# Water Potability Prediction Using Supervised

# Machine Learning Algorithms

Rohith Chityala

Faculty of Engineering, Environment and Computing

Coventry University

Coventry, England chityalar@uni.coventry.ac.uk

**Abstract— In this paper, the application of different supervised machine learning algorithms to predict potability of water is discussed. The three Machine Learning algorithms used in this study are Random Forest algorithm, XGBoost algorithm, and K Nearest Neighbour algorithm. These algorithms are applied on a dataset containing parameters that define the quality of water. There is total 3276 rows and 10 columns in this dataset out of which 9 columns are independent variables. Whereas 1 variable i.e., potability is considered as dependent variable or target variable. Before implementing the machine learning models, an exploratory data analysis has been performed on the dataset, which also includes removal of null value and standardising the data. The dataset has been split into training and testing sets in the ratio of 70:30. The machine learning models performance is evaluated and compared based on their accuracy scores. Among the three algorithms applied, Random Forest has shown the best performance with an accuracy of 69%. The XGBoost algorithm, and K Nearest Neighbour algorithm have resulted with an accuracy of 66% and 65%. The performance of these models is visualised by another evaluation metric known as Confusion Matrix.**

I. INTRODUCTION – PREDICTING WATER QUALITY

Water is an essential part of Earth’s hydrosphere. It plays a vital role in keeping all forms of life alive. Water is available in various forms on earth such as rivers, lakes, seas, oceans, ponds, etc. It is important to maintain the quality of water at different levels depending on its usage. To lead a healthy lifestyle, it is important to make sure that the quality of water is suitable for drinking. But due to various types of pollution in recent years, the quality of water has been declining at an alarming rate. Hence, water quality has become a burning issue in today’s world, where many governmental and non-governmental organisations are working hard to improve the quality of water. A report published by the United Nations (UN) estimates that the deaths caused due to water contamination every year is 1.5 million. It is also estimated that 80% of the health issues in developing countries are caused due to water contamination. So, it is critical to analyse and test the quality of water from time to time. There are different ways of analysing the quality of water with the traditional method being collecting and testing samples from various water bodies. Many developed countries are using impactful water treatment methods before making it available for drinking. The traditional methods of testing or analysing the quality of water are often seemed to be time-consuming and expensive.

So, researchers are finding new ways to test or predict the quality of water to minimise the costs and any risks included. In today’s world, data plays a crucial role in solving many problems. Data related to different parameters of water that define its quality is being collected with the use of IoT monitoring devices. With the development and application of data science and machine learning techniques on the data, it is being possible to predict certain factors that can provide solutions to a problem. This study focuses on implementing different machine learning algorithms on the water data to predict if it is potable or not.

The structure of the report is as follows: Chapter 1 gives an introduction about the problem and proposed techniques to tackle it. Chapter 2 provides a brief description of previous studies published related to the problem. Chapter 3 introduces methods chosen to analyse the problem. Chapter 4 discusses the dataset and tools used to implement this study. Chapter 5 describes how the implementation has been done. Chapter 6 reviews the results obtained followed by Chapter 7 which discusses conclusion and future work of this study.

II. LITERATURE REVIEW

Different papers have been published in recent years that studied and experimented various methodologies to address the water quality problem. Below mentioned is a review of some of these papers.

(Amir et al., 2018) published a paper that discusses application of Machine Learning algorithms such as ANN, Group method of data handling (GMDH), and SVM. The dataset used in this study consists of data related to Tireh river in Iran. The accuracy of GMDH model was found to be slightly less than ANN and SVM models. Based on the DDR evaluation metrics used, SVM proved to be the best performer among the three models implemented.

A study proposed by (Aldhyani et al., 2020) is an implementation of advanced Artificial Intelligence algorithms to predict the water quality index (WQI) and water quality classification (WQC). Different neural network algorithms used in this study are LSTM, Non-linear auto aggressive neural network (NARNET), SVM, K-NN, and Naïve Bayes classifier.

The dataset used for this study is a collection of data related to drinking water supply in India. The NARNET model has shown a better performance with 96% accuracy to predict WQI values. On the other hand, SVM has gained highest accuracy of 97% to predict WQC values.

Another study proposed by (Khoirul et al., 2021) reviews the prediction of water quality by applying two machine learning algorithms which are Naïve Bayes and Decision Tree. These models were evaluated using cross-validation techniques. Naïve Bayes algorithm was further divided into four types. Compared to Naïve Bayes algorithm, Decision tree algorithm has shown highest accuracy score of 97%.

III. METHODOLOGY

Machine learning is the concept where a machine is designed to learn from the data without any human interference. Based on the design and application, machine learning is divided into three types- Supervised learning, Unsupervised learning, and Reinforcement learning. Unlike Unsupervised algorithms, Supervised algorithms are those that use labelled data to predict or classify the outcome. In this study, three different types of Supervised learning algorithms are used to predict potability of water. A brief explanation of these algorithms is described in the below section.

*A. RANDOM FOREST CLASSIFIER*

Random Forest is a type of ensemble learning technique that involves combining vast number of decision trees to predict an outcome rather than using a single decision tree. The ensemble works in parallel and assigns classifications based on a majority vote. When compared to a single classifier, this enhances classification accuracy in an ensemble technique. The advantages of Random Forest include its ability to perform both classification and regression tasks and it can handle the data imbalance problem very well.

*B. K-NEAREST NEIGHBOUR (KNN)*

K-Nearest Neighbour often referred to as K-NN is a supervised machine learning algorithm that predicts or classifies the incoming data by calculating the distance between new and existing data points. The letter K in K-NN denotes the number of nearest points for which the distance needs to be calculated. It can be used to solve both regression and classification problems. Despite being the simplest machine learning algorithm, K-NN is known as a lazy learner and the computational cost is high compared to other algorithms.

*C. XGBOOST ALGORITHM*

eXtreme Gradient Boost or XGBoost algorithm is an ensemble learning technique that is scalable, distributed gradient-boosted decision tree (GBDT) machine learning library.

It provides reliable results by using successively created short decision trees and a highly flexible training strategy that minimises overfitting. Gradient boosting techniques have been developed in an idea to improve a single weak model by combining it with other weak models.

*D. EVALUATION METRICS*

* Accuracy

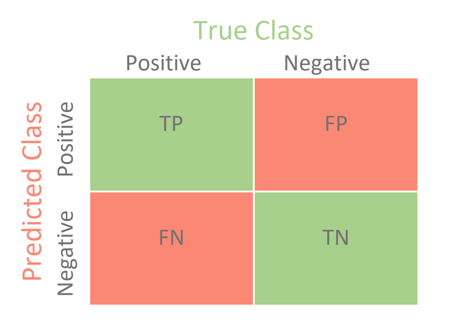
Accuracy score is a type of evaluation metrics used to measure the performance of a machine learning model. It is the ratio of total number of true values predicted to the total number of values predicted by a model. The below shown equation represents the formula of accuracy score. The total number of true values predicted include the True Positive and True Negative values.

Accuracy =

* Confusion Matrix

A Confusion matrix of a machine learning model indicates the predicted number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values by a machine learning algorithm. Confusion matrix acts as a baseline for different evaluation metrics such as Precision, Recall, and Accuracy scores. Below shown Figure 1 is a representation of 2\*2 confusion matrix.

Figure 1 *Confusion Matrix*



Note: From Confusion Matrix for binary classification by ([https://miro.medium.com/max/700/1\*fxiTNIgOyvAombPJx5KGeA.png](https://miro.medium.com/max/700/1*fxiTNIgOyvAombPJx5KGeA.png))

IV. THE DATASET

The data used in this study is a dataset available publicly on Kaggle. It consists of 3276 observations of 10 different variables such as Ph value, hardness, solids, chloramines, sulfate, conductivity, organic carbon, trihalomethanes, turbidity, and potability. Of all the 10 variables, 9 are considered as predictor variables with potability being the target variable. Potability variable is divided into two different classes- 0 and 1, where 0 denotes water is not potable and 1 denotes the water is potable. The below shown table 1 represents the list of features present in the dataset, the number of unique values in each variable, and their data types.

Table 1 *List of Variables*

|  |  |  |  |
| --- | --- | --- | --- |
| S.No | Name of the Variable | Number of unique values | Data type |
| 1 | pH | 2785 | Float64 |
| 2 | Hardness | 3276 | Float64 |
| 3 | Solids | 3276 | Float64 |
| 4 | Chloramines | 3276 | Float64 |
| 5 | Sulfate | 2495 | Float64 |
| 6 | Conductivity | 3276 | Float64 |
| 7 | Organic\_carbon | 3276 | Float64 |
| 8 | Trihalomethanes | 3114 | Float64 |
| 9 | Turbidity | 3276 | Float64 |
| 10 | Potability | 2 | Int64 |

The proposed methodology is implemented in Jupyter notebook using Python language. Different libraries have been imported to accomplish the study’s objectives. NumPy and Pandas are imported to perform any operations on the dataset. Matplotlib and Seaborn are imported to generate visualisation plots. The machine learning algorithms and their evaluation metrics have been imported using scikit-learn library.

V. EXPERIMENTAL SETUP

Firstly, the dataset has been loaded into the Jupyter Notebook using the Pandas library. A statistical analysis of the data has been performed using the describe function. The results obtained by using this function include various statistical parameters such as minimum value, maximum value, 25th percentile, 75th percentile, median, count, mean, standard deviation value of every single variable.

1. *TARGET VARIABLE ANALYSIS*

As a part of exploratory data analysis, the outcome variable, Potability has been analysed visually using bar plot and pie chart. The table shown below represents the number of 0s and 1s present in the Potability variable.

Table II *Shape of Output variable*

|  |  |
| --- | --- |
| Type of Class | Number of values |
| 0 | 1998 |
| 1 | 1278 |

The figure 1 and figure 2 shown below are the count plots obtained using matplotlib and seaborn library. It is found that 39% of the output data is Potable and 61% of the output data is Not Potable.

Figure 1 *Pie Chart Showing percentage of 0s and 1s*

Chart, diagram, sunburst chart

Description automatically generated

Figure 2 *Bar Graph showing count of 0s and 1s*

Chart, bar chart

Description automatically generated

1. *NULL VALUE ANALYSIS*

In every dataset, the probability of presence of null values is high. It is advantageous to clean the data before implementing the machine learning models. Cleaning the data often includes filling missing or null values and outlier treatment. In this study, it has been found that there are certain number of null values present in the dataset, which is shown in the below table III. Only three variables, ph, trihalomethanes, and sulfate consisted of null values.

These null values have been filled by mean value of that specific variable.

Table III *Count of Null Values*

|  |  |
| --- | --- |
| Variable | No of Null Values |
| pH | 491 |
| Hardness | 0 |
| Solids | 0 |
| Chloramines | 0 |
| Sulfate | 781 |
| Conductivity | 0 |
| Organic\_carbon | 0 |
| Trihalomethanes | 162 |
| Turbidity | 0 |
| Potability | 0 |

1. *CORRELATION ANALYSIS*

Correlation analysis is a type of statistical analysis that deals with understanding the relation between all the variables present in the dataset. This analysis explains the hidden relationships present between any two variables. It is measured by calculating a variable called Correlation Coefficient. The correlation coefficient value is measured in the range of -1 to +1, where value 1 indicates there is a perfectly linear positive correlation and -1 denotes perfectly linear negative correlation. In this study, correlation has been analysed by visualising heatmap as show in below figure 3.

Figure 3 *Heatmap of Correlation*

A picture containing graphical user interface

Description automatically generated

D) *TRAIN-TEST SPLIT*

To make the data ready to fit the machine learning models, standardizing of the data has been done. Standardizing the data is a type of feature scaling technique that makes the data internally consistent to implement the models. It basically rescales all the variables to a single scale or range of values. This avoids any bias action that can be occurred. After standardizing the data, the mean and variance of all the variables is reduced to 0 and 1. This has been achieved by importing Standard Scaler function from sklearn library. After standardizing the data, splitting of the dataset has been done. The data has been split into training and testing sets of data in the ratio of 70:30. 70% of the data is considered as training data and 30% is considered as testing data. The models are first trained on the training data and their performance is calculated by fitting the models on testing set of data. The below shown table IV represents the shape of training and testing sets of data.

Table IV *Shape of Train and Test data*

|  |  |
| --- | --- |
| X\_train | (2293, 9) |
| X\_test | (983, 1) |
| y\_train | (2293, 1) |
| y\_test | (983, 1) |

VI. EXPERIMENTAL RESULTS

In this section, the machine learning models implemented, and their accuracy scores calculated are discussed. All the algorithms have been implemented by importing respective functions from sklearn library.

Random Forest Classifier has shown an accuracy score of 69% approximately. The below shown figure 4 is a 2\*2 confusion matrix obtained for Random Forest classifier.

Figure 4 *Confusion Matrix of Random Forest Classifier*

Chart, treemap chart

Description automatically generated

Random Forest has predicted the actual potable water as potable 130 times and not potable as 547 times.

When XGBoost classifier has been applied on the dataset, an accuracy score of 66% is obtained. The below shown figure 5 is a confusion matrix of the XGBoost classifier. We can see that compared to Random Forest, XGBoost has predicted a greater number of true positive values.

Figure 5 *Confusion Matrix of XGBoost Classifier.*

Chart, treemap chart

Description automatically generated

K-Nearest Neighbour classifier has performed with an accuracy score of 65%, which is slightly lower than the XGBoost algorithm. The below shown figure is the Confusion Matrix obtained for the K-NN classifier implemented. We can see from the confusion matrix that the number of True negatives predicted by K-NN is higher compared to other two algorithms.

Figure 6 *Confusion Matrix of K-Nearest Neighbour Classifier*

Chart, treemap chart

Description automatically generated

VII. CONCLUSION AND DISCUSSION

In this paper, the prediction of potability of water is analysed by applying machine learning algorithms. Initially, the importance of water and why it is vital to maintain the quality of water is studied. To overcome the limitations of traditional water quality prediction techniques, an implementation of machine learning algorithms has been experimented to predict the water potability. The dataset used to implement the algorithms is taken from Kaggle. It contains of 3276 rows and 10 columns. The output column or target variable is the Potability column which consists of two unique values 0 and 1. The dataset is loaded into the Jupyter notebook platform by applying Python language. As a part of understanding and cleaning the data, exploratory data analysis has been performed on the dataset. The missing values present in the dataset have been replaced by the mean value of the respective variables. The implementation of machine learning algorithms was done by standardising the data, initially. After standardizing the data, the dataset has been split into train and test sets. The three chosen machine learning algorithms, Random Forest Classifier, K-Nearest Neighbour, and XGBoost algorithm have been applied on the dataset using scikit-learn library. Of all the three algorithms implemented, Random Forest has shown a greater accuracy score of 69% compared to other two models. The K-NN algorithm and XGBoost algorithm’s performance is slightly varied with an accuracy score of 66% and 65%.

Every study comes with few limitations. One of the limitations in this study is the non-availability of variety kind of dataset. A dataset with diverse range of parameters or with a higher size can be used to obtain higher accuracies of machine learning algorithms. Research could be done on increasing the accuracy scores of the algorithms by using optimisation techniques such as Gridsearchcv and Randomsearchcv. Optimisation techniques often result in improved accuracy scores of the machine learning models. Different evaluation metrics such as precision, recall, F-1, AUC-ROC score, and curve can also be used to evaluate the performance of models.

VIII. APPENDIX

Dataset Link: <https://www.kaggle.com/code/emrearslan123/water-potability-prediction/data>

Code Link:

REFERENCES

[1] Amir Hamzeh Haghiabi, Ali Heidar Nasrolahi, Abbas Parsaie; Water quality prediction using machine learning methods. Water Quality Research Journal 1 February 2018; 53 (1): 3–13. doi: https://doi.org/10.2166/wqrj.2018.025

[2] Aldhyani, T., Al-Yaari, M., Alkahtani, H., & Maashi, M. (2020). Water Quality Prediction Using Artificial Intelligence Algorithms. *Applied Bionics And Biomechanics*, 2020, 1-12. https://doi.org/10.1155/2020/6659314

[3] M. I. Khoirul Haq, F. Dwi Ramadhan, F. Az-Zahra, L. Kurniawati and A. Helen, "Classification of Water Potability Using Machine Learning Algorithms," *2021 International Conference on Artificial Intelligence and Big Data Analytics*, 2021, pp. 1-5, doi: 10.1109/ICAIBDA53487.2021.9689727.

[4] Fu, Zhao, "Water Quality Prediction Based on Machine Learning Techniques" (2020). *UNLV Theses, Dissertations, Professional Papers, and Capstones.* 3994. http://dx.doi.org/10.34917/22110053

[5] Siyuan Liu, Yunfan Yang, Muhan Yao, "A comparison of five commonly used machine learning algorithms applied to potable water detection," *Proc. SPIE 12160, International Conference on Computational Modeling, Simulation, and Data Analysis (CMSDA 2021), 121601R* (11 March 2022); https://doi.org/10.1117/12.2627657

[6] Najah Ahmed, A., Binti Othman, F., Abdulmohsin Afan, H., Khaleel Ibrahim, R., Ming Fai, C., & Shabbir Hossain, M. et al. (2019). Machine learning methods for better water quality prediction. *Journal Of Hydrology*, 578, 124084. https://doi.org/10.1016/j.jhydrol.2019.124084

[7] Haghiabi, A., Nasrolahi, A., & Parsaie, A. (2018). Water quality prediction using machine learning methods. *Water Quality Research Journal*, 53(1), 3-13. https://doi.org/10.2166/wqrj.2018.025

[8] R. Zhao, "The Water Potability Prediction Based on Active Support Vector Machine and Artificial Neural Network," *2021 International Conference on Big Data, Artificial Intelligence and Risk Management (ICBAR),* 2021, pp. 110-114, doi: 10.1109/ICBAR55169.2021.00032.

[9] Ighalo, J., & Adeniyi, A. (2020). A comprehensive review of water quality monitoring and assessment in Nigeria. *Chemosphere,* 260, 127569. https://doi.org/10.1016/j.chemosphere.2020.127569

[10] Confusion Matrix for Your Multi-Class Machine Learning Model.

Medium. (2022). Retrieved 21 April 2022, from https://towardsdatascience.com/confusion-matrix-for-your-multi-class-machine-learning-model-ff9aa3bf7826.