artificial Neural Networks (ANN), and Support Vector Machines (SVM). However, we found out that all earlier high-accuracy studies focused on algorithms that might not be accurate in all locations and environmental circumstances, and they cannot be generalized to be used in other regions or hydrogeological datasets. Also, they used more than two classes to classify the water quality. In this project, we will expand our dataset to be more useful in other regions. We will also use binary classification which means that we will have two classes identifying if the water can be consumed or not. We will consider all attributes while focusing more on the pH of water since it plays an important role and affects many chemical and physical processes in water. Consequently, it will be necessary to do more work to facilitate the prediction of water quality severity and to improve its accuracy.

As an example of data pre-processing techniques applied, there are several stages, such as data cleaning, which involves removing missing or inconsistent data, detecting outliers, and dealing with imbalanced datasets. We used a combination of random over-sampling techniques to avoid biased results resulting from an imbalanced dataset. So, after implementing preprocessing techniques and machine learning algorithms, empirical results have revealed that Gradient Boosting Machine other classifiers with 75.6% accuracy. 66% of precision and 77% of recall with 9 features only.

The paper is organized as follows: section 2 contains the literature review and its outcomes, and section 3 contains a description of the proposed models which includes Random Forest, Gradient Boosting Machine, and Support Vector Machine. Section 4 contains the empirical studies which include the dataset description, experimental setup, criteria for performance evaluation, tools, and optimization strategy. Moreover, section 5 contains the results and discussion and lastly, section 6 contains the conclusion and the suggested future work that can be implemented to improve this work further.

1.1 Problem Statement

Water quality has become a cause for concern in recent years due to increasing pollutants, chemicals, and other substances entering our waterways from sources including agricultural runoff, industry disposal, and sewage overflows. This has resulted in reduced water quality, leading to health problems such as skin irritation, gastrointestinal illness, and even cancer in some cases [2]. In addition, these contaminants have an adverse effect on aquatic ecosystems, with declines in fish populations and implications for wildlife that rely on the water source. Furthermore, due to the lack of accurate testing methods, it is difficult to detect these contaminants. As a result, there is an urgent need for solutions to mitigate the resulting public health risks posed by poor water quality and ensure access to clean, safe drinking water. Also, the need for classifying cases based on severity levels improves accuracy in predicting water quality severity. Our aim is to create a machine-learning algorithm that can offer the environmental industry this kind of service. Thus, contribute to society by assisting in the resolving of water quality-related concerns.

1.2 Motivation

This project is motivated by the following reasons:

- The need for increased recreational opportunities for fishing, swimming, and other aquatic activities.
- Reduce the socioeconomic impact of water quality pollution by participating in the effort.
- The need to increase public health and safety by more accurately monitoring the quality of drinking water.
- The need to greatly reduce costs associated with collecting and analyzing water data samples.
- The need to create faster responses to environmental disasters or contamination events.

1.3 Objectives

The Study objectives for this work can be stated as follows:

- Create an effective and automated machine learning model to reduce errors and improve success rates.
- Prepare a clean dataset by using some preprocessing techniques.
- Determine the potability of the water using different machine-learning algorithms.
- Increase the accuracy in the prediction of the water quality model.
- Improve the results by using feature selection techniques to minimize the complexity of the model and make it easier to interpret.
- Contribute to society by helping resolve an issue related to water quality.

1.4 Limitation

The dataset did not mention DO (Dissolved Oxygen) and BOD (Biochemical Oxygen Demand) tests that may play a critical role in determining the water quality. In further work, we may focus on these two tests with the pH level of the water for more accuracy.

2 Literature Review

Needless to say, water is essential to life, so taking measures to ensure optimal water quality can have far-reaching impacts on keeping people safe from illness due to contaminants that may exist in the water supply. Many studies have been carried out to help us understand how to measure and analyze different aspects of water quality. In this section, we examine researchers' efforts in recent years and discuss their findings and possible future work.

In a study by Muharemi *et al*. [3], water supply companies have adapted warning systems that detect contamination using sensors that can monitor the water's quality and send notifications whenever the water quality changes. The study is being conducted to provide better results with the highly imbalanced dataset. Moreover, the dataset was extracted from a German public water company and is a time series data set containing 122,334 entries. A set of models was trained using the following classification algorithms: Logistic Regression (LR), Linear Discriminant Analysis (LDA), Artificial Neural Networks (ANNs), Support Vector Machines (SVM), Deep Neural Networks (DNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM). As for the feature selection, embedded methods using classifiers have been employed. These classifiers select features based on the highest measured information gain. Specifically, the method can measure the importance of the features by using Random Forest (RF) classifiers. In addition, since the dataset is time series data, a time series cross-validator has been used to evaluate the model. It is different from the traditional k-fold cross-validation method, as the former returns successive sets as supersets of the previous ones. After running experiments on training models by implementing each algorithm, it was found that the Support Vector Machines algorithm was the best-performing one with 98.8% accuracy.

In a study made by Ahmed *et al*. [4], supervised learning algorithms to estimate water quality based on a minimum number of parameters with economical sensors to predict water quality, because water quality parameter sensors are exorbitant. A total of eight regression algorithms and ten classification algorithms were used in the study to predict the Water Quality Index (WQI) and Water Quality Class (WQC). In more detail, Multiple Linear Regression, Polynomial Regression, Gradient Boosting, Ridge Regression, and Gaussian Naïve Bayes are some of the algorithms employed in the study. As part of the dataset collected from the Pakistan Council of Research in Water Resources (PCRWR), 663 samples were collected from 13 different sources of Rawal Water Lake between 2009 and 2012. Based on the experimental results, indicates that Gradient Boosting was the most efficient regression algorithm with a mean Absolute Error

(MAE) of 1.9642, Mean Squared Error (MSE) of 7.2011, Root Mean Square Error (RMSE) of 2.6835, and Residual Standard Error (RSE) of 0.7485. Moreover, in classification algorithms, Multi-layer Perceptron (MLP) was more accurate than the other algorithms, with an accuracy of 85% and a precision of 56%.

In a paper prepared by Neha R *et al*. [5], multiple machine learning models were used to analyze a method for water quality detection in the domain of water resources in India, such as lakes, ponds, and rivers. In conducted research, two datasets were used that had been collected from different river stations and had a total of eight water quality parameters such as pH, Dissolved Oxygen (DO), and Electrical Conductivity (EC). As for the models used, two supervised learning algorithms are SVC and Naive Bayas, and also the third algorithm was the Decision Tree. As stated in their research objectives, the goal was to analyze the performance of the three models, to decide on the most efficient model; Balanced Accuracy Score and Confusion Matrix were used. The result obtained shows that the Decision Tree model was the most efficient, as it scored 98.5% accuracy.

In a paper prepared by Gupta *et al*. [6], a machine learning model to predict the Water Quality Index (WQI) for the groundwater sample was proposed. This model aims to evaluate the water quality in the Mid Gangetic Region (South Bihar plain) of India using a machine learning algorithm and a Water Quality Index. For predicting WQI, the study adopts the Elaboration Likelihood Model (ELM) for WQI modeling. Furthermore, the model is compared to three other hybrid models (BBO-ELM, RBF-ELM, and OS-ELM). A collection of historical river water quality data was used to develop and test the models. There are two parts to the dataset in this model. Firstly, the training dataset consists of a randomly chosen collection of data used to create the model. Secondly, the testing dataset consists of a randomly selected group of data utilized to test the accuracy of the created model. Based on the experimental results, it indicates that BBO-ELM is an effective approach in judging water quality and can be effectively used as a substitution for the assessment. Moreover, this model yielded the highest R^2 value (0.955) among others during validation, indicating its effectiveness as well. Therefore, it may be concluded that this is the most appropriate hybrid model for evaluating the WQI.

In a paper prepared by Kouadri *et al*. [7], A suggested machine learning model to forecast the Water Quality Index (WQI) for enhancing groundwater resource management strategies. The model aims to simplify WQI computations and illustrate water quality variation in critical situations where the necessary analyses are lacking. A total of eight different models were used in the study to predict the water quality parameter (WQI). Therefore, they used a dataset provided by the Directorate of Water Resources (DRE) of the State of Illizi were relied on with 114 samples taken from 57 exploited wells of 6 different layers Eight artificial intelligence algorithms were used to calculate the weight of the study: random forest (RF), M5P tree, multilinear regression (MLR), random subspace (RSS), locally weighted linear regression (LWLR), additive regression (AR), artificial neural network (ANN), and support vector regression (SVR). According to a

comparison of performance assessment metrics, the MLR model has better accuracy compared to other models, and the RF model has a lower error rate. As a result, they found that the RF algorithms could be a reliable and affordable model to improve groundwater quality control strategies in a dry area of southeast Algeria. Therefore, these algorithms might not be accurate in all locations and environmental circumstances, and they cannot be generalized to be used in other regions or hydrogeological datasets [5].

In a study conducted by Malek *et al*. [8], they used the Water Quality Index (WQI) which is based on a number of parameters to classify the quality of water into three classes: Clean, Slightly polluted, and Polluted to detect water quality, as water pollution is considered a serious problem in Malaysia. In the study, samples were collected along 8 areas of the Kelantan River in Malaysia. Seven classification algorithms were used with 658 observations from the dataset; these models used 13 parameters as input. As a result, it was found that the Gradient Boosting (GB) model scored a 94.9% accuracy rate, which is the highest among other models, while the Decision Tree (DT) scored 86.22%, which is the lowest accuracy of the seven models used.

A paper prepared by Lee H. *et al*. [9], depicts the Republic of Korea's control of the water quality of its rivers with successive impoundments. It aims to create a framework based on machine learning to rank total phosphorus (TP) management methods based on machine learning (ML). Therefore, they used datasets from the serial impoundment made by three dams, a sewage treatment plant, a tributary, and a meteorological station that consisted of 40 input variables and one output variable with 546 raw data records. There are four ML techniques' effectiveness assessed through comparison analysis: extreme gradient boosting (XGBoost), random forest (RF), long short-term memory (LSTM), and deep neural network (DNN). Additionally, the LSTM-based model is selected as the best forecasting model for Euiam Lake. As a result, they suggest that a reliable predictive model can be created for a study site with distinct seasonal patterns of temperature and rainfall as well as significant seasonal variations in water quality, making it easier to choose management priorities. Therefore, by quantitatively determining the reduction level of the chosen priority, it is anticipated that the proposed method could be used as a practical, cost-effective tool for water quality management.

In a paper prepared by Derdour *et al*. [10], protecting the environment and sustaining water quality for human consumption is essential, hence the development of non-traditional economic techniques. Specifically, the study aims to prove that machine learning methods can predict water quality. By using a water sample dataset from the Wilaya of Naama in Algeria, 151 samples were used for examination, and 18 formed the basis of testing and confirmation for the prediction model. Multiple algorithms were used to classify these samples, specifically, Decision Tree, Ensemble Tree, K Nearest Neighbors, Discrimination Analysis Classifier, and Support Vector Machine (SVM). Moreover, the k-fold cross-validation method was used to evaluate the prediction model's robustness. Furthermore, the best result of the model was obtained by using Linear Support Vector

Machine on raw data with 94.7%. However, the same algorithm obtained even higher accuracy with the standardized training data, at 95.4%. Finally, the study was limited by the low sample size, as only 151 samples were used for training.

Water quality is closely related to environmental, economic, and health concerns. Therefore, it is a necessity to monitor water quality and understand the factors that impact it. This study, written by Suwadi et al. [11], shows improvement in the prediction of water quality by reducing the number of features involved in the test and employing three different algorithms (information gain, correlation, and symmetrical uncertainty to quantify the weight value). Therefore, they used datasets from the Langat Basin in Selangor with 29 attributes and 907 samples. The outliers were detected in WEKA using the unsupervised attribute interquartile range before extreme values and outliers were removed. To quantify the weight value in the study, three algorithms from the filtering approach category were applied: information gain, correlation, and symmetrical uncertainty. This DOE-WQI dataset contains DO, NH3-NL, BOD, COD, SS, and pH. With the available dataset and selected features, the classification problem of predicting water quality could be solved with DO, NH3-NL, BOD, COD, SS, and pH. Furthermore, they found that DO (Dissolved Oxygen) and BOD (Biochemical Oxygen Demand) both play a critical role in predicting and determining the quality of water.

A study written by Hoque et al. [12], The purpose of the study is to identify the most effective algorithm for measuring WQI (water quality index) and clean drinking water. Eight stand-alone regression learning algorithms were used in this study (DT, LR, Ridge, Lasso, SVR, RF, ET, and ANN). The data that was used in the study was collected in India, it has six water quality parameters, and two sets of derivative features from the original features are used as inputs for the algorithms to learn the patterns and predict the WQI. The findings suggest that LR and Ridge are the most effective regression algorithms for water quality prediction systems. In addition, the water quality rating scale is the most appropriate input for the model. The results show that LR and Ridge were trained using the water quality rating scale.

Since water is a crucial source of living it is very important to know the quality of it before consumption. In a research paper prepared by Kaddoura, S [13]. a number of machine learning (ML) models were used to evaluate the water quality based on certain parameters as water represents more than two-thirds of the earth’s surface, it can be hard to check WQ manually. Water quality (WQ) parameters were categorized into 3 categories: Physical, Chemical and Biological. These parameters were used to predict the water quality and how safe it is for consumption. In the research paper, a dataset consisting of 3276 records was collected, preprocessed and split it into training data and testing (80:20) data to be used in categorizing the water into four main categories: potable water, palatable water, contaminated water (polluted) and infected water. The researcher found out that SVM algorithms scored the highest accuracy with 73% followed by ANN with an accuracy value of 72%.

Although there are a variety of tests available, the aim of our study is to improve the efficiency of predicting the quality of water by reducing the features and focusing more on measurements of the water's pH level. Maintaining optimal pH levels is essential for maintaining water quality and ecosystem health [14]. A pH test helps protect aquatic plants and animals and keeps drinking water and groundwater safe.

2.1 Literature Review Summary Table

Ref	Dataset		Approach	Preprocessing	Methods	Accuracy	Feature Extraction	Notes
	Domain	Size						
[3]	Naama, Algeria.	169 samples.	Machine Learning	Not Mentioned	- Decision Tree - Ensemble Tree -K-Nearest Neighbors - Discrimination Analysis Classifier - Support Vector Machine	95.4%	- pH. -Electrical conductivity. - Mineralization . - Magnesium.	-
[4]	Rawl Watershed, Pakistan.	663 samples.	Machine Learning	- Cleaning -Normalization	- Regression - Classification	85%	- pH. - Turbidity. - Temperature. -Total dissolved solids.	The study intends to validate the viability of its usage in real-time water quality detection systems by achieving a decent accuracy with a limited number of parameters.
[5]	Kelantan River, Malaysia.	In 2005-2020	Machine Learning	-Outlier detection - Exploration - Z-Score -Normalization	-K-Nearest Neighbors - Support Vector Machine - Decision Tree	94.9%	-Oxygen concentrations in solutions. - Biochemical oxygen	-

					- Naive Bayes - Random Forest -Gradient Boosting		demand. - Ammoniacal nitrogen. - pH. -Total suspended solids. -Chemical oxygen.	
[6]	The South Bihar Plain, India	156 samples.	Machine Learning	Not Mentioned	Elaboration Likelihood	95%	- pH -Total dissolved solids. -Electrical conductivity. - Calcium. - Magnisum. -Total hardness. - Nitrate. - Sulfate.	-
[7]	State of Ilizi's Directorate of Water Resources.	114 samples.	Machine Learning	Not Mentioned	- Multiple Linear Regression - Random Forest - M5 Prime - Residual Sum of Squares - Autoregressive -Artificial Neural Networks - Support Vector Regression -Local Weighted Linear Regression	99%	-Total dissolved solids. - Temperature. - Calcium. - Magnesium - Nitrate. -Pollution indicators.	The study offers a backup plan in the event that crucial data is not available.
[8]	Narmada River, India.	In 2017-2018	Machine Learning	Not Mentioned	- Support Vector Machine - Decision Tree - Naive Bayes	87.1%	- pH -Electrical conductivity -Dissolved oxygen. -Biochemical oxygen demand.	-

[9]	Three dams, a sewage treatment facility, a tributary and a weather station.	546 raw data.	Machine Learning	Pre-treatment	- XGBoost - Random Forest - Long Short-Term Memory - Deep Neural	92%	-Chemical oxygen demand. -Suspended solids. -Total phosphorus. -Water temperature. - pH -Total nitrogen. Electrical conductivity. -Oxygen demand. - Outflaw. - Biochemical oxygen demand.	For a serial impoundment system with distinct seasonal patterns of temperature, precipitation, and water quality, a reliable predictive model was created to aid in management decisions.
[10]	Thuringer Fernwasserver-sorgun company, Germany.	122,334 samples.	Machine Learning	- Integration -Normalization - Cleaning - Imputation -Noise identification	-Logistic Regression -Linear Discriminant Analysis - Support Vector Machine -Neural Networks -Recurrent Neural Networks -Deep Neural Networks -Long-Short-Term Memory - Random Forest	0.04-0.08	- Temperature. -Chlorine Dioxide. -Redox potential. - Flow rate. - pH value. -Electrical conductivity. - Turbidity.	-
[11]	Langat Basin, Selangor.	907 samples	Machine Learning	- Cleaning -Normalization - Sampling	Optimized Forest Classifier	98.8%	-Electrical conductivity. - pH. -Dissolved oxygen. - Magnesium.	-
[12]	India.	1991 samples	Machine Learning	Not Mentioned	- Decision Tree - Regression	95.64%	- Temperature. -Dissolved	

					-Linear Regression - Ridge - Lasso - Random Forest - Support Vector Machine -Artificial Neural Networks -Extra Tree Regression		oxygen. - pH. - Conductivity. - Biochemical oxygen demand. - Nitratenan.	
[13]	Public Domain	3276. samples.	Machine Learning	-Removing Empty Inputs	-Random Forest -Gradient Boosted Trees -XBGoost -Logistic Regression	97%	-Physical -Chemical -Biological	Biological features such as bacteria were considered in water quality parameters

Table 1 Literature Review Summary

2.2 Literature Survey Outcome

After reading and reviewing research papers on water quality detection, we found some gaps and areas we can explore and fix. None of the previous studies had classified water potability into two classes only. Therefore, this study aims to deploy binary classification by using two classes only (Potable Water, Non-potable Water). Moreover, we also found that many attributes are being used for detection, however, our study aims to focus on the pH level feature. Moreover, the dataset used in this project can achieve higher quality by implementing different approaches as we wish to do.

3 Description of Proposed Techniques

Machine learning offers a variety of algorithms, each takes data and turns it into something useful, as well as making predictions that humans cannot make without the help of technology [15]. However, these algorithms' performance depends on multiple factors that need to be studied before making a choice for the right algorithm. Namely, this project aims to observe a dataset of different bodies of water and classify the water as potable or not. Furthermore, the choice of the algorithm also depends on the input's type, size, and whether it's clean or dirty. Finally, the linearity of the data is important, as non-linear complex data require more training. Given the steps

required to choose a machine learning algorithm, three algorithms have been picked to train the data, and they are briefly described below:

3.1 Random Forest

In 1974, Leo Breiman and Adele Cutler first introduced the Random Forest (RF) algorithm. Classification and regression problems can be solved using supervised machine learning algorithms [16]. Working with ensemble learning, it solves complex problems by combining many classifiers. Random Forest algorithms have many decision trees. Each decision tree is built using a random subset of features from the dataset, which makes it less likely to overfit. In a decision tree algorithm, a training dataset is divided into branches, which are further subdivided into other branches. This method produces a more accurate result by averaging all of the individual predictions [17].

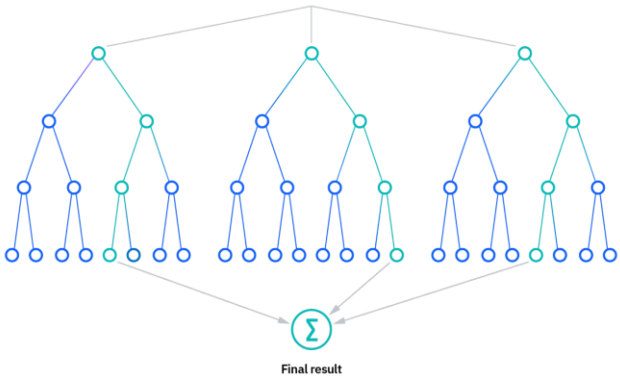


Figure 1 General Visualization of Random Forest in Use

Figure 2 shows an example of how the Random Forest algorithm works for a binary classification problem that classifies if the animal in the picture is a Dog or a Cat. It has N decision trees; each tree will produce an output based on the features it was exposed to. Finally, after counting how many outputs belong to each class, Random Forest predicts and chooses based on the most repeated prediction [18].

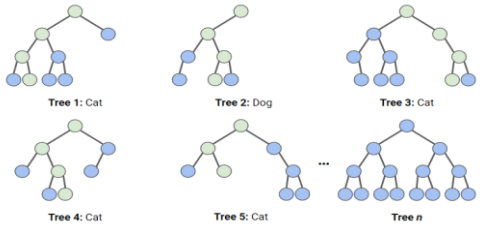


Figure 1 Binary Classification Using Random Forest [18]

3.2 Support Vector Machine

Support Vector Machine (SVM) was put forward in 1992 by computer scientist researcher Vladimir Vapnik and his co-workers. SVM is a supervised linear model that is applicable to both regression and classification problems. Since data can differ in many ways, SVM has evolved to handle non-linear data using the kernel [19]. In our model, we will use linear SVM, given that SVM works efficiently with numeric data and automatically creates higher dimensions of the features and summarizes this in the output [20]. In addition, linear SVM is a powerful method for applications requiring high-dimensional data, including text classification, air pollution modeling, and water quality monitoring.

Basically, by developing an optimal hyperplane, SVM is able to classify numerous binary classes with clarity [21]. As seen in Figure 2, support vector points are the locations where the support vectors are closest to the hyperplane, and margins are the spaces between the vectors and the hyperplane. By changing the vectors' positions, the hyperplane will also change, so SV points are crucial to finding the hyperplane [21]. Technically this hyperplane also known as a hyperplane that maximizes margin, and it can be written as:

$$x = b + \sum_{i \text{ is support vector}} \alpha_i y_i a(i) \bullet a$$

We can improve classification accuracy and lower the chance of overfitting by finding the hyperplane that maximizes the distance between data points of different classes.

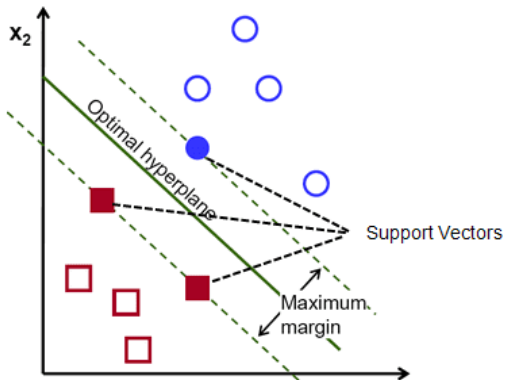


Figure 2 Support Vector Machine (SVM) Algorithm [22]

3.3 Gradient Boosting Machine

Gradient Boosting Machine (GBM) was put forth in 1999 by Jerome Friedman, a statistician at Stanford University. Using a variety of weak prediction models, commonly decision trees, GBM is a form of machine learning a method that combines them to produce a more precise model [23]. Also, it is used for regression and classification algorithms that only need 100 data, and it enhances the model's overall performance by identifying and fixing any mistakes made by inexperienced learners [24]. It consists of three basic components: a loss function, a weak learner for making predictions, and an additive model for combining weak learners. [25].

Creating a base model to classify the observations in the training dataset is the first stage of gradient boosting. To achieve that, we have to compute the loss function, which is expressed by the following formula:

$$L = -\sum_{i=1}^{n} y_i \log(p) + (1-p)\log(1-p)$$

Then we need to transform the loss function into a log (odds), which will be the initial leaf node. In order to get the full tree, we will calculate the pseudo-residual, which can be computed using this formula [26]:

$$Residual = Observed - Predicted$$

Following that, we will estimate the target category using every tree in the ensemble. Next, we will calculate the new residuals set and repeat this step until we reach a point where the number of iterations specified by the hyperparameter is matched. Finally, when trained, we will make the final prediction using all the trees in the ensemble [27].

4 Empirical Studies

4.1 Dataset

The synthetically generated dataset provides over 3276 water samples with 9 distinct attributes that help determine the water quality. To classify the water samples, it is either zero or one, where zero indicates non-potable water and 1 is potable water.

Potability	**Number of occurrences**
0	1998
1	1278

Table 2 Number of Occurrences for Each Case

Feature	Feature description
pH	Used to specify the acidity or basicity of water where water with a pH less than 7 is considered acidic and greater than and less than 14 is considered basic.
Hardness	Is caused by the number of magnesium and calcium dissolved in the water.
Solids (total dissolved solids)	The amount of organic and inorganic materials in the water. where the maximum limit is 1000 mg/l and the desired limit for these solids is 500 mg/l.
Chloramines	Chlorine and chloramines are disinfectants used in water. Where 4 mg/l are considered safe for consumption.
Sulfate	A naturally occurring mineral that is highly concentrated in the sewer water and much less in freshwater.
Conductivity	How much energy flows through the water, and as per standards electrical conductivity should not exceed 400 µS/cm.
Organic carbon	It is derived from decaying natural organic matter and synthetic sources. Drinking water with less than 2 mg/l is considered safe.
Trihalomethanes	Are found in water treated with chlorine. Where THMs up to 80 ppm is considered safe.
Turbidity	It depends on the number of solids in the water. Turbidity of 5.00 NTU is recommended.
Potability	Determines the water quality, where 0 indicates non-potable water and 1 indicates potable water.

Table 3 Feature Description

4.1.1 Statistical Analysis of Dataset

Table 4 demonstrates the statistical analysis of the dataset. The mean, the standard deviation, the minimum, the maximum and the median.

No.	Feature	MEAN	STD	MIN	MEDIAN	MAX
1	pH	7.080795	1.594320	0.000000	7.036752	14.000000
2	Hardness	196.369496	32.879761	47.432000	196.967627	323.124000
3	Solids	22014.092526	8768.570828	320.942611	20927.833607	61227.196008
4	Chloramines	7.122277	1.583085	0.352000	7.130299	13.127000
5	Sulfate	333.775777	41.416840	129.000000	333.073546	481.030642
6	Conductivity	426.205111	80.824064	181.483754	421.884968	753.342620
7	Organic carbon	14.284970	3.308162	2.200000	14.218338	28.300000
8	Trihalomethanes	66.396293	16.175008	0.738000	66.622485	124.000000
9	Turbudity	3.966786	0.780382	1.450000	3.955028	6.739000

Table 4 Statistical Analysis of Dataset

No.	Feature	Count
1	pH	2785
2	Hardness	3276
3	Solids	3276
4	Chloramines	3276
5	Sulfate	2495
6	Conductivity	3276
7	Organic carbon	3276
8	Trihalomethanes	3114
9	Turbidity	3276

Table 5 Count of Features

4.2 Experimental Setup

In this section, we will explain how we build WQ (Water Quality) classification models, using a publicly available dataset. The procedure begins with preprocessing. Start with separating potable water from non-potable water. Then fill the missing value of the Potable class with the median of the values and use the same technique for non-POTABLE water values. The reason we use the median over other techniques is that the median is less sensitive to extreme values/outliers than others. For example, as the median is not influenced by extreme values at its tail, it can better capture the typical or "central" value of the data if the distribution is skewed [28]. Also, the median gives us more accuracy than the mean and mod. Using the Pandas skew function is utilized to check correlation values, which are considered normal when they lie between 0.5 to -0.5. Skewing is performed based on skewness values, and Box-Cox is used to remedy non-normal distribution. Due to direct partitioning, the data becomes unbalanced, which is resolved through SMOT over-sampling. Normalization scales features within the same range. After acquiring a clean dataset cross-validation was used with 10k-fold. Next, the data undergoes training on three machine learning models: Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting Machine (GBM). The trained model is then compared to the trained data results. The estimator (model) was also fitted to the training data by iterating through specified hyperparameters with GridSearchCV. To reduce variables, the embedded method is used for feature selection. Performance measures such as accuracy, precision, recall, and most effective parameters are considered in selecting the model with the highest performance.

.

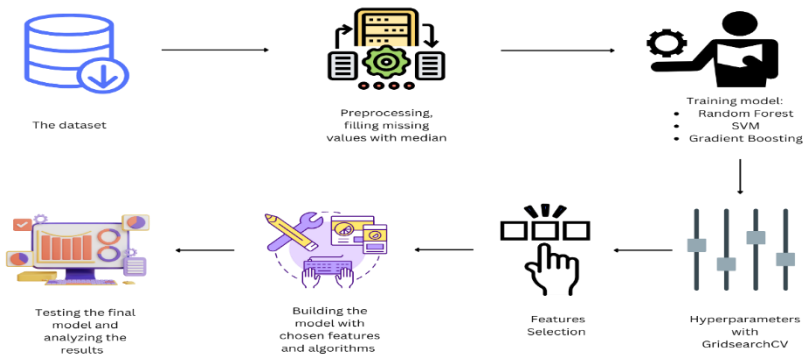


Figure 3 WQ Classification Model

4.3 Criteria for Performance Evaluation:

To evaluate classification performance, three methods were used, namely accuracy, precision, and recall. In terms of accuracy, we measure how many correct predictions were made compared to how many observations there were. Precision measures the ratio of the positive predictions that are correct to the number of positive predictions overall. Similarly, recall can also be called sensitivity or True Positive Rate, which measures how many positive predictions were correctly compared to how many positive observations were made. Accuracy, Precision, and Recall equations [29].

$$\text{Precision} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)+ False Positive (FP)}}$$

$$\text{Recall} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}}$$

$$\text{Accuracy} = \frac{\text{True Positive (TP)} + \text{True Negative(TN)}}{\text{True Positive (TP)} + \text{True Negative(TN)} + \text{False Positive (FP)} + \text{False Negative (FN)}}$$

4.4 Tools

As part of our first experiment, we will implement our proposed model through Random Forest machine learning algorithms. We will implement Python programming using the Jupyter Notebook environment with NumPy, Pandas, and MATLAB libraries. To ensure the validity of our development procedures, three laptops will be used: Lenovo yoga core i7 11th Gen, Hp core i7 11Gen, and Hp core i7 10th Gen (all 64-bit).

4.5 Optimization

It is necessary to modify the parameters of the learning algorithms in order to generate models that perform a classification problem optimally. For this purpose, grid search with cross-validation was used. Grid search defines all possible combinations of parameter values in order to explore all parameters simultaneously. Using a 10-fold cross-validation, we validated the models for each combination, and we calculated their average accuracy. Below are the tables listing the optimal parameters for each classifier based on the grid search.

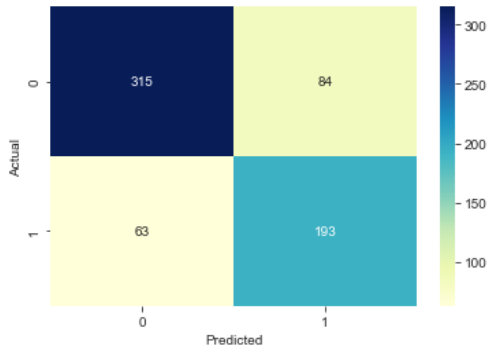
Random Forest	Optimal value chosen	Confusion Matrix
max_depth	10	
max_features	‘auto ‘	
min_samples_leaf	4	
n_estimators	5	

Table 6 Random Forest Optimal Parameters Values

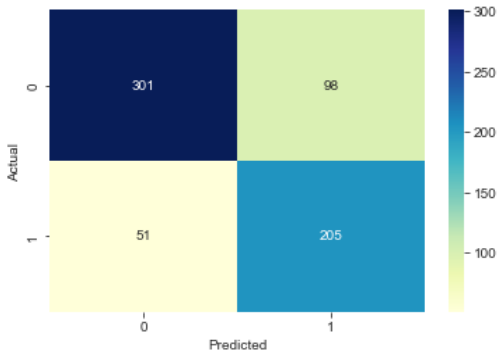
Gradient Boosting	Optimal value chosen	Confusion Matrix
max_depth	3	
Max_features	‘Auto ‘	
min_samples_leaf	2	
min_samples_split	5	
n_estimators	100	

Table 7 Gradient Boosting Optimal Parameters Values

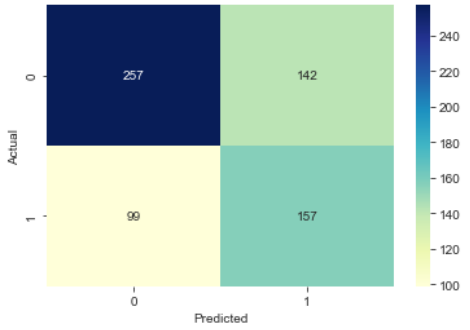
Support Vector Machine	Optimal value chosen	Confusion Matrix
C	1	
coef0	0.5	
degree	1	
gamma	0.1	
kernel	'rbf '	

Table 8 Support Vector Machine Optimal Parameters Values

5 Result and Discussion

An analysis and discussion of the results of the water quality classification are presented in this section. In this study, we evaluated the effects of feature selection and used 10-fold cross-validation to analyze the outcomes of training/testing and data partitioning. When all three classifiers were examined, it was found that GB consistently outperformed all other algorithms in each evaluation. As for RF, it was almost as accurate as GB. In contrast, SVM was the least accurate and scored the lowest in each evaluation. The results related to recall show that RF was better than GB. However, our research outcomes differ slightly from the literature reviews. We used PH levels and Sulfate to evaluate water quality, which yielded higher accuracy for some algorithms. In addition, we achieved even greater accuracy by utilizing all features.

	Accuracy	Recall	Precision	F1-score
Random Forest	76%	73%	67%	70%
Gradient Boosting Machine	77.1%	79%	67%	73%
Support Vector Machine	63%	61%	52%	56%

Table 9: Accuracy Using pH and Sulfate Feature

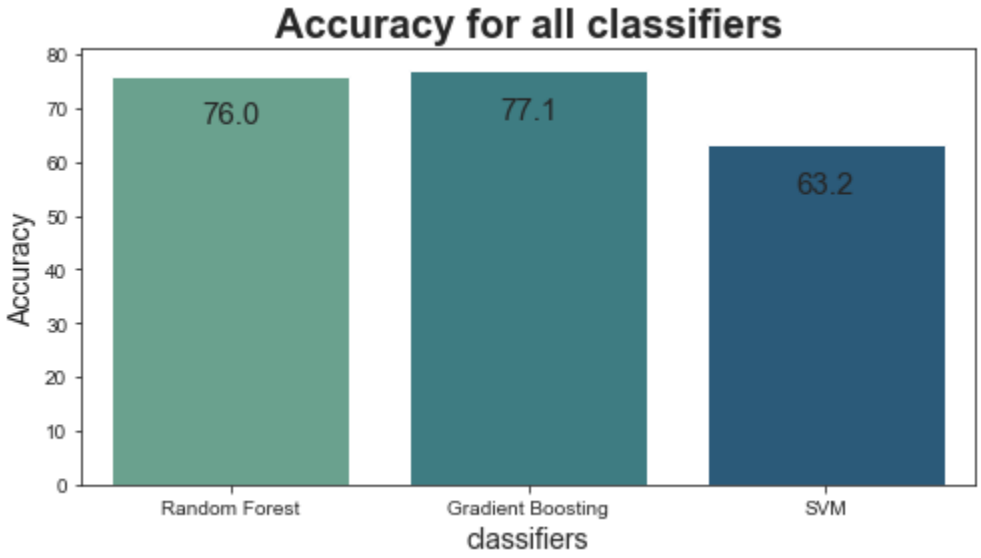


Figure 4 Accuracy Using pH and Sulfate Features

5.1 Results of Examining the Effect of Feature Selection

This table compares the model with feature selection and the algorithm with all features. As we can see from the table, the embedded method has a small impact on the results. While feature selection reduces the accuracy of the RF and GB by 1%, and increases the accuracy of the SVM by 5%.

	Random Forest	**Gradient Boosting Machine**	**Support Vector Machine**
Using features: All	76%	77.1%	63%
Using features: 2 features (Ph and Sulfate)	75.1%	75.6%	67%

Table 10 : Comparison between the accuracy with two feature (Ph and Sulfate) and with all the feature.

Output of Embedded method

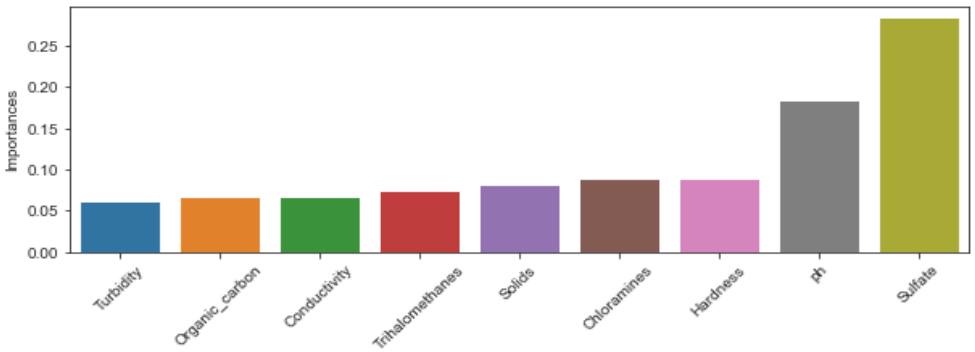


Figure 5 Output of Embedded method

Table 11 shows the most effective performance of all models after embedding. Embedded algorithms give the PH and Sulfate as the best subset of features. This increases the accuracy of certain algorithms like SVM while decreasing others like Random Forest and Gradient Boosting. In the end, we conclude that all the features of the water are essential, and it gives us higher accuracy and better results than the other study that uses the same data.

	Accuracy	**Recall**	**Precision**	**F1-score**
Random Forest	75.1%	76%	65%	70%
Gradient Boosting Machine	75.6%	77%	66%	71%
Support Vector Machine	67%	64%	56%	60%

Table 11 Accuracy After Implementing Embed Method

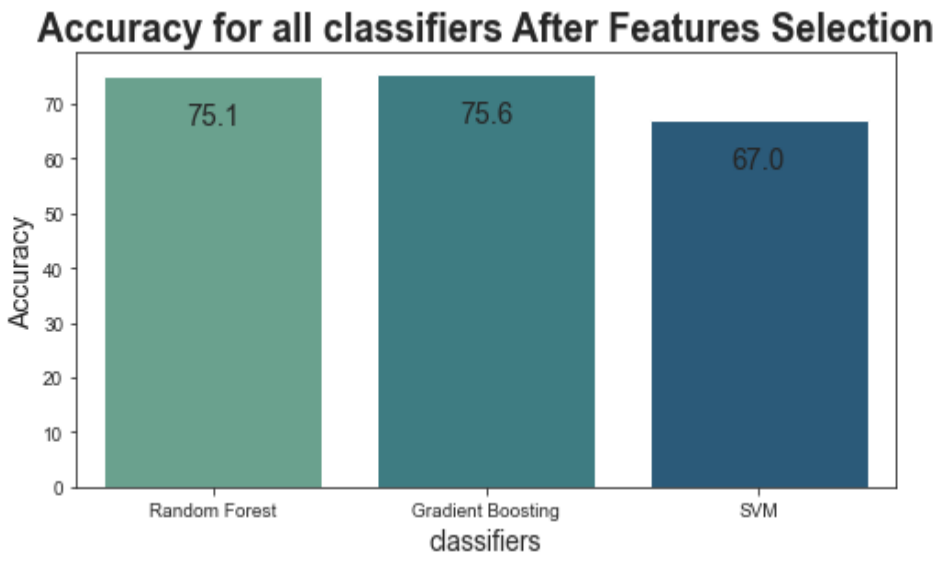


Figure 6 Accuracy After Embedded Method

6 Conclusion

Water is a crucial source of living, and examining each water source can be difficult and time-consuming. Besides, it requires a certain level of expertise. However, considering the use of Machine Learning (ML) for such problems can help in analyzing and detecting water quality (WQ) in a short time at a low cost. In our study, a dataset consisting of 3276 water samples and 9 attributes was used in the following models: Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting Machine (GBM). Based on the models and how they performed, it founds that Random Forest (RF) outperformed the other algorithms with an accuracy value of 75.1%, 76% Recall, 65% Precision, and 70% F1-Score. The discussed methods in this paper can be used to help in detecting water quality and classify if it is potable or not in an easier way compared to examining it without the use of Machine Learning. In addition to the proposed work, in future studies, a multiclass classification can be used to classify water quality in more detailed categories[2].

7 References

[1] L. A. Freeman, D. R. Corbett, A. M. Fitzgerald, D. A. Lemley, A. Quigg, and C. N. Steppe, "Impacts of urbanization and development on estuarine ecosystems and water quality," *Estuaries Coast.*, vol. 42, no. 7, pp. 1821–1838, 2019.

[2] F. Fernández-Luqueño *et al.*, "Heavy metal pollution in drinking water - a global risk for human health: A review," *Afr. J. Environ. Sci. Tech.*, vol. 7, no. 7, pp. 567–584, 2013.

[3] Muharemi, F., Logofătu, D., & Leon, F. (2019). Machine learning approaches for anomaly detection of water quality on a real-world data set. *Journal of Information and Telecommunication*, *3*(3), 294–307. https://doi.org/10.1080/24751839.2019.1565653

[4] Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient water quality prediction using supervised machine learning. *Water (Switzerland)*, *11*(11). https://doi.org/10.3390/W11112210

[5] Radhakrishnan, N., & Pillai, A. S. (2020). *Comparison of Water Quality Classification Models using Machine Learning*. 1183–1188. https://doi.org/10.1109/ICCES48766.2020.9137903

[6] Gupta, A. N., Kumar, D., & Singh, A. (2021). Evaluation of Water Quality Based on a Machine Learning Algorithm and Water Quality Index for Mid Gangetic Region (South Bihar plain), India. *Journal of the Geological Society of India*, *97*(9), 1063–1072. https://doi.org/10.1007/S12594-021-1821-0/METRICS

[7] Kouadri, S., Elbeltagi, A., Islam, A. R. M. T., & Kateb, S. (2021). Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast). *Applied Water Science*, *11*(12), 1–20. https://doi.org/10.1007/S13201-021-01528-9/TABLES/9

[8] Malek, N. H. A., Yaacob, W. F. W., Nasir, S. A. M., & Shaadan, N. (2022). Prediction of Water Quality Classification of the Kelantan River Basin, Malaysia, Using Machine Learning Techniques. *Water 2022, Vol. 14, Page 1067*, *14*(7), 1067. https://doi.org/10.3390/W14071067

[9] Lee, H. W., Kim, M., Son, H. W., Min, B., & Choi, J. H. (2022). Machine-learning-based water quality management of river with serial impoundments in the Republic of Korea. *Journal of Hydrology: Regional Studies*, *41*, 101069. https://doi.org/10.1016/J.EJRH.2022.101069

[10] Derdour, A., Jodar-Abellan, A., Pardo, M. Á., Ghoneim, S. S. M., & Hussein, E. E.

(2022). Designing Efficient and Sustainable Predictions of Water Quality Indexes at the Regional Scale Using Machine Learning Algorithms. *Water (Switzerland)*, *14*(18). https://doi.org/10.3390/w14182801

[11] Suwadi, N. A., Derbali, M., Sani, N. S., Lam, M. C., Arshad, H., Khan, I., & Kim, K.-I. (2022). *An Optimized Approach for Predicting Water Quality Features Based on Machine Learning*. https://doi.org/10.1155/2022/3397972

[12] Palade, V., Kharroubi, S. A., Tchounwou, P. B., Mohd, J., Hoque, Z., Azlina, N., Aziz, A., Alelyani, S., Mohana, M., & Hosain, M. (2022). Improving Water Quality Index Prediction Using Regression Learning Models. *International Journal of Environmental Research and Public Health 2022, Vol. 19, Page 13702*, *19*(20), 13702. https://doi.org/10.3390/IJERPH192013702

[13] Kaddoura, S. (2022). Evaluation of Machine Learning Algorithm on Drinking Water Quality for Better Sustainability. *Sustainability (Switzerland)*, *14*(18). https://doi.org/10.3390/su141811478

[14] *pH | US EPA*. (n.d.). Retrieved May 20, 2023, from https://www.epa.gov/caddis-vol2/ph

[15] *How to Choose a Machine Learning Algorithm: A Simple Step-By-Step Guide*. (n.d.). Retrieved May 9, 2023, from https://labelyourdata.com/articles/how-to-choose-a-machine-learning-algorithm

[16] Understanding Random Forest. How the Algorithm Works and Why it Is… | by Tony Yiu | Towards Data Science
https://towardsdatascience.com/understanding-random-forest-58381e0602d2
Accessed: 2023-05-11

[17] onesmus Mbaabu. "Introduction to Random Forest in Machine Learning | Engineering Education (EngEd) Program." *Section.io*, 11 December 2020, https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/.
Accessed 10 May 2023.

[18] Random Forest Classification with Scikit-Learn | by Adam Shafi | datacamp
https://www.datacamp.com/tutorial/random-forests-classifier-python
Accessed: 2023-05-10

[19] Support Vector Machine (SVM) Algorithm - Javatpoint

https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm
Accessed: 2023-05-10

[20] Chauhan, V. K., Dahiya, K., & Sharma, A. (2019). Problem formulations and solvers in linear SVM: a review. *Artificial Intelligence Review*, *52*(2), 803–855. https://doi.org/10.1007/S10462-018-9614-6/METRICS

[21] SUPPORT VECTOR MACHINES(SVM). Introduction: All you need to know… | by Ajay Yadav | Towards Data Science
https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589
Accessed: 2023-05-10

[22] Support Vector Machines: Types of SVM [Algorithm Explained] | upGrad blog
https://www.upgrad.com/blog/support-vector-machines/
Accessed: 2023-05-11

[23] The Evolution of AI: Tracing the History and Progress of Gradient Boosting in Artificial Intelligence – TS2 SPACE
https://ts2.space/en/the-evolution-of-ai-tracing-the-history-and-progress-of-gradient-boosting-in-artificial-intelligence/
Accessed: 2023-05-10

[24] GBM in Machine Learning - Javatpoint
https://www.javatpoint.com/gbm-in-machine-learning
Accessed: 2023-05-10

[25] Gradient Boosting - Definition, Examples, Algorithm, Models
https://www.wallstreetmojo.com/gradient-boosting/
Accessed: 2023-05-11

[26] Gradient Boosting for Classification | Paperspace Blog
https://blog.paperspace.com/gradient-boosting-for-classification/
Accessed: 2023-05-10

[27] Gradient Boosting Decision Tree Algorithm Explained | by Cory Maklin | Towards Data Science
https://towardsdatascience.com/machine-learning-part-18-boosting-algorithms-gradient-boosting-in-python-ef5ae6965be4

Accessed: 2023-05-10

[28] Python - Replace Missing Values with Mean, Median & Mode - Data Analytics https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/ Accessed: 2023-05-29

[29] *Precision and Recall | Essential Metrics for Data Analysis*. (n.d.). Retrieved May 20, 2023, from https://www.analyticsvidhya.com/blog/2020/09/precision-recall-machine-learning/

Plagiarism Report

ml

ORIGINALITY REPORT

13%	10%	9%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Hye Won Lee, Min Kim, Hee Won Son, Baehyun Min, Jung Hyun Choi. "Machine-learning-based water quality management of river with serial impoundments in the Republic of Korea", Journal of Hydrology: Regional Studies, 2022 Publication	1%
2	www.ncbi.nlm.nih.gov Internet Source	1%
3	link.springer.com Internet Source	1%
4	www.researchgate.net Internet Source	1%
5	Nur Afyfah Suwadi, Morched Derbali, Nor Samsiah Sani, Meng Chun Lam, Haslina Arshad, Imran Khan, Ki-Il Kim. "An Optimized Approach for Predicting Water Quality Features Based on Machine Learning", Wireless Communications and Mobile Computing, 2022 Publication	1%