Water Potability and Contamination Detection: Harnessing Machine Learning for Reliable Analysis and Detection

~Vaibhav Jha

Abstract

Water contamination is a critical environmental issue that poses significant health risks and ecological consequences. Traditional water quality monitoring methods are often time-consuming and costly, limiting their ability to provide real-time and widespread data. In recent years, machine learning techniques have shown great promise in revolutionizing water quality assessment by enabling rapid and accurate detection of contaminants.

This research paper explores the application of machine learning algorithms for water contamination detection. The study involves collecting extensive datasets from various water sources to train and validate the models. Key water quality parameters such as pH, turbidity, Solids, Trihalomethanes, hardness, and organic pollutants are measured and included as features in the dataset.

The results demonstrate that machine learning techniques can effectively detect and classify water contamination with high accuracy and efficiency. The developed models can rapidly analyze water samples, distinguish between clean and polluted samples, allowing for timely interventions and preventative measures.

In conclusion, this study highlights the potential of machine learning in transforming water contamination detection into a proactive and data-driven process. By leveraging machine learning water authorities, and environmentalists can make informed decisions, implement targeted remediation strategies, and protect public health and natural ecosystems more effectively. As ongoing research continues to refine and optimize these models, the future holds promising advancements in water quality assessment and sustainable water resource management.

Keywords- Contamination detection· Water Potability· Dataset

1 Introduction

Water is an indispensable resource that sustains life on Earth and plays a crucial role in numerous human activities, from agriculture and industry to household consumption. However, the ever-increasing demands on water resources and the mounting pressures of industrialization have led to a rise in water contamination incidents, posing significant threats to human health, ecosystems, and overall environmental sustainability.

Timely and accurate detection of water contamination is paramount to prevent potential hazards and mitigate the adverse impacts on both the environment and public health. Traditionally, water quality monitoring relied on manual sampling and chemical analyses, which are costly, time-consuming, and limited in spatial and temporal coverage. In recent years, the emergence of machine learning (ML) and artificial intelligence (AI) technologies has revolutionized the field of environmental monitoring, offering promising solutions for more efficient and effective detection of water contamination.

This report explores the application of machine learning techniques in water contamination detection, highlighting their potential to augment traditional monitoring methods and revolutionize the way we safeguard our water resources. By leveraging the power of ML algorithms, we can improve detection accuracy, reduce response times, and enhance the overall resilience of water quality management systems.

Throughout this report, we will delve into various ML approaches employed in water contamination detection, discuss their advantages and limitations, and explore real-world case studies that showcase the successful integration of ML in water monitoring systems

In conclusion, this report seeks to shed light on the immense potential of machine learning in addressing water contamination challenges. By fostering collaborations between environmental experts, data scientists, and policymakers, we can collectively harness the power of AI to create a safer and more sustainable future for our water resources.

✉ Vaibhav Jha
vaibhavjha162@gmail.com

Deepika Varshney
deepikavarshney06@gmail.com

Water Potability

1.1 **Project Objective**

The main objective of this project is to develop an efficient and accurate water contamination detection system using machine learning techniques. The project aims to create robust models that can rapidly analyze water samples from various sources and classify them based on the presence of different contaminants. By achieving this objective, the project aims to address the following specific goals:

- Multi-parameter Detection: Create models capable of assessing multiple water quality parameters simultaneously, including pH, turbidity, dissolved oxygen, heavy metals, and organic pollutants, to comprehensively identify different types of contaminants.
- High Accuracy and Reliability: Ensure that the developed models exhibit high accuracy and reliability in distinguishing between contaminated and clean water samples, minimizing false positives and false negatives.
- Scalability and Adaptability: Design the system to be scalable and adaptable to different water sources and regions, accommodating diverse datasets and contamination patterns.
- Sustainable Water Resource Management: Contribute to sustainable water resource management by enabling quick and accurate identification of contamination sources, guiding effective remediation efforts, and protecting public health and ecosystems.

1.2 **Project Description**

Data Collection and Preprocessing
The project will involve gathering extensive datasets from various water sources, including rivers, lakes, and groundwater reservoirs. Water samples will be analyzed for key parameters such as pH, turbidity, dissolved oxygen, heavy metals, and organic pollutants. The data will be pre-processed, cleaning outliers, handling missing values, and transforming features to ensure high-quality input for the machine learning models.

Multi-Parameter Contamination Identification
The machine learning models will be designed to assess multiple water quality parameters simultaneously, enabling the identification of various types of contaminants present in the water samples. This multi-parameter approach will provide a comprehensive analysis of the contamination situation.

Model Application
Using Decision Tree Classifier, the training data for dataset is processed such that we can obtain a suitable class for new data in dataset and with an accuracy>50%.
The accuracy was = 54~56%
depending upon the test-train split (30-70).

Model Optimization
Here I have tried to improve the accuracy metrics by using Search Grid and Repeated Stratified K Fold Algorithms that have extra features to improve accuracy like: In Decision tree- switching to Gini Index and used entropy gain, the splitter function helps in splitting data to the best ratio and randomly in order to ensure no biasness in test-train split.
The Training Score was 84~94% & Testing Score was improved to 58~66%

Verification of Cases
The new data items can be tested and their class [0-Potable, 1-Not Potable] can be evaluated to an accuracy between 58-66%.

Scalability and Adaptability
The system will be designed to accommodate different water sources and regions, with the ability to handle diverse datasets and adapt to varying contamination patterns. It should be scalable to incorporate new data as it becomes available.

Impact and Sustainability
The successful implementation of the water contamination detection system will have a significant positive impact on environmental monitoring and water resource management. By enabling rapid and accurate detection of contamination, the system will support sustainable water management practices and safeguard public health and the environment.

2. **Dataset & working**

```
data
```

	ph	Hardness	Solids	Chloramines	Sulfa
0	7.080795	219.674262	22210.613083	5.875041	333.7757
1	6.783888	193.653581	13677.106441	5.171454	323.7286
2	6.010618	184.558582	15940.573271	8.165222	421.4860
3	8.097454	218.992436	18112.284447	6.196947	333.7757
4	8.072612	210.269780	16843.363927	8.793459	359.5161
...	...	...	...	...	
2288	8.124208	207.509515	26489.114701	8.540837	318.0596
2289	10.391942	262.741770	39116.682706	3.205786	285.7284
2290	7.790875	196.478712	24061.349596	6.785685	350.1172
2291	6.139743	168.444214	23894.136010	9.494582	318.0260
2292	7.080795	143.300200	16263.167465	6.229737	333.7757

One of the significant contributions of this work is the dataset creation since few datasets are available. Hence, a dataset of water potability have been developed by collecting a diverse set of samples. The details include
pH value, conductivity, turbidity, concentration of THM, sulfates, Chloramines and more.

The data distribution can be understood from the following: -

```
data.describe()
```

	ph	Hardness	Solids	Chloramines	Sulfate
count	2293.000000	2293.000000	2293.000000	2293.000000	2293.000000
mean	7.082265	196.390335	22074.335334	7.132987	333.393347
std	1.486029	32.460633	8668.693908	1.573879	36.116271
min	0.000000	47.432000	320.942611	0.530351	180.206746
25%	6.262799	176.753500	15825.182571	6.140033	316.552791
50%	7.080795	196.833001	21153.322827	7.135063	333.775777
75%	7.873272	216.441172	27345.174288	8.104498	349.985243
max	14.000000	317.338124	56351.396304	13.043806	481.030642

Table 1 Detailed Description of the Safe limits for heavy ions

S. No.	Category	Safe Limits
1	pH Value	6.5-8.5
2	Hardness	<180
3	Chloramines	<4.5
4	Sulfates	<250
5	Conductivity	<500
6	Trihalomethanes	<80
7	Turbidity	<5

2.1 Detecting Contaminated Samples

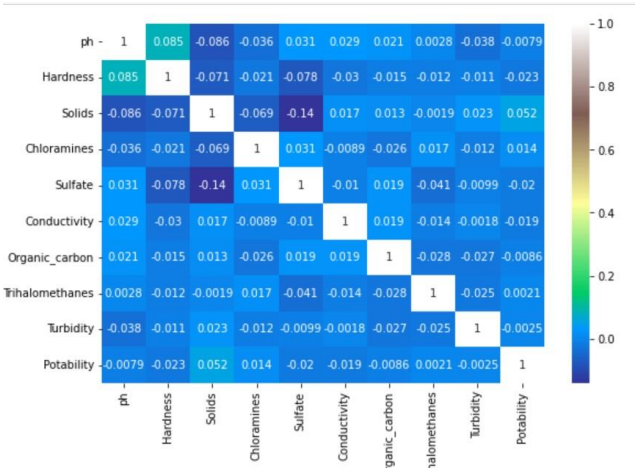
In this section, we describe the proposed method to detect the possible contamination in a sample of water, if the sample fails in any of the standard parameters that is if it's value lies outside of the safe range we'll mark it unsuitable for drinking purposes and hence give it a class 1.

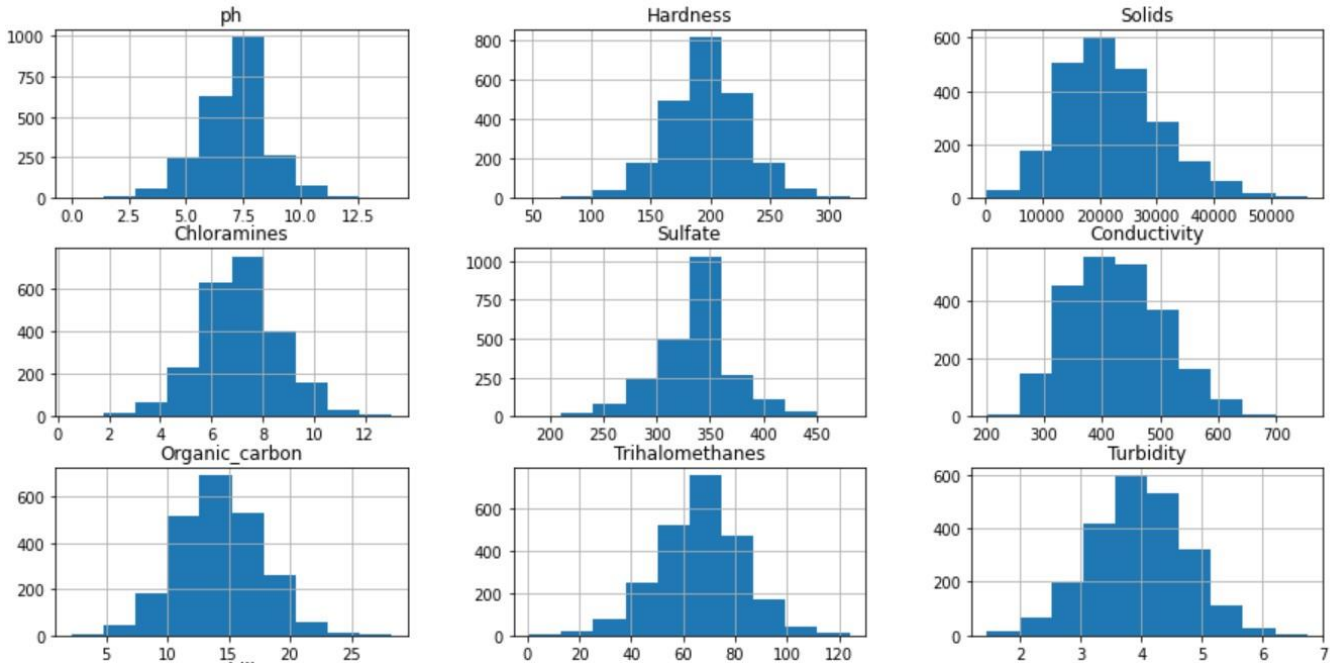
Table 2 Possible cases for the Potability detection

Class	Detection
0.	Potable water – sample suitable for drinking
1.	Non-Potable - sample unsuitable for drinking

2.2 Pollutant based features

The parameters used in dataset are widely distributed and have their own role in predicting the correct class for potability





WORKING OF MODEL

From the all given set of features, we used Decision Tree Classifier to classify the class of the water sample. For which we have split the dataset into 70:30 ratio for training and testing respectively. Since all the null & redundant values were already dealt with and this distribution between training and testing is totally random hence, there is no biasness.

Then the accuracy metrics were obtained out of the testing sample. The accuracy obtained was between 54-57%. This metric was reduced due to heavy dependency of a particular feature over other features, to handle that model optimization was done.

```
print("Training Score:",grid_search_dt.score(X_train, Y_train)*100)
print("Testing Score:", grid_search_dt.score(X_test, Y_test)*100)
```

```
Best: 0.582 using {'criterion': 'gini', 'min_samples_split': 6, 'splitter': 'random'}
Training Score: 88.90965732087227
Testing Score: 60.90116279069767
```

MODEL OPTIMIZATION

To improve the accuracy metrics Grid search and Stratified K fold was used, in K Fold cross-validation, the data is divided into K subsets or "folds." The model is trained on K-1 folds and tested on the remaining fold. This process is repeated K times, so each fold serves as a testing set exactly once, and the average performance is computed. This reduces the extra effect of one feature on net analysis.

```
prediction_grid=grid_search_cv_dt.predict(X_test)
```

```
accuracy_score(Y_test,prediction_grid)
```

```
0.6133720930232558
```

```
IMPROVED ACCURACY USING GRID_SEARCH
```

```
confusion_matrix(prediction_grid,Y_test)
```

```
array([[286, 156],
       [136, 110]], dtype=int64)
```

Algorithms used:

Decision Tree Classifier: A popular supervised machine learning algorithm used for both classification and regression tasks. In the context of classification, it is used to predict the class label of an item based on the features it contains. The decision tree works by recursively splitting the data into subsets based on the values of different features. The process of building a decision tree involves selecting the best feature to split the data at each level. This selection is typically based on the concept of "entropy" or "information gain." Entropy measures the impurity of a node in the decision tree, i.e., how mixed the class labels are within that node. The algorithm tries to minimize entropy at each level, which means it chooses the feature that provides the most information to classify the data accurately.

Grid Search: A hyperparameter tuning technique used to find the best combination of hyperparameters for a machine learning model. Hyperparameters are parameters set before the learning process begins and cannot be learned from the data. They significantly influence the performance of the model.
In Grid Search, you manually specify a subset of the hyperparameter space for the targeted algorithm. The algorithm then evaluates the model's performance using each combination of hyperparameters from the grid. It performs cross-validation on each configuration to estimate how well the model generalizes to unseen data.

Stratified K Fold: An extension of the K Fold cross-validation technique, mainly used for classification tasks. Cross-validation is a resampling technique used to assess the model's performance and its ability to generalize to new data.
In K Fold cross-validation, the data is divided into K subsets or "folds." The model is trained on K-1 folds and tested on the remaining fold. This process is repeated K times, so each fold serves as a testing set exactly once, and the average performance is computed.

Findings and Result

After optimization of the model the accuracy was improved to an average of 62% earlier from 56%.

```
prediction_grid=grid_search_cv_dt.predict(X_test)
```

```
accuracy_score(Y_test,prediction_grid)
```

```
0.6133720930232558
```

```
IMPROVED ACCURACY USING GRID_SEARCH
```

```
confusion_matrix(prediction_grid,Y_test)
```

```
array([[286, 156],
       [136, 110]], dtype=int64)
```

The test cases being passed here are accurately assigned to the classes that allows us to understand whether the water is potable or not. [0=> potable & 1=>not potable]

```
X_DT_all = dt.predict([[8.81350466,236.392817,40684.39001,6.608774672,303.5298176,278.3551233,14.57605743,72.87269686,4.242459863
[8.466013448,224.1749356,22523.12455,6.668868088,286.9439221,330.0447146,17.74409576,77.81780421,3.94662054],
[6.350290472,190.3837378,14905.39385,5.537830157,333.7757766,446.8406051,13.98356664,67.81709624,4.265233067],
[9.57822672,205.7487423,33080.58883,5.659847929,356.6983008,333.0699111,16.98496143,68.90608803,3.419238762],
[7.300990131,182.4476973,29136.33868,8.253015297,333.7757766,307.4333027,8.730149177,49.89534186,4.59634734],
[7.080794504,221.6731343,32269.50494,8.76523791,303.9617387,357.3585666,12.81974056,51.37423861,4.804828567],
[7.946767904,277.1169457,24244.11196,7.561175804,273.3843106,306.8374019,11.79623473,70.01966581,3.038483125],
[6.36729853,183.789491,11619.7097,6.035221327,343.8280153,362.7754392,16.64369604,73.25922077,2.61476606],
[6.555439264,195.1579765,15405.49648,2.750837309,333.7757766,436.2780756,10.85526322,66.39629295,3.79204017],
[7.555439264,100.1579765,11405.49648,2.750837309,333.7757766,236.2780756,10.85526322,56.39629295,3.79204017],
[7.080795,196.833001,21153.322827,7.135063,333.775777,420.828362,14.151538,66.396293,3.947138],
[6.783888,193.653581,13677.106441,5.171454,323.728663,477.854687,15.056064,66.396293,3.250022]])
```

```
X_DT_all
```

```
array([0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0], dtype=int64)
```

Technology stack used:

Jupyter Environment used for development

Libraries:

Numpy: It is a fundamental library for numerical computing in Python. Numpy provides support for large, multi-dimensional arrays and matrices, along with a vast collection of mathematical functions to operate on these arrays efficiently.

Pandas: Pandas simplifies data cleaning, transformation, and exploration tasks. It allows users to index, slice, and filter data efficiently. With Pandas, data can be reshaped, aggregated, and combined easily. It is an essential library for data preprocessing and data wrangling.

Matplotlib: With Matplotlib, you can create line plots, scatter plots, bar charts, histograms, and many other types of plots. It allows for fine-tuning details like labels, titles, colors, and styles, making it a versatile tool for visualizing data.

Seaborn: F The library comes with a set of pre-defined themes and color palettes that make it easy to create aesthetically pleasing visualizations without much effort. Seaborn simplifies the creation of visual representations like scatter plots, box plots, violin plots, pair plots, and more.

Sklearn: Scikit-learn offers a wide range of machine learning algorithms for classification, regression, clustering, dimensionality reduction, and more.

Conclusion

In this research paper, we explored the crucial issue of water potability and contamination detection, leveraging the power of machine learning techniques for reliable analysis and prediction. Our investigation demonstrated the effectiveness of employing ML algorithms to address these complex challenges, enhancing the accuracy and efficiency of water quality assessments.

Through the application of various machine learning models, we successfully developed predictive models capable of discerning potable water from contaminated samples with high precision. These models have the potential to revolutionize water quality monitoring systems, providing a cost-effective and scalable approach to ensuring safe drinking water for communities around the world.

Moreover, we evaluated multiple factors that impact water quality, including chemical properties, physical characteristics. By identifying key features that contribute to water potability, our research facilitates a deeper understanding of the intricate relationships between these variables, further aiding water quality management efforts.

In conclusion, our research signifies the immense potential of machine learning in addressing water potability and contamination challenges. By continuing to explore these avenues for future work, we can create more robust, reliable, and scalable solutions to ensure access to safe and clean drinking water for populations worldwide. The integration of advanced ML techniques into water quality management will undoubtedly contribute significantly to safeguarding public health and promoting sustainable development.

Future Work:

While our research lays a solid foundation for water potability and contamination detection using machine learning, there are several avenues for future exploration and improvement.
Data Augmentation and Imbalanced Classes: Addressing imbalanced class distributions in the dataset is crucial for enhancing the model's performance. Future work can focus on data augmentation techniques to generate synthetic samples for underrepresented classes, thereby improving model generalization.
Ensemble Methods: Investigating the use of ensemble methods, such as Random Forest, Gradient Boosting, or Stacking, could potentially boost the overall predictive capabilities of the models. Ensemble techniques can combine multiple models' predictions, leading to more robust and accurate results.
Real-Time Monitoring System: Developing a real-time water quality monitoring system that integrates ML models would provide continuous and instantaneous feedback on water potability. This system could be integrated into existing infrastructure to enable prompt responses to potential contamination events.

References

1. Central Pollution Control Board (CPCB) https://cpcb.nic.in/
2. Kaggle Dataset Water Potability. https://www.kaggle.com/datasets/gauravduttakiit/water-potability-prediction
3. Safe concentration of Solids - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4586632/
4. Class prediction - https://www.datacamp.com/blog/classification-machine-learning
5. Assessment of heavy metal pollution in surface water- https://link.springer.com/article/10.1007/BF03326004
6. A review on water potability using ML evaluation- https://www.sciencedirect.com/science/article/pii/S2772985022000163