e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:08/August-2023 Impact Factor- 7.868 www.irjmets.com

WATER POTABILITY PREDICTION USING MACHINE LEARNING

Nirmala Malagi[*1]

[*1]Student, Department of MCA, Visvesvaraya Technological University Belagavi, Belagavi, Karnataka, India

DOI : https://www.doi.org/10.56726/IRJMETS44413

ABSTRACT

Water quality is essential for preserving environmental balance and safeguarding human health. The machine learning approach has been shown to have huge promise for highly accurate and effective water quality prediction. This research offers a thorough analysis of these techniques for forecasting water quality. The model is developed using a random forest and a support vector machine. Exploratory data analysis is utilized to understand the dataset, while preprocessing techniques are utilized to deal with outliers and null values. The evaluation metric is used to evaluate every model. The random forest achieved an accuracy of 69.20%.

Keywords: Exploratory data Analysis, Random Forest, Support Vector, Entropy, WHO

I. INTRODUCTION

Water is a important natural source for life to exist on Earth. 70% of the world is covered with water, 97% of which is ocean water; just 3% of this water is readily accessible to organisms that require fresh water to survive. Only 0.5% of the world's water is readily accessible since 2.5% of it is inaccessible because of the ice caps in the polar regions, soil, atmosphere, and significant levels of pollution. The water purity is degrading because of the home sewage (including nitrate and phosphate), the discharge of solid waste, including trash, rubbish, and electronic waste, and thermal pollution have all contributed to the degradation of the quality of this accessible water. Consuming this contaminated water will have a harmful impact on both human health and water resources. Also affected is the national economy. The main aim of this work is to design a model that can assess water and determine whether it is safe to drink or not. The World Health Organization (WHO) developed the standards for water quality. Hardness, pH, chloramines, solids, organic carbon, conductivity, sulfate, turbidity, and trihalomethanes are the criteria that were employed. The support vector machine(SVM) and random forests, are employed in this project. To determine a model's efficiency, accuracy scores are utilized. The ultimate model, a random forest, is selected with an accuracy of 69.20%.

II. METHODOLOGY

Data collection

The gathering of data that most effectively addresses the problem. Inaccurate data will lead to inaccurate findings from the model. Both static and real-time data used in data projects. It might be images, unlabeled and labeled data, etc.

https://www.kaggle.com/datasets/adityakadiwal/water-potability

The water_potability dataset for the system was utilized from Kaggle and consists of 3276 samples with 9 water quality parameters and 1 target column.

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0
...	...	...	...	...	...	...	...	...	...	...
3271	4.668102	193.681735	47580.991603	7.166639	359.948574	526.424171	13.894419	66.687695	4.435821	1
3272	7.808856	193.553212	17329.802160	8.061362	NaN	392.449580	19.903225	NaN	2.798243	1
3273	9.419510	175.762646	33155.578218	7.350233	NaN	432.044783	11.039070	69.845400	3.298875	1
3274	5.126763	230.603758	11983.869376	6.303357	NaN	402.883113	11.168946	77.488213	4.708658	1
3275	7.874671	195.102299	17404.177061	7.509306	NaN	327.459760	16.140368	78.698446	2.309149	1

3276 rows × 10 columns

Figure 1: Dataset

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:08/August-2023 Impact Factor- 7.868 www.irjmets.com

Factors Affecting Water Quality

pH: Potential hydrogen, or pH, is a unit of measurement for determining whether or not acid and basic are in balance. The Water's pH ranges from 6.5 to 8.5, according to the WHO.

Hardness: The ability of water to turn into soap is another way to assess hardness. According to the report of the WHO, drinking is safe at 60 to 120 mg/L.

Solids: The total amount of dissolved solids in the water is referred to as a solid. The usual TOS range for potable water is 500–1000 mg/L, according to the report of the WHO. Chloramines: Chloramine is a disinfectant that water suppliers add to the water purification process to kill bacteria and viruses. The WHO recommends a chloramine concentration of up to 4 mg/l as preferred for drinking water.

Sulfate : The concentration increases in drinking water have a direct influence on health.

Conductivity: The drinking water conductivity is only up to 400 S/cm, according to the WHO.

Trihalomethanes: The THM concentrations up to 80 ppm are good for human consumption.

Turbidity: The WHO states that the water's turbidity level is 5 NTU.

Exploratory Data Analysis

This step for better understand the relationships between the dataset and each individual dataset. Many data scientists utilize EDA to obtain extensive analyses of the information about datasets. Finding patterns, identifying anomalies, testing a concept, or verifying presumptions are all beneficial. In this project, it was found that 61% of the available dataset is non-potable and 39% is potable. Based on the study, every parameter in the dataset is regularly distributed. Examining the association between each feature and the target feature reveals that each feature has a unique effect on the target feature

Data Pre-processing

1. The dataset utilized for this study contains some missing values: trihalomethanes (162), pH (491), and sulfate (781). The KNNImputer is used, which replaces the null values with those of its neighbors.
   The method takes the n_neighbors parameter, which takes the number of neighbors to be considered while calculating values. The "distance" value of the weight parameter determines whether to consider the data points by the inverse of their distance. The fit and transform methods used perform imputation of values.
2. The parameters in the given dataset had a few outliers, which were eliminated after being calculated using the IQR method.

IQR=Q3-Q1

Q1=25 percentile Q3=75 percentile

Any values less than the lower limit and greater than the upper limit are treated as outliers.Lower_limit=q1-1.5*IQR

Upper_limit=q3-1.5*IQR

Models used

Random Forest

The problems that involve classification and regression may be solved with this supervised learning approach. It works on the theory of collaborative learning, which involves mixing many classifiers to address challenging issues and enhance model performance. It takes a minimum amount of time for training in comparison with different algorithms. The system functions well and offers accurate outcome predictions despite the large dataset. When a sizable chunk of data is missing, it can nevertheless be accurate. It is found in the Sklearn library.

Streps involved in Random Forest

1.The model is trained using "entropy" criteria. It helps with random feature selection. Its value lies between 0 and 1.0-pure splitting, which means that all data belong to the one class.1-impure spilit, i.e. occurrences from different classes

$$E = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

P_i is the probability of randomly selecting samples.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:08/August-2023 Impact Factor- 7.868 www.irjmets.com

2. It uses 600 decision trees with a maximum depth of 10 sub-nodes per tree. 0.9 percent of features are used while splitting. The minimum of 10 samples is needed to split. Random state set to 42 to reproduce the identical set of samples while training and testing

3.The system is trained utilizing a function by passing x_train, which consists of ph, sulfate, chloramines, conductivity, turbility, trihalomethanes, organic carbon, and solids. The y_train contains potability.

Support Vector Machine

The working process of the SVM model:

1.The SVM algorithm uses the RBF kernel with c = 1 and gamma = 0.1 to train the model. The parameter 'c' is utilized to prevent the data from being incorrectly categorized, and 'gamma' describes the influence of a single training sample on a model.

2.The model is trained using an inbuilt function passing X_train and Y_train parameters, where X_train contains hardness, pH, sulfate, chloramines, trihalomethanes, turbidity, organic carbon a, and conductivity, and Y_train contains potability.

Hyperplane: The hyperplane, which divides datapoints in a 2-dimensional feature space, is a straight line. It creates a hyperplane in higher-dimensional spaces.

Support Vectors: These are the data values that determine the location and orientation of the hyperplane, which is crucial for identifying the decision boundary.

Kernel: The data is converted into higher dimensions, which helps separate data into different categories. In this project, the RBF kernel is used. The RBF kernel calculates the similarity between the pair of points.

The equation of the RBF kernel

$$K(X_1, X_2) = exp(-\frac{||X_1 - X_2||^2}{2})$$

Here, X_1 and X_2 indicate the two points that essential to be analyzed.

$||X_1\text{-}X_2||^2$ indicates the distance among two points.

When the K value lies between 0 and 1, where 0 indicates dissimilarity and 1 indicates similarity.

III. MODELING AND ANALYSIS

Model and Material which are used is presented in this section. Table and model should be in prescribed format.

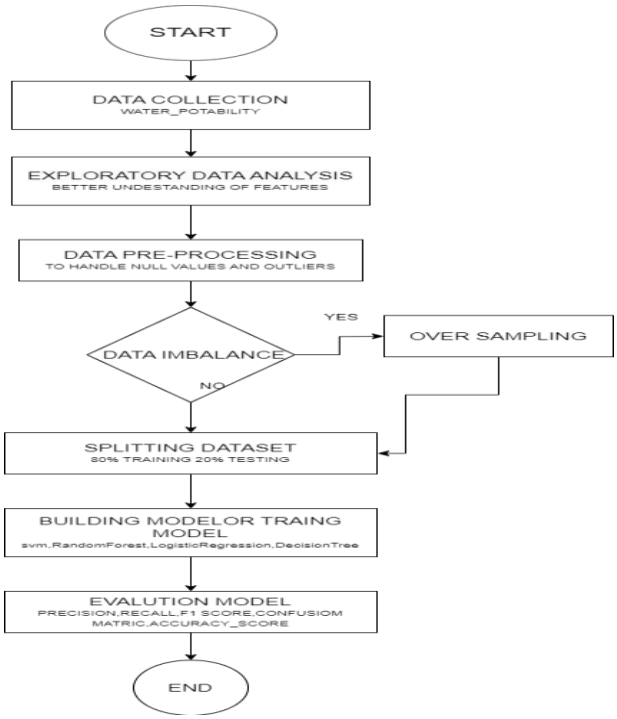

Figure 2: Flow Chart

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:05/Issue:08/August-2023 Impact Factor- 7.868 www.irjmets.com

IV. RESULTS AND DISCUSSION

The assessment metric for each model utilized in this proposed system is specified in above Table 3. SVM has the maximum accuracy of 69.05% and random forests achieves an accuracy of 69.20%.

Table 1. Models Evaluation Metrics

Random Forest	KNNImputer for handling null values.	69.20%
	IQR for removing outliers	
Support Vector Machine	KNNImputer for handling null values.	69.05%
	IQR for removing outliers	

V. CONCLUSION

In this project, the purity of water is evaluated depending on the parameters of water. The present system is trained utilizing two distinct machine learning algorithms: the SVM and random forest classifier. The behaviour of a model is examined using accuracy_score. Random forest achieved 69.20 percent compared to svm and chosen as final model. Water quality estimation can be by adopting numerous machine learning techniques.

VI. REFERENCES

[1] Yafra Khan, Chai Soo See Predicting and Analyzing Water Quality using Machine Learning: A Comprehensive Model", IEEE, 2016

[2] Kathleen Joslyn, John Lipor, " A Supervised Learning Approach to Water Quality Parameter Prediction and Fault Detection", IEEE 2018.

[3] Amir Hamzeh Haghiabi, Ali Heidar Nasrolahi and Abbas Parsaie," Water quality prediction using machine learning methods", Water Quality Research Journal,2018.

[4] Mr. Anbuchezhian, Dr. R. Venkataraman, Mrs. V. Kumuthavalli," Water Quality Analysis and Prediction using Machine Learning Algorithms", JETIR,2018.

[5] Jinal Patel, Charmi Amipara, Tariq Ahamed Ahanger , Komal Ladhva ,Rajeev Kumar Gupta, Hashem O. Alsaab , Yusuf S. Althobaiti ,and Rajnish Ratna," A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI", Hindawi,2022

[6] Heming Gao, Yuru Li, Handong Lu, Shuqi Zhu," Water Potability Analysis and Prediction", in Highlights in Science Engineering and Technology, vol 16,2022.

[7] Md. Saikat Islam Khan, Nazrul Islam, Jia Uddin, Sifatul Islam, Mostofa Kamal Nasir," Water quality prediction and classification based on principal component regression and gradient boosting classifier approach", Journal of King Saud University – Computer and Information Sciences,2022.