**Drinking Water Potability Prediction**

**Group members:- Group no.08**

**Navneet kaur (0792027)**

**Navdeep kaur (0795879)**

**Harmanpreet kaur sumal (0795660)**

**Manpreet kaur (0794006)**

* **Content:-**
* Summary/Abstract
* Introduction
* Methodology
* Data Analysis
* Findings and Interpretations
* Recommendations
* Conclusion
* Appendix
* References
* **Abstract:-**

Predicting drinking water quality involves collecting comprehensive data on water sources and potential contaminants, analyzing historical records to identify trends and patterns, utilizing monitoring systems to measure key parameters in real-time, employing predictive modeling techniques to forecast future water quality based on historical data, conducting source water assessments to understand potential risks, performing risk assessments to evaluate the overall health risk, ensuring regulatory compliance with standards and guidelines, and implementing preventive measures and interventions to deliver safe drinking water to the public.

* **Introduction:-**

Every living thing in the world needs access to water to survive. Research shows that while we can last three weeks without food, we cannot go three days without water. Studies show that around 66% of the Earth is made up of water, with only 1% of that water being fresh or useful and the rest being saline or salt water. Even though water is necessary for our survival, some water, like ocean water, cannot be consumed. Even while contaminated water is causing health problems for individuals all throughout the world, but this article has focus on India. India is one of the nations that has many health problems since the water is unsafe to consume.

* According to a United Nations (UN) statistic, over 1.5 million individuals every year pass away and polluted water is to blame for 80% of health issues in underdeveloped nations.
* Each year, 2.5 billion illnesses and 5 million fatalities are reported. A nation's prosperity and wealth are largely dependent on its access to water. However, one of the new issues in the modern world is the significant decline in water levels during the past few decades.
* According to the World Health Organisation, 97 million Indians, second only to Chinese citizens, lack access to clean drinking water. The World Bank calculates that 21% of infectious infections in India are caused by filthy water.
* **Objective:-**
* The main objective of the project is predicting the quality of water.
* Project also aims a reliable, affordable, consumer oriented and sustainable potable water supply and sanitation services.
* The system will help reduce people’s consumption of poor-quality water and consequently curtail horrific diseases such as typhoid and diarrhea.
* To understand what constitutes safe, potable water and distinguish between potable and non-potable water by applying machine learning techniques.
* In this case, our application can improve water pollution in different water bodies.
* Due to health issues, we want to ensure that water is clear, lacks salt, free of odor and flavor ,free of chemicals and so on.
* **Methodology(**Material and Methods)

**Description about Dataset: -**

The dataset used in this project is collected from Kaggle in India. We have 3276 different Water bodies and 899 rows and 12 columns in our dataset . 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.

The parameter in the dataset defines as below: -

* **pH:** pH is a measure of the acidity or alkalinity of water. It quantifies the concentration of hydrogen ions in the water and is measured on a scale from 0 to 14. A pH value below 7 indicates acidity, a value of 7 is neutral, and a value above 7 indicates alkalinity.
* **Total coliform:** Total coliform is a group of bacteria that are commonly found in the environment, including soil, vegetation, and animal feces. The presence of total coliform bacteria in water is an indicator of potential contamination from sources like sewage or fecal matter.
* **Temperature:** Temperature refers to the measure of how hot or cold the water is. It can influence the solubility of chemicals, the growth of microorganisms, and the overall water quality.
* **Dissolved oxygen:** 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.
* **BOD (Biochemical Oxygen Demand):** BOD is a measure of the amount of oxygen consumed by microorganisms while decomposing organic matter in water. High BOD levels can deplete dissolved oxygen, leading to harmful effects on aquatic organisms.
* **Conductivity:** Conductivity is often used as a proxy for assessing water salinity and overall water quality.

A screenshot of a computer

Description automatically generated with low confidence

**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. 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.

**A screenshot of a computer code

Description automatically generated with medium confidence**

**Data Preprocessing:**

* Preprocessing the data is a crucial step in getting the models to perform better. The dataset is cleaned up in this step by removing any extraneous or redundant data.
* For the machine learning models, this data is meaningless. Preprocessing aids in boosting the learning models' effectiveness. Preprocessing not only helps to improve the model's performance but also to speed up computation. Several missing values were discovered in the dataset for this study during the data preprocessing phase.

**Tools:-**

* Jupyter notebook using Python for creating the ML models, correlation, and comparing the accuracy.
* Tableau for Visualization.
* **Data Analysis:-**

The data has been analyzed using Tableau and Dashboard given below:-

A screenshot of a computer

Description automatically generated with medium confidence

In this dashboard, there is a combination of Line chart, Bar, Map and Mixture of Bar chart and Line chart.

* **Bar and Line Chart:-** There were fluctuations in the average value of Conductivity and B.O.D according to years.
* **Line Chart:-** It shows the average values of Total Coliform increased as reach to peak point(2,171,564 )with the passes of year.
* **Bar Chart:-** It depicts the highest and lowest average values of Conductivity according to states Which was 47,156 and 885 in Triveni Sangam and GHAGGAR Punjab respectively.
* **Map Graph:-** Map graph is showing the average value of Total Coliform in the Water according to states. By help of this map we can understand Water is more contaminated in Punjab and Odisha.
* **Findings and Interpretations:-** Modelling and the prediction of water quality have played a pivotal and significant role in saving time and consumption in lab analysis. There are various studies that used machine learning models for modelling and predicting WQ. Ahmed et al. applied the FFNN model to predict WQI, and 25 parameters were used as input data. Gazza et al. Applied machine learning to predict WQI, and 23 input parameters were considered. Saki Zadeh employed 16 parameters. Rankovic et al. proposed an artificial intelligence model to predict WQ using 10 input parameters. Umair Ahmed et al. Various machine learning models for WQI and WQC, and four parameters were used as input data. It is noted that the polynomial regression model is good for predicting WQI, whereas the multi-layer perceptron (MLP) model is suitable for classifying WQC(Water Quality Certification).

There are some limitations of Developed models, it can be difficult to use and expensive to predict water quality. Therefore, the management or agencies has to develop effective strategy which can remove the contamination resource and also need of alternative source to meet the community demand.

* **Recommendations:-**As we visualized in Tableau there is more Contaminated water in India so, modelling and predicting water quality using AI algorithms is very important for the protection of the environment. The artificial intelligence models were developed to predict and classify water quality for drinking by employing data from rivers collected from different locations in Indian states.

**Further Process: -** We completed the visualization in Tableau until today. In further step, we are going to work in Python notebook using ML techniques(Logistic Regression and KNN) to predict the water quality and improve the accuracy.

A diagram of data processing

Description automatically generated with medium confidence

* **Conclusion:-** 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. After Visualization in Tableau, we come to know that how much the average value of each feature in different states according to year. After getting tableau information we can represent that where we need to pay more attention to improve the quality of water. 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. According to the result of the research, the features of water are hardly related to each other. 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. After Visualization in Tableau, we come to know that how much the average value of each feature in different states according to year. After getting information tableau we can represent that where we need to pay more attention to improve the quality of water.
* **Appendix** :- <https://github.com/HarmanSumal/Capstone>
* **References:-**

* <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>