**DRinking water potability predection**

CAPSTONE PROJECT DAB 422

Group No.08

Group members: -

Name: Navneet Kaur

E-mail id: - sukhdeepsinghkahlon96@gmail.com

Name: Navdeep Kaur

E-mail id: -navdeepsaggu93@gmail.com

Name: Harmanpreet Kaur Sumal

E-mail id: - sumalharman543@gmail.com

Name: Manpreet Kaur

E-mail id: - [manpreetk01061999@gmail.com](mailto:manpreetk01061999@gmail.com)

Table of Contents

[Abstract 3](#_Toc141534839)

[Introduction 3](#_Toc141534840)

[Description about Dataset 4](#_Toc141534841)

[Methods of Analysis 5](#_Toc141534842)

[Analysis with Machine learning techniques 8](#_Toc141534843)

[Results 21](#_Toc141534844)

[Discussion 22](#_Toc141534845)

[Conclusion 23](#_Toc141534846)

[Future Work 23](#_Toc141534847)

[Acknowledgements 23](#_Toc141534848)

[References 24](#_Toc141534849)

[Appendices 24](#_Toc141534850)

Abstract

Water is one of the largest resources on earth. People need water to sustain life, including drinking water. It is important to know whether drinking water - human life resource - is enough for everyone now and in the future. However, water resources are not evenly distributed everywhere on the planet. While the water resource is rich in some countries and regions, it is not enough for some other regions. The analysis of different region’s water resources should be done individually. In this project, we are going to analyze the potability of water by using an Indian water potability dataset from Kaggle. Protecting and caring for water is one of the most critical environmental problems today. This research aims to design an intelligent system using machine learning models to improve water quality and predict whether it is safe to be used as drinking water. Several models of machine learning algorithms are compared to find the best model to be used for the accuracy of prediction of water quality. In this project, we compare K-Nearest Neighbour, Logistic Regression, Random Forest, and XGB models to get the best model for water potability prediction.  According to the research, the features of water are not related to each other. All the features should meet a specific standard to get potable water.

# Introduction

Water is utilized for drinking, household usage, food production, or leisure, safe and readily available water is critical for public health. Improving supplies of water, and improved management of water resources, might help countries thrive and reduce poverty. There are many reasons why water is deteriorating because in our India there are many industrial areas so the release of pollutants in rivers is the main reason for water deteriorating. There are many other reasons for water deteriorating like people’s garbage (plastics), the unwanted things in rivers, their nearest ponds, lakes, and in sea, and due to plastics and unwanted garbage, there are some toxic occurrences. So, for all these reasons, water is deteriorating nowadays. Contaminated water and inadequate sanitation have been related to diseases such as typhoid, dysentery, polio, cholera, hepatitis, and diarrhoea. People are exposed to preventable health dangers due to a lack of, inadequate, or poorly managed water and sanitation facilities. It is especially the case in health facilities, at which water shortage, hygiene, and cleanliness assistance exposes staff and patients to viruses and bacteria. Globally, 15% of people get a virus throughout a stay in the hospital, only with numbers becoming very higher at lower areas. The choice of drinkable water must be decided with great care. Many domain-acknowledgements are required to address this challenge. In this case, this system gets built in that manner to comprehend as a supply of data as much as feasible while retaining generality. In India, however, industrial and home pollutants have contaminated 70% of accessible water.

Approximately 80% of the local population and 20% of the urban population do not have access to clean drinking water. Three-quarters of the nation’s children’s health issues are infectious diseases and environmental factors, mainly water supply and sanitation. Diarrhoea is responsible for 46% of mortality in children under the age of five, with water-related disorders accounting for a large amount of this. According to Ethiopia’s Ministry of Health, 6000 children die each day from diarrhoea and dehydration.

Description about Dataset

The dataset used in this project is collected from certain historical locations in India from Kaggle. This dataset has total of 3276 samples were collected and analyzed for 8 important hydro-chemical parameters, which are pH value, D.O(dissolved oxygen), temperature, B.O.D. (Biochemical Oxygen Demand), conductivity, Nitrate nan N, Fecal coliform, and Total coliform. We have 1991 rows and 12 columns in our dataset.

We select this dataset because people are facing many health issues because of contaminated water in India as compared to other countries and we can take Canadian dataset to predict the water quality but in Canada it is not concern able issue. Water is safe to drink in Canada. Especially, this dataset includes the locations and years so by comparing locations we can know that where water is more contaminated and according to years we can know that water is going to pollute or not as year increases. In capstone project part one we Predict the quality of water only using the parameters, we did not include the locations and years so without locations and years we cannot specify that where water is not safe to drink and what is the reason behind the contaminated water.

In this project, we are going to analyze data by using tableau. So that we can present visualization of data within different graphs

The parameter in the dataset defines as below: -

1. **PH value: -** PH is an important parameter in evaluating the acid–base balance of water. It is also an indicator of acidic or alkaline condition of water status. Who has recommended a maximum permissible limit of pH from 6.5 to 8.5. current investigation ranges were 6.52–6.83 which are in the range of WHO standards.
2. **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 number of dissolved solids in water determines the electrical conductivity. Electrical conductivity (EC) 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.
3. **D.O: -**Dissolved oxygen refers to the amount of oxygen gas dissolved in water. It is essential for aquatic life and is an indicator of water quality. Adequate dissolved oxygen levels are necessary to support healthy ecosystems.
4. **Temperature: -** It includes Palatability, viscosity, solubility, odors, and chemical reactions are influenced by temperature. Thereby, the sedimentation and chlorination processes and biological oxygen demand (BOD) are temperature dependent. It also affects the biosorption process of the dissolved heavy metals in water. Most people find water at temperatures of 10-15°C most potable.
5. **Biochemical oxygen demand (BOD): -** BOD is a measure of the amount of oxygen consumed by microorganisms while decomposing organic matter in water. It quantifies the level of organic pollutants and indicates the water's capacity to support aerobic life. High BOD levels can deplete dissolved oxygen, leading to harmful effects on aquatic organisms.
6. **Total Coliform: -**Total coliforms include bacteria that are found in the soil, in water that has been influenced by surface water, and in human or animal waste.
7. **Fecal Coliform: -** Fecal coliforms are the group of the total coliforms that are present specifically in the gut and faces of warm-blooded animals. Because the origins of fecal coliforms are more specific than the origins of the more general total coliform group of bacteria, fecal coliforms are considered a more accurate indication of animal or human waste than the total coliforms.
8. **Nitrate nan N: -**Nitrate (NO3−) is well-known environmental pollutant that not only arises naturally, but also is released by several anthropogenic exercises. These anthropogenic activities include the manufacture and use of nitrate fertilizers, fossil fuel combustion (occur as atmospheric deposition, hereinafter AD) and releases of both domestic and industrial sewage systems and modification in nitrogen-fixing crops in natural vegetation.

Methods of Analysis

In this project, the usage Tableau for visualization of data and the ML models to predicting the quality of water after applying different models so that we can find the best models to predict the quality of water whether it is potable or not. In related work, they are using ARIMA model to predict the water quality, but we are not familiar with this model. So, we are not going to use this technique. We will apply Random Forest, KNN, Logistic Regression and XGB ML models to predict the quality of water and compare the accuracy which model is providing highest accuracy and best fit to predict the quality of water.

**Average value of total coliform in the water according to states**

A map of india with different states

Description automatically generated

Figure . According to the above map chart, Punjab and Odisha have greater total coliform in their water, meaning there are more bacteria in those two states, which has a significant impact on the potability of their water.

**Average of total coliform with respect to years**

A graph with a line

Description automatically generated

Figure . This line graph is showing the average of total coliform according to years. This line graph shows that the value of total coliform remains constant from 2002 to 2011 but after that it started increases and reach to the peak point that was 2,171,564 in the end of year.

**Dashboard**

A screenshot of a graph

Description automatically generated

Figure 3. The above dashboard represents the analyses in different graph forms.

## Analysis with Machine learning techniques

**Correlation between the features**

Correlation analysis is to find possible correlations between all the features to find the dependent features using commonly obtainable features. A correlation matrix is a table that displays the correlation coefficients for different characteristics.

**A screenshot of a computer

Description automatically generated**

Figure . In this plot, fecal and total coliform is highly correlated with each other so here we can remove fecal coliform because it will be included in total coliform.

**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. Like 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. After applying random forest on our dataset, we get the 100% accuracy for training data and 90% for testing data and the accuracy of out of bagging score is 91%. Here, our model is providing 100% accuracy on training dataset that means our model is going to overfit.

**Accuracy of Random Forest after hyper parameter tuning**

Before hyperparameter tuning we get the 100% accuracy of training data it doesn’t mean our model is performing perfectly on unseen data. Sometimes it indicates overfitting which means the model memorized the training data instead of learning the underlying patterns and relationship. So, to prevent overfitting, we did hyperparameter tuning and after that we get the 98% accuracy on training data and 90% for test data.

A graph of a number of different colored bars

Description automatically generated

Figure .In this scenario model is predicting non-drinkable, good and poor water quality classes with high precision, recall and F1 score it indicates that the model is not able to differentiate between these classes effectively.

**Feature importance with Random Forest**

Feature importance is necessary because it helps in understanding the relative significance or contribution of different features in predicting the target variable or outcome of a machine learning model.

**A graph of blue rectangular bars

Description automatically generated**

Figure 6. It can be observed from the graph the B.O. D importance is on top with 0.35 which means if we neglect this feature than it put bad impact on the overall accuracy of the model. So B.O.D. parameter has more influence on our model during predicting the water quality as compared to other parameters.

**Logistic Regression**

Logistic regression is a statistical algorithm used to model and predict binary or categorical outcomes. In the case of predicting the quality of water, logistic regression can be applied to determine whether the water is drinkable or non-drinkable based on a set of input features. The training accuracy is 75% and testing is 74% with logistic regression.

A graph of a number of percentages

Description automatically generated

Figure .In the above bar graph, on the high precision, recall and F1 score our logistic regression model is predicting non-drinkable class.

**Confusion Matrix**

**A screenshot of a graph

Description automatically generated**

Figure . The rows of the confusion matrix represent the actual classes, and the columns represent the predicted classes and describe about the misclassify classes. In this case, we have five classes: Excellent, Good, Poor, Very Poor, and non-Drinkable.

**KNN**

K-nearest neighbours is a supervised machine learning algorithm used for both classification and regression tasks. We applied KNN on our dataset to predict the quality of water. We get the accuracy 62% on training data and 56% on testing data. In this case, the 62% accuracy on the training set indicates that the model is capturing some patterns in the training data, but it may not be able to generalize well to new, unseen water quality samples. The accuracy of 56% on the testing set suggests that the model's performance is limited, and it is struggling to make accurate predictions on unseen water quality samples.

A graph of percentages of poor and poor

Description automatically generated

Figure 9.The above graph is based on high precision, recall and F1 score KNN model is predicting the quality of water is good and poor, it indicates that the model is not able to differentiate between these classes effectively.

**Confusion Matrix**

A blue squares with black numbers

Description automatically generated

**XGBBoost**

XGBoost is a powerful machine learning algorithm that belongs to the family of gradient boosting methods. It is widely used for classification and regression tasks due to its effectiveness and high performance. After applying XGB on our dataset we get the 100% accuracy on training dataset and 92% on test data. There is also overfitting in our dataset. To prevent this, we do hyperparameter tuning so that our model doesn’t learn about underlying the patterns and relationship. After hyperparameter tuning we get the 99% on training and 91% on testing data. The decrease in accuracy on the training data from 100% to 99% suggests that the model has become slightly less prone to overfitting. It is fitting the training data slightly less closely, which can potentially help it perform better on new, unseen data.

A graph of a number of students

Description automatically generated

Figure 10. In this scenario XGBoost model is predicting non-drinkable, good and poor water quality classes with high precision, recall and F1 score ,it indicates that the model is not able to differentiate between these classes effectively .This model is also overfitting we did hyperparameter tuning but it’s still not performing well.

**Confusion Matrix for XGBoost**

**A blue squares with white text

Description automatically generated**

**Confusion Matrix for XGBoost after hyperparameter tuning**

**A screenshot of a graph

Description automatically generated**

**Flow Process**

Project Discussion and Select topic for Project.

Collection of Data

Read the Articles and identify the problem.

Develop a Plan for Analysis

Clean and Prepare Data for Analysis

Feature Selection

Visualization of Data with Tableau

Prepared Midinterim Report and Presentation



Model Development

Model Evaluation

Optimization and Reinforcement

Deployment and Integration

Continue updating and Monitoring.

Evaluating the performance of the Models

Finish the Model

Results

After applying some ML algorithms, we got the accuracy as given below:

|  |  |  |
| --- | --- | --- |
| **Models Name** | **Accuracy on training data** | **Accuracy on testing data** |
| Logistic Regression | 0.75 | 0.74 |
| KNN Classifier | 0.62 | 0.56 |
| Random Forest | 0.98 | 0.90 |
| XGBoost | 0.99 | 0.91 |

Discussion

The result from this experiment is performed on the water quality dataset and the result indicated that the accuracy of the XGBoost and random forest is better with or without hyperparameter tuning than other two models. In the prior research there was no location, and the features were different where we got the accuracy only 65.70% but now its increased by almost 33%. Now, we face a problem of overfitting in two of our models (XGBoost and Random Forest) and for preventing this we used hyperparameter tuning but still they perform as same. So, here we can say that logistic regression is performing well as compared to others as it performed well on both training and testing .

A graph of a test

Description automatically generated

# Conclusion

Predicting usable water is crucial for protecting the ecosystem and avoiding pollution. To sustain optimum public health, clean drinking water must be made available. The potability of the water can be ensured by drinking it from reliable sources. Accurately forecasting drinking water becomes challenging. To avoid making predictions that are incorrect, the optimum learning algorithm is required. We used different machine models to precisely forecast and categorise the Water Quality Index (WQI) using a dataset made up of 3276 samples from different sources of rivers and lakes in India The potability of drinking water can be predicted using an intelligent model based on four distinct machine learning algorithms named Logistic Regression, KNN classifier, Random Forest and XGBoost using 8 common criteria where XGBoost and Random Forest perform superior to others . Thus, our model will assist in lowering the amount of poor-quality water that people consume, hence reducing the spread of terrible illnesses like typhoid and diarrhoea. This method allows for focused actions for environmental and public health management, which holds potential for more effective water quality monitoring, particularly in areas with limited resources. The suggested machine learning models aids the preservation of water resources and the general health of ecosystems by improving monitoring efforts and promoting sustainable water management practises.

# Future Work

After applying all the models on the selected dataset to predict the quality of water we got the highest accuracy with random forest and XGBoost but these models are overfitting and to prevent this we did hyperparameter tuning but we did not got much difference in the accuracy and also with precision ,recall and F1 score these models are predicting non drinkable ,good and poor quality classes which means these models are not able to differentiate between these classes effectively. This might be happening due to imbalanced data, feature relevance, model complexity and so on. To overcome this on can do feature engineering, model selection like SVM ,neural network and others, error analysis.

# Acknowledgements

We began working on our project to forecast the potability of drinking water in May and finished it in July. We started by gathering the India dataset before looking for related research. We choose to use analysis in Tableau and create various machine learning models after reading all relevant material. Finally, we discovered the potability of water using random forest, XGBoost, KNN and logistic regression where logistic regression was well performed model.

References

[Water Quality Assessment in Terms of Water Quality Index](http://pubs.sciepub.com/ajwr/1/3/3/index.html)

[Modelling and Prediction of Water Quality by Using Artificial Intelligence](https://www.mdpi.com/2071-1050/13/8/4259/pdf)

# Appendices

<https://github.com/HarmanSumal/Capstone>