https://doi.org/10.58864/mrijet.2023.10.1.8

Srinidhi Iyer
Department of ECE, FET,
Manav Rachna International Institute of Research and Studies, Faridabad, Haryana, India
E-mail: srinidhi.iyer18@gmail.com

Simran Kaushik
Department of ECE, FET,
Manav Rachna International Institute of Research and Studies, Faridabad, Haryana, India
E-mail: kaushiksimran827@gmail.com

Poonam Nandal
Professor, Department of CSE, FET,
Manav Rachna International Institute of Research and Studies, Faridabad, Haryana, India
E-mail: poonamnandal.fet@mriu.edu.in

Water Quality Prediction Using Machine Learning

Abstract: *This paper shows the use of ML algorithms for the prediction of water quality. The model is trained on Water Quality dataset from Kaggle and it consists of key features such as, pH value, hardness, solids etc. Algorithms used were SVM, Random Forest and Decision Tree. Also, hyperparameter tuning was done in SVM for improving the accuracy using Grid Search technique. The Random Forest algorithm out-performed the others with an accuracy of 68%. Hence, it shows that ML can be used for predicting the quality of water.*

Keywords: *Water Quality prediction, SVM, Random Forest, Decision Tree, Hyperparameter Tuning, Grid Search*

I. INTRODUCTION

On Earth, the basic necessity for the functioning of all the lifeforms is water. Earth being the only planet to support life, it is safe to say that water is the reason behind it. About 96% of all earth's water, about 71% of the earth's surface is water-covered and the oceans hold only 2.5% for our essential things. Every day, the human body must require the nearly 3-litre water. We are facing the problem of water pollution irrespective of the fact that we see and have rivers and lakes around the city.

Water Cycle is one of the processes in which water changes its state one to another having the same amount of water. Pollution of water is any change, major or minor in the biological, physical or chemical properties of water that leads to a detrimental consequence eventually of any living organism.

Some of the water pollutions are caused by Direct Sources, that directly releases dangerous by-products and wastes without treating then into the nearest water source, such as factories, waste management facilities, refineries, etc and Indirect Sources include pollutants that infuse in the water bodies through groundwater or via the atmosphere through acidic rain or soil.

There are some specific water quality standards that can be used to indicate the quality of the water bodies such as Stream Standards, Effluent Standards, Drinking water standards, etc. Moreover, other applications/usages possess their standards for water specification. For example, irrigation water transferred to plants or soil must be neither too saline nor contain toxic materials thus destroying the ecosystem.

There are different properties based on some specific industrial processes for the water quality for the industrial purposes. Some of the natural resources of fresh water, such as ground and surface water, are low-priced resources. However, industrial/human activities and other natural processes can pollute these resources.

Therefore, surveillance of water quality is important as if not properly monitored, it can harm the living organisms. There are several manual methods available for the prediction of the water quality but they are expensive and consumes a lot of time. This method has limitations too. For example, if a farmer wants to check the quality of water for supplying it to the crops, then he has to wait for a week or more for all the manual work to take place for its measurement.

However, sensors can be used to take the value of different parameters to predict the water quality manually. This will take a lot of time and therefore, ML technology can be one of the solutions to provide the predictions.

Causing a rate of increased acceleration, today technology is at rapid pace, allowing faster progress and change. For the prediction and modelling of the Water Quality, several methodologies are proposed. Artificial Intelligence, Machine Learning, Deep learning, etc are the few technologies that are emerging at the fastest rate. Therefore, there methodologies can be used in the prediction which includes statistical approaches, visual modelling, analysing and predictive algorithms, etc.

Modelling and prediction of water quality are available in two types which are non-mechanism and mechanism models. In this paper, several Machine Learning algorithms were tested for the prediction of water quality.

II. LITERATURE SURVEY

In 2018, Ali Heidar Nasrolahi along with Amir Hamzeh Haghiabi and Abbas Parsaie predicted the Water Quality of a river bed in Iran Tireh River by taking pH, Na, Ca, Mg, etc such components into consideration. Performance was tallied by using several ML and DL algorithms. It was observed that results of SVM was the front runner and gave the best accuracy. ANN gave acceptable accuracy for practical purposes.[1]

In 2019, Umair Ahmed et.al explained ways to efficiently predict water quality using supervised Machine Learning. Harrowing diseases have been in increased proportions due to the depreciation and deterioration of water quality at an alarming rate which was a direct impact of rapid urbanization and industrialisation. Their research monitors and works with supervised Machine Learning algorithms to calculate Water Quality Index (WQI) and Water Quality Class (WQC), the former being a singular index which describes the general quality of water and the latter being the derivative and distinctive class on the basis of WQI.[2]

In 2020, Mohammed Al-Yaari et.al illustrated the use of Artificial Intelligence algorithms along with the performance of each used algorithm. As we know, for the protection of the environment, predicting and modelling of the quality of water is immensely important. In the methodology they proposed, to predict WQI, artificial intelligence algorithms, such as, NARNET and LSTM were used. Along with this, KNN, SVM and Naïve Bayes algorithms were also implemented. They used a dataset with 7 relevant and significant features and statistical parameters were used to develop the model and evaluate them.[3]

In 2020, Navideh Noori et.al explained the water quality prediction using SWAT-ANN coupled approach. For solving environmental problems Machine Learning algorithm such as Artificial Neural Networks is being used widely. They illustrated the application of SWAT-ANN for water quality prediction.[4]

In 2022, Jin-Won Yu et.al explained the use of AI algorithms for the water quality prediction. Combined the power of data decomposition, fuzzy C-means clustering and bidirectional gated recurrent model for the prediction of water quality.[5]

In 2022, Manisha Koranga et.al discussed the use of Machine Learning Algorithms for water quality prediction for Nanital Lake, Uttarakhand. Analysed the use of machine learning algorithms and used eight regression algorithms and nine classification algorithms. Three algorithms Random Forest, SVM and Stochastic Gradient Descent comes out to be the most effective machine learning algorithms.[6]

Reviewing the literature shows that artificial intelligence techniques have been proposed for water conservation projects for which water quality prediction and assessment plays an important role. Hence, this paper presents a designed algorithm for the prediction of Water Quality considering the concentration, pH, duration factors.

III. DATA PREPROCESSING AND WATER QUALITY INDEX PREDICTION

To improve the water quality, pre-processing phase plays a vital role in data analysis. For the calculation of the Water Quality Index, the most significant factors are taken into consideration. For the system's superior accuracy, Data Normalization Techniques has been used.

IV. SUPPORT VECTOR MACHINE

One of the most popular supervised learning algorithms is Support Vector Machine, it can be used for Regression and classification problems. Widley, it is used for Classification problems in Machine Learning.

Creation of the decision boundary (which is the best area or plane or line) that helps to sort n-dimensional data space into classes. This helps us to put the new query point in the accurate category in the future. Whenever there's a new query point, it is compared to the decision boundary and is classified accordingly. This is the main goal of Support Vector Machine. Decision boundary which will suit the best for a particular dataset is called a Hyperplane.

Hyperplane consists of extreme points/vector. Extreme cases indicate datapoints which lie in all the extremities such data points are termed as support vectors. Since, the whole algorithm is based on these extremities it referenced as Support Vector Machine.

In the SVM algorithm, we use the loss function it helps us to the maximize the margin hinge loss.

V. DECISION TREE

Decision tree is a supervised machine learning algorithm which can be used for both classification and regression.

It is mostly preferred to solve classification problems in which data has to be classified into different categories. It has a tree like structure with internal node acting as features, branches as decision rule and leaf nodes as the outcome. It is basically a graphical representation of all the possible outcomes for a given problem based on the given conditions. It simply asks a question and based on the answers it further splits the trees. In order to build a tree, it uses CART algorithm.

VI. RANDOM FOREST

Random Forest is a classifier that contains number of decision tree on various subsets. It is a supervised machine learning algorithm used for both classification and regression.

To improve the predicted accuracy of the dataset it takes the average of the decision trees.

It is based on the concept of ensemble learning, which is basically a process of combining several classifiers to solve complex problems and to improve the performance of the model. To have the higher accuracy and prevent the problem of overfitting, greater number of trees are used.

VII. CONFUSION MATRIX

For classification problems confusion matrix is used on a wide scale. it is used for multiclass classification problems and binary classifications as well. Counts from predicted and actual values is represented by confusion matrices.

In confusion matrices True Negative is represented by "TN" it shows the number of negative examples which were labelled correctly. In the same way, True Positive is represented by "TP" it shows the number of positive samples which were labelled correctly. False Positive is represented by "FP" it shows the number of actual negative samples which were classified as positive. And False Negative is represented by "FN" it shows the number of actual positive samples classified as negative. For evaluation of WQI model we use Accuracy, Precision, Recall, Specificity, Mean Square Error, Sensitivity.
These statistical measurements are as

mentioned below:

- Accuracy=TP+TN/TP+FP+FN+TN
- Precision=TP/TP+FP
- Recall=TP/TP+FN
- F1-score=2*P*R/(P+R)

VIII. RESULT

In this section, dataset used along with use of various Machine Learning algorithms for prediction is highlighted.

The dataset is used is from Kaggle. This dataset consists of water quality metrics for 3276 different water bodies. Key features used to tally the results are: pH value, hardness, solids, Chloramines, Sulphates, Organic carbon, Trihalomethanes, Turbidity, Potability. The results of the different machine

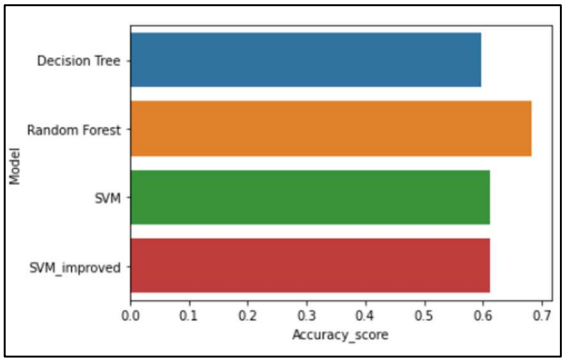
learning algorithms, namely, SVM, Decision Tree, Random Forest are discussed based on parameters like accuracy, precision, recall and F1-score.

The main aspect taken into consideration is accuracy.

Accuracy=TP+TN/TP+FP+FN+TN

The results of the algorithms are mentioned below:

	Model	Accuracy_score
1	Random Forest	0.682604
2	SVM	0.613428
3	SVM_improved	0.613428
0	Decision Tree	0.598169



IX. CONCLUSION

For the protection of the environment and the human health, Water quality prediction plays a very important role. With the advancement in the technology, Artificial and machine learning models can be used for the prediction of the same to make human life healthier and easier. In this paper, investigation of different Machine Learning algorithms on Water Quality Prediction dataset has been done.

The simulation result shows that Random Forest algorithm outperforms the other two algorithms namely Support Vector Machine and Decision Tree. The validation on test dataset provides 68% accuracy which is 8% better than other algorithms. This clearly shows the effectiveness of the technique in the prediction of Water Quality. Accuracy can further be enhanced by training the model with larger number of samples.

REFERENCES

[1] Haghiabi, A. H., Nasrolahi, A. H., &Parsaie, A. (2018). Water quality prediction using machine learning methods.

[2] Umair Ahmed, Rafia Mumtaz, Hirra Anwar, Asad A. Shah, Rabia Irfan and Jose Garcia-Nieto (2019). Efficient Water Quality Prediction Using Supervised Machine Learning.

[3] Mohammed Al-Yaari, Hasan Alkahtani and Mashael Maashi (2020). Water Quality Prediction using Artificial Intelligence Algorithms.

[4] Navideh Noori, Latif Kalin and Sabahattin Isik (2020). Water Quality prediction using SWAT-ANN coupled approach.

[5] Jin Won Yu, Ju-Song Kim, Xia Li, Yun-chol Jong, Kwang-Hun Kim and Gwang-Il Ryang (2022). Water quality forecasting based on data decomposition, fuzzy clustering and deep learning neural network.

[6] Manisha Koranga, Pushpa Pant, Tarun Kumar, Durgesh Pant, Ashutosh Kumar Bhatt and R.P. Pant (2022). Efficient water quality prediction models based on machine learning algorithms for Nainital Lake, Uttarakhand.