Logo

Description automatically generated with medium confidence

A picture containing text

Description automatically generated

**Interm Report**

drinking water potability prediction

DAB 322: CAPSTONE PROJECT

Icon

Description automatically generated with medium confidence

**GROUP NO:-08**

**GROUP MEMBERS:-**

**Navneet kaur (0792027)**

**Navdeep kaur (0795879)**

**Harmanpreet kaur sumal (0795660)**

**Manpreet kaur (0794006)**

**CONTENTS**

[**INTRODUCTION 2**](#_Toc38048878)

**Objective**

[**DESCRIPTION ABOUT THE DATASET 4**](#_Toc38048879)

**Material and Methods**

[**MODEL IMPLEMENTATION 20**](#_Toc38048892)

**EDA (EXPLORATORY DATA ANALYSIS)**

[**DECISION TREE 24**](#_Toc38048896)

**RANDOM FOREST CLASSIFIER**

**HYPER-PARAMETER TUNING**

[**RELATED WORK 6**](#_Toc38048880)

**Results**

**Discussion**

**Conclusion**

[**REFERENCES 29**](#_Toc38048899)

**Introduction:-**

Water quality analysis is a complex topic due to the different factors that influence it. This concept is inextricably linked to the various purposes for which water is used. Different needs necessitate different standards. There is a lot of study being done on water quality prediction. Water quality is normally determined by a set of physical and chemical parameters that are closely related to the water's intended usage. Water that meets the predetermined parameters for a specific application is considered appropriate for that application. If the water does not fulfil these requirements, it must be treated before it may be used.

Water is a very essential resource for the survival of every living being in the world. According to research, we can survive without food for 3 weeks, but we can’t last 3 days without water. Although water is essential for us to live, all the water in the world is not drinkable for example, the ocean. If our drinking water contains dissolved salts in high concentration it may cause severe side effects, kidney problems, or even lead to death, if consumption is long term. Water in the different regions has different properties. In fact, the consequences of polluted drinking water are so dangerous and can badly affect health, the environment, and infrastructures. As per the United Nations (UN) report, about 1.5 million people die each year because of contaminated water-driven diseases. In developing countries, it is announced that 80% of health problems are caused by contaminated water. Five million deaths and 2.5 billion illnesses are reported annually . According to the studies, approximately 66% of the Earth is made up of water with the availability of fresh or usable water being only 1%, while the rest of the water is saline or salt water. Water is an integral part of the prosperity and wealth of a nation. However, the level of water has been falling considerably during the last few decades, which is one of the emerging problems in the modern world.

Although people are facing many health issues of contaminated water in all over the world but here we are going to discuss about India. India is one of the those country which facing lots of health issues because water is not safe to drink.The World Health Organization estimates that 97 million Indians, second only to China, do not have access to safe drinking water. As a result, the World Bank estimates that unclean water is responsible for 21% of infectious diseases in India**.** So for that let’s make an ML program that can predict whether the water is potable or not. In this dataset, there are certain parameters like pH, hardness, sulfates which can be taught to our machine to predict the water sample is safe to drink or not. The availability of a safe and sufficient quantity of drinking water is a crucial part of basic healthcare since drinking water quality has a significant effect on the health of people. The level of components in potable water must not threaten consumer health or reduce its usefulness . The following characteristics of good water quality should be met by water.

* Free of harmful organisms.
* Clean and clear (low turbidity).
* Lack of saline.
* Devoid of substances that provide an unpleasant flavor or smell.
* Devoid of substances that might have harmful effects on human health.
* Low levels of substances such as lead that are immediately hazardous or have negative long-term effects.
* Free of chemicals that could damage the water supply system or taint washed-in clothing.

**Objective:-**

The main objective of this work is**to measure water quality using machine learning algorithms.** A Water Quality Index (WQI) is a numeric expression used to evaluate the quality of a given water body. In this paper the following water quality parameters were used to evaluate the overall water quality in terms of the WQI. These parameters were as pH, conductivity, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic\_carbon, Trihalomethanes, Turbidity. These parameters are used as feature vector to represent the water quality. In paper two kinds of classification algorithms, namely Decision Tree(DT), Random Forest(RF) were employed to predict the water quality class.

Water quality protection or improvement is essential to support ecosystems, human health. We are going to compare out results with one of previous research. Although, we used only two machine learning algorithms in our work but we are going to improve the accuracy as compared to that research. Here is the link of that research which is related to our dataset and topic.

https://www.kaggle.com/code/zeynepozisil/potability-of-water-project

<https://www.kaggle.com/code/zeynepozisil/potability-of-water-project>

**Description about Dataset:-**

The dataset used in this study is collected from certain historical locations in India. The dataset has 9 significant parameters, namely, pH, conductivity, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic\_carbon, Trihalomethanes, Turbidity. We have 3276 rows and 10 columns in our dataset after removing null values the number of rows are 2011 and columns is 10.Data was collected by the Indian government to ensure the quality of the supplied drinking water. The target class is potable. It has two values, ‘0’ or ‘1’, where ‘0’ denotes that the water is not safe for drinking and ‘1’ denotes that it is safe for drinking. These are the features in out dataset.

Graphical user interface, application

Description automatically generated

The parameter in the dataset defines as below:-

* **PH value:-**  PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status. WHO has recommended a maximum permissible limit of pH from 6.5 to 8.5. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards.
* **Hardness:-** Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.
* **Solids (Total dissolved solids - TDS):-** Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals produced an unwanted taste and diluted color in the appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized. The Desired limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which is prescribed for drinking purposes.
* **Chloramines:-**Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.
* **Sulfate:** **-**Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. The principal commercial use of sulfate is in the chemical industry. Sulfate concentration in seawater is about 2,700 milligrams per liter (mg/L). It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.
* **Conductivity:** **-**Pure water is not a good conductor of electric current rather it's a good insulator. Increase in ions concentration enhances the electrical conductivity of water. Generally, the amount of dissolved solids in water determines the electrical conductivity. Electrical conductivity (EC) actually measures the ionic process of a solution that enables it to transmit current. According to WHO standards, EC value should not exceed 400 μS/cm.
* **Organic\_carbon:** **-**Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to the US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is used for treatment.
* **Trihalomethanes:-**THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water.
* **Turbidity:** **-**The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.
* **Potability:-** Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable. (0) Water is not safe to drink and (1) Water is safe to drink.

**Talking about the ethics(5C’s : Consent , Clarity, Consequences ,Control , Consistency)** in data collection.

**Control:-** As it is clear that the data is collected from Kaggle, it is open so it will not hamper the Control of ethics.

**Consent** part has also verified since the data is free from bias and equal consent has been taken that the data will be used for the purpose of research only.

**Consequences** , as far as this is concerned, the results are clear and there will no manipulation in the data and the results will be carried out for the purpose of finding insights only.

**Clarity** has been explained clearly since the data is transparent in nature this will not hamper the confidentiality concerns.

**Consistency** , on the other hand with each result , they will be similar in the nature and this will not deviate the results in any cost. All the aspects of ethics have been taken into this during this capstone project and it has been made sure that that whatsoever the results will come , it will be helpful for the betterment of the analysis that has been carried out previously.

Material and Methods:-

**EDA(EXPLORATORY DATA ANALYSIS)**

EDA helps us to understand the nature of the data we are working with and identify patterns and relationships that may exist between variables. The purpose of EDA is to describe the main features of the data, identify any unusual patterns or outliers, and highlight any missing or inconsistent values.

The main techniques used in EDA include data visualization, descriptive statistics, and hypothesis testing. Data visualization techniques can include scatter plots, histograms, box plots, and heat maps, among others, to help us visualize the distribution of data and relationships between variables Descriptive statistics can be used to summarize the central tendency, spread, and shape of the data, and to identify any unusual values or outliers. Hypothesis testing can be used to determine whether there is a statistically significant relationship between variables or whether any observed patterns are due to chance.

EDA is an important step in the data analysis process as it can help to inform subsequent modeling and analysis steps. By gaining a better understanding of the data we are working with, we can make more informed decisions about which modeling techniques to use, how to preprocess the data, and how to interpret the results.

Table

Description automatically generated

**Data Preprocessing:-** The processing phase is very important in data analysis to improve the data quality. In this phase, the WQI has been calculated from the most significant parameters of the dataset. Since the dataset contains missing values, data preprocessing is needed to deal with this problem. Then, water samples have been classified on the basis of the WQI values. For obtaining superior accuracy. Data preprocessing is an important step to obtain better performance from the models. In this step, the unnecessary or redundant data are removed from the dataset. These data have no meaning for the machine learning models. Preprocessing helps to enhance the efficacy of the learning models. Not only does preprocessing help to enhance the performance of the model, it also helps to reduce the computational time. In this research, during the data preprocessing, we came to learn that there are several missing values in the dataset.

**Distribution of features:-**

The histogram distribution of the nine features used for training the machine learning models. The tenth attribute is the target class with potable and not potable values. The histogram helps to understand the distribution of each feature in the dataset. It shows how frequently a value/feature appears in a relatively unbiased way. The given features PH have a normal distribution and they are not skewed. Moreover, the distribution is unimodal and symmetric. The given range of these features is different, and the occurrence of each feature is helpful to determine the center of a particular feature.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

**Distribution of Potability:-** It has two values, ‘0’ or ‘1’, where ‘0’ denotes that the water is not safe for drinking and ‘1’ denotes that it is safe for drinking.

Chart, bar chart

Description automatically generated

**Correlations between features:-** We used correlation analysis to find possible correlations between all the features in order to find the dependent features using commonly obtainable features. A correlation matrix is a table that displays the correlation coefficients for different characteristics. In a table, the matrix represents all possible value pairs. It is also good for spotting and displaying trends. the correlation plot, it is clear that there is not any linear correlation for a feature with the target or within the features themselves

Graphical user interface

Description automatically generated with medium confidence

**DATA PROCESSING - SPLITTING THE DATA INTO TRAIN AND TEST SETS:-**

Before we teach our machine we need to process the data. Here we split the data into data inputs (X has the properties of water) and data outputs (y show whether the water is consumable or not). This is our data structure:-

Graphical user interface, text, application

Description automatically generated

**Decision Tree:-** DT is a renowned machine learning algorithm extensively utilized for regression and classification problems. The selection of the root node at each level is a problem in the decision tree. Decision Tree Classifier model on the X\_train and y\_train datasets and makes predictions on the X\_test dataset. It then calculates the accuracy score of the model on the test dataset and prints it out using f-strings. It also classification report of the model's performance on the test dataset. The accuracy score of the model on the test dataset is 60%. The classification report provides precision, recall, F1-score, and support for each class in the dataset.

**The accuracy of the Test set is : 60.0%**

**Model Evaluation**

1. **Classification Report**

Table

Description automatically generated

Precision is the percentage of the result that are relevant . Recall is the percentage of the total relevant results correctly classified by the algorithm. and if the F scores reaches best value i.e. 1 , it means perfect precision and recall.

For instance , from the above classification report it can be concluded that precision here is 60% which means our model will have 60% relevant results whereas , 70% of recall means it will generate 70% relevant results out of the total classified values in the case of non-potability water(0). Same is the case with potability water(1).

Also, after running the confusion matrix code , following output was produced.

1. **Confusion Matrix**

**Confusion Matrix:**

**[[ 282 118 ]**

**[ 142 114 ]]**

Fig : Confusion matrix for Random forest classifier

From the above matrix we can identify that our model performs well since the number of correct predictions(282,118) is greater than the number of incorrect prediction(142,114).

**Random forest:-** The tree-based classifier RF combines several poor apprentices (poor learners) to generate very accurate predictions. To train different decision trees utilizing diverse bootstrap samples, RF uses bootstrap bagging . A bootstrap sample is generated using the sub-sampling of the training dataset, where the size of the training and test sample dataset is the same. Similar to other ensemble classifiers, RF uses decision trees for making predictions. At each stage, the identification of the root node is a challenging task for the development of decision trees. Random Forest Classifier model on the X\_train and y\_train datasets and makes predictions on the X\_test dataset. It then calculates the accuracy score of the model's performance on the test dataset. The classification report provides precision, recall, F1-score, and support for each class in the dataset. Compared to the Decision Tree Classifier, Random Forest Classifier shows better accuracy.

**The accuracy of the Test set is : 66.0%**

**Model Evaluation**

1. **Classification Report**

Table

Description automatically generated

Precision is the percentage of the result that are relevant . Recall is the percentage of the total relevant results correctly classified by the algorithm. and if the F scores reaches best value i.e. 1 , it means perfect precision and recall.

For instance , from the above classification report it can be concluded that precision here is 66% which means our model will have 66% relevant results whereas , 88% of recall means it will generate 88% relevant results out of the total classified values in the case of non-potability water(0). Same is the case with potability water(1).

Also, after running the confusion matrix code , following output was produced.

1. **Confusion Matrix**

**Confusion Matrix:**

**[[ 350 50 ]**

**[ 176 80 ]]**

Fig : Confusion matrix for Random forest classifier

From the above matrix we can identify that our model performs well since the number of correct predictions(350,50) is greater than the number of incorrect prediction(176,80).

**Hyperparameter tuning**

Hyperparameter tuning refers to the process of selecting the best combination of hyperparameters for a machine learning model to achieve optimal performance on a given task. Hyperparameters are adjustable parameters that control the behavior of a machine learning algorithm, such as the number of hidden layers in a neural network, the learning rate of the optimizer, the number of trees in a random forest, or the regularization parameter in linear regression.

Hyperparameter tuning involves searching for the best combination of hyperparameter by testing various values using a systematic approach such as grid search, random search or Bayesian optimization. The goal is to find the combination of hyperparameter that maximizes the performance of the model on the validation set or minimizes a given metric such as mean squared error or accuracy.

Hyperparameter tuning is a crucial step in machine learning, as it can significantly improve the performance of a model and prevent overfitting or underfitting. It is often an iterative process that requires domain expertise and a good understanding of the dataset and the algorithm.

random\_grid, which contains all possible combinations of hyperparameters. It can be passed as an argument to a function that performs hyperparameter tuning, such as GridSearchCV or RandomizedSearchCV, to search for the best set of hyperparameters for a random forest classifier model.

The best set of hyperparameter found by the randomized search, as well as the corresponding cross-validation score. These hyperparameters can be used to build a final random forest classifier model that hopefully performs better than the default model.

**Accuracy of random forest after using hyper parameter tuning:-**

Table

Description automatically generated with medium confidence

**Related work:-**

Although research has shown that various machine learning models, including deep neural networks, kernel models, fuzzy logic, genetic programming, neuro-inference models, and others, have been utilized to design surface water quality phenomena . There are still a number of new classifiers that have not yet been investigated. For the conservation of the water environment, water quality prediction is very important. Authors developed a water quality assessment approach based on long short-term memory (LSTM) and IGRA, taking into account the multivariate correlation and temporal sequence of the water quality data . The first suggestion made by IGRA was to choose features that have a higher absolute correlation being predicted. Second, an LSTM-based prediction model was created, with the indicators collected by IGRA serving as its inputs. Results show promising output for water quality prediction.

Traditional water quality prediction approaches used machine learning and statistical characteristics (normal distribution) and techniques and has not achieved good results. Contrarily, artificial intelligence-based approaches have shown better results as there is no need to determine the relation of dependent variables with independent ones . The authors applied a neural network model to determine the quality of groundwater and provide an improved water quality monitoring system for drinking purposes. Authors investigated quality indicators for potability by applying explainable artificial intelligence . Many researchers have concentrated their efforts on using many variables as a function model after the realization of the significance of monitoring and forecasting the changing water quality. Artificial neural networks (ANNs) such as MLP and radial bias have been employed by researchers for water quality prediction and have achieved satisfactory results . A water quality index was also produced using an ANN and five significant and widely accepted water quality indicators . The literature includes studies on using artificial intelligence systems to predict the water quality index .

The recurrent neural network (RNN) is the most widely used deep learning model for the analysis of time-series data. An improved RNN with a significant capacity for information acquisition and archival is LSTM, which has been used extensively for predicting water quality . In order to perform a thorough predictive study of the water quality in the next time period, the authors developed a prediction technique based on the bidirectional LSTM that takes into account the reliance at many time scales .

For accurate real-time water quality prediction, the researchers introduced a novel feature selection and classification approach in . The complexity of the suggested approach is decreased by using a learning-based model and quantum teaching to choose the best possible collection of characteristics. The authors proposed two tree-based hybrid models, namely, XGBoost and RF, to provide more precise short-term water quality prediction and they also introduced a novel data denoising technique (CEEMDAN) .

Numerous versions of models have been utilized by researchers to determine water quality; they are still facing challenges in the process. The literature review indicates that there is a limited number of studies using deep neural networks in predicting water quality, especially for drinking purposes.

**Results :-** After applying some ML algorithms we got the accuracy as given below:-

|  |  |
| --- | --- |
| Model Name | Accuracy |
| Random Forest | 0.6036585365853658 |
| Decision Tree | 0.6554878048780488 |
| Best Score using Randomized search | 0.6694656488549617 |
| Best Score using Grid search | 0.6744274809160304 |
| Accuracy from Random Forest after hyper parameter tuning | 0.6570121951219512 |

## **Discussion:-** The result from this experiment is performed on the water quality dataset and the result indicated that the the accuracy of the random forest is better than decision tree. On executing these algorithms on the prescribed dataset, the accuracy level of these algorithm is 60%(decision tree) and 66%(random forest). AS the future scope, we desire to examine more additional attributes along with different classification techniques on the water quality rate dataset.

**Conclusion:-**  
 Water quality relates to everyone’s life. The adequacy of water resources not only affects people's life safety, but also deeply affects the development and stability of society. Also, the government of each region should predict the sustainability of their water resources. Water used by humans should be guaranteed to be enough and safe. Different standards apply to different uses of water. When it comes to drinking water, the standards need to be particularly strict. Not every water resource meets these standards. In addition to the special substances contained in some specific type of water, there are nine main factors that affect the potability of water, including, PH value, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic carbon, Trihalomethanes and Turbidity. According to the result of the research, the features of water are hardly related to each other. There are no general rules between each distribution of the factors. For example, the range of the non-potable distributions of PH value and organic are larger than the range of the potable distributions, while for the other factors except Trihalomethanes (the distribution of potable and non-potable is almost the same), the range of potable distribution is larger than the non-potable distributions. Qualification of one feature does not increase the possibility of the qualification of another feature. Only if every factor meets the potable standard, the water can be drunk. By using different machine learning models, the authors got a model in the end which can help to predict whether a water resource is potable. Although there are some other model like logistic regression model and so on but in that time we the accuracy of our model on random forest classifiers and discission tree by comparing past research and the authors got a model in the end which can help to predict whether a water resource is potable. As a result, the accuracy of these model is improved as compared with previous result but The model still needs improvement.

**Refrences:-**

* [https://downloads.hindawi.com/journals/cin/2022/9283293.pdf](%20https:/downloads.hindawi.com/journals/cin/2022/9283293.pdf)
* <https://www.researchgate.net/publication/365495813_Water_Potability_Analysis_and_Prediction>
* <https://europepmc.org/article/PMC/PMC9514946>
* [https://www.researchgate.net/publication/363721702\_A\_Machine\_Learning-Based\_Water\_Potability\_Prediction\_Model\_by\_Using\_Synthetic\_Minority\_Oversampling\_Technique\_and\_Explainable\_AI](https://www.researchgate.net/publication/363721702_A_Machine_Learning-)
* <https://www.kaggle.com/code/zeynepozisil/potability-of-water-project>