Water Potability

AGENDA

Abstract

Introduction

Methodology

Results

Conclusion

ABSTRACT

Water is a crucial and indispensable resource for sustaining human life, and maintaining its quality is of utmost importance for the well-being of individuals. When drinking water becomes contaminated, it poses severe health risks, including diseases like diarrhea, cholera, and various other waterborne ailments. As a result, ensuring safe and clean water becomes crucial to promote public health. Recent findings indicate that a significant number of approximately 3,575,000 people lose their lives each year due to water-related illnesses. Therefore, accurate prediction of water potability has the potential to substantially reduce the incidence of such diseases. Notably, machine learning algorithms have emerged as powerful tools for effectively predicting water quality, enabling timely and precise monitoring of water resources. This Project focuses on multiple algorithms to forecast water potability based on the physicochemical properties of water samples obtained from the Drinking Water dataset. This dataset comprises nine distinct parameters, namely pH, hardness, solids, chloramines,sulfates, trihalomethanes, organic carbon, conductivity, and turbidity. By employing various algorithms, such as Random Forest, SVC, XGBoost, DecisionTree and KNN, we aim to determine the potability of drinking water. Notably, the SVC algorithm demonstrates superior performance compared to traditional ML models,achieving an accuracy of 0.69, precision of 0.94, Recall of 0.69, and F1 score of 0.80.Additionally, the SVC algorithm also performs well, yielding an accuracy of 69%.Consequently, this Project holds significant promise in providing reliable water quality data to Projecters, water management personnel, and policymakers, thereby enhancing the effectiveness of water potability monitoring.

INTRODUCTION

Water, an indispensable resource for all life forms on our planet, holds immense importance in the realms of economy, ecology, and human well-being. The provision of safe and uncontaminated drinking water is of paramount significance in safeguarding good health and preventing waterborne illnesses. The existence of harmful bacteria, viruses, parasites, and chemicals in contaminated water can give rise to a range of diseases and infections,including diarrhea, dysentery, typhoid, polio, cholera, and hepatitis A. Shockingly, the WHO has estimated that approximately 485,000 individuals succumb to diarrhea-related complications caused by consuming contaminated drinking water on an annual basis. Furthermore, polluted water sources can also contribute to the development of chronic health conditions, such as cancer and developmental disorders.

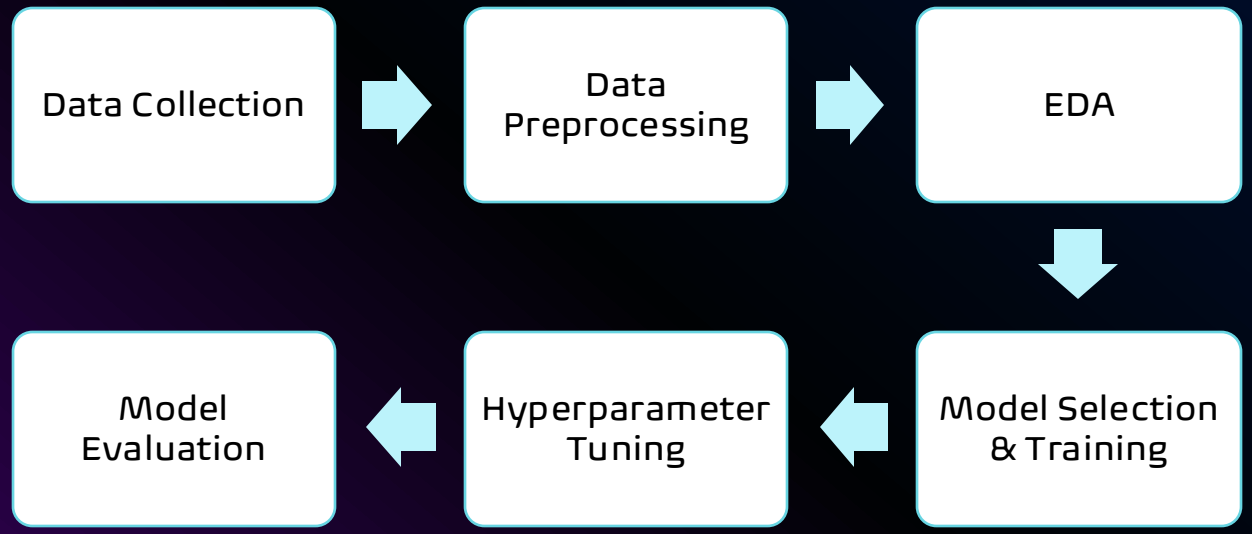
In the context of India, a distressing report has brought to light that, each year number of people suffering from waterborne diseases is nearly 37.7 million including children [4]. Disturbingly, domestic and industrial pollutants have contaminated approximately 70% of the available water, leaving nearly 80% and 20% of the rural and urban populations respectively without safe drinking water. Moreover, the global community is grappling with significant challenges pertaining to water scarcity and the deteriorating quality of water, adversely affecting millions of individuals worldwide. Alarming data from World Health Organization for the year 2018 report says that people consuming fecal matter contaminated water is about 2 billion. Hence, in order to achieve sustainable development, promote a healthy existence, and eradicate poverty, ensuring universal access to clean and safe water is of utmost importance.

In regions where water treatment facilities are inadequate, such as developing nations and rural areas, the ability to predict water potability is of utmost importance. Traditional methods of monitoring water quality, which involve costly and time-consuming laboratory and statistical investigations, have proven to be inefficient. Therefore, there is a pressing need for a more efficient and cost-effective alternative. This Project aims to propose and assess the viability of an approach based on machine learning for accurately predicting water potability in real-time. In last few years, machine learning algorithms have made significant advancements in predicting water quality, enabling more precise and efficient monitoring. Various classification techniques, including SVC, RF, DT and KNN can be employed to predict the potability of water.
For this Project, along with the above algorithms, XGBoost is also applied to the Drinking Water Dataset.

The primary focus given here is on predicting the potability of drinking water based on its physicochemical characteristics. The Project seeks to develop a model that can provide accurate and timely data on the quality of drinking water, enabling policymakers and water resource managers to implement preventive measures and ensure the availability of safe drinking water for the general public. Also, the Project aims to compare the performance and accuracy of various algorithms used for predicting water quality.

METHODOLOGY

The task of ensuring water potability is complex, as it involves numerous physical, chemical, and biological factors that impact the quality of drinking water. Machine learning techniques have emerged as valuable tools for forecasting water quality and assessing water potability. This project introduces a strategy that harnesses ML models for water potability prediction. To develop a more accurate model for predicting water potability wisely, thereby facilitating efficient water management and ensuring the availability of clean drinking water within communities is a primary objective of this project. The flow of this project is described in Figure. The proposed methodology encompasses data collection, preprocessing, handling missing values and outliers, data normalization, model construction, performance evaluation, and model optimization through hyperparameter tuning.



DATA COLLECTION

The project utilized available dataset on institution portal as the primary source of data. This dataset comprises 3276 observations of water quality collected from different locations and includes nine distinct **Parameters Unit Standards** physicochemical parameters, namely pH, hardness, solids, chloramines, sulfates, trihalomethanes, organic carbon, conductivity, and turbidity, along with a target feature portability, which is used to make a prediction using various machine learning algorithms.

Drinkable Water Quality Standards

Parameters	Standards & Units
pH	6.5 - 8.5
Organic Carbon	2 mg/L
Chloramines	4 ppm
Turbidity	5 NTU
Trihalomethanes	80 µg/L
Sulfate	250 mg/L
Hardness	300 mg/L
Conductivity	500 µS/cm
Solids	1000 mg/L

The standard values for each water quality parameter recommended by WHO and the EPA are represented in Table. If the values of these parameters exceed their standard limit, then that water is not suitable for drinking.

DATA PREPROCESSING

- ✓ Data Transformation & Cleaning
- ✓ Normalization
- ✓ Detections of outliers

This is one of the crucial steps in any machine learning algorithm as it transforms raw data into a structured format suitable for analysis by machine learning algorithms. This process encompasses various essential steps that ensure the quality and relevance of the data, thereby impacting the accuracy of the resulting models.
The initial step in data preprocessing involves addressing missing values, which can introduce complexities during the training phase. One common approach is to impute missing values with the mean value of the corresponding feature. However, it is crucial to assess whether this method is appropriate for the specific dataset and problem atand. Alternative imputation techniques such as KNN or regression imputation may be more suitable in certain cases.Identifying and handling outliers is another critical aspect of data preprocessing, as they can significantly affect the outcomes of statistical analysis and machine learning algorithms.
Feature selection is a crucial component of data preprocessing, which identifies relevant and informative features from data. By reducing complexity and preventing overfitting, it enhances model performance.Another vital step in data preprocessing is feature scaling, which standardizes the features to a consistent scale.
This can improve the performance of certain algorithms, such as those based on distance or gradient descent. Standardization, min-max scaling, and robust scaling are among the commonly utilized techniques for feature Scaling.It seeps there is no outliers or noise data in dataframe.

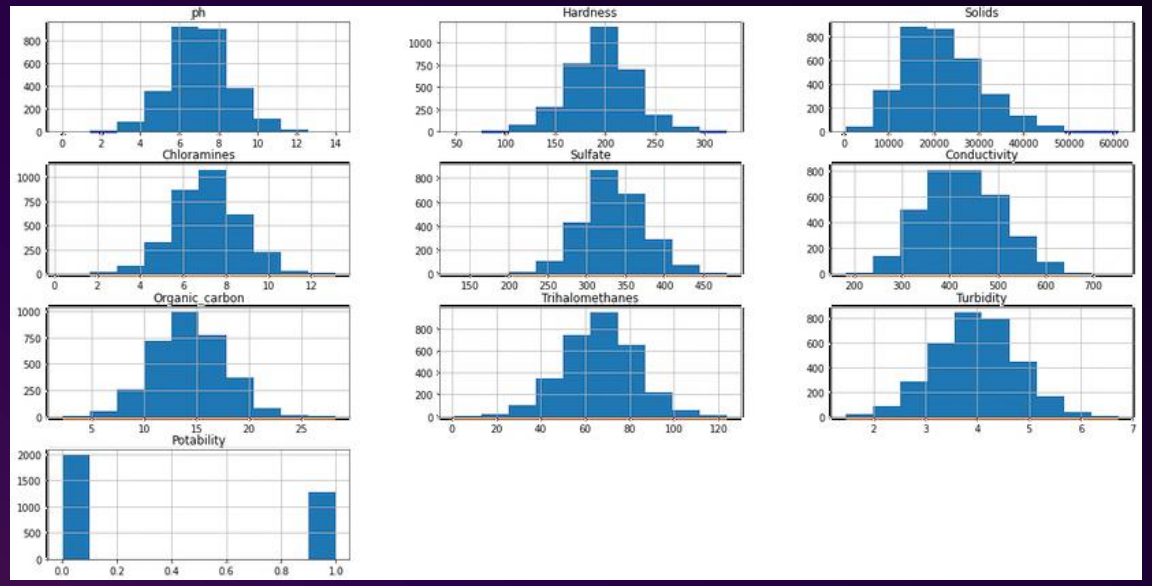
	min	mean	std	max
ph	0.000000	0.505771	0.104997	1.000000
Hardness	0.000000	0.540231	0.119263	1.000000
Solids	0.000000	0.356173	0.143968	1.000000
Chloramines	0.000000	0.529963	0.123921	1.000000
Sulfate	0.000000	0.581699	0.102669	1.000000
Conductivity	0.000000	0.427940	0.141336	1.000000
Organic_carbon	0.000000	0.463026	0.126750	1.000000
Trihalomethanes	0.000000	0.532673	0.127938	1.000000
Turbidity	0.000000	0.475853	0.147548	1.000000
Potability	0.000000	0.390110	0.487849	1.000000

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is a critical initial step to deeply understand a dataset's structure, patterns, and characteristics through visualization and statistical methods. EDA helps identify crucial elements like missing values, outliers, variable distributions, and relationships, enabling informed decisions about data cleaning, feature engineering, preprocessing, and model selection, ultimately leading to more accurate and robust ML models.

➢ Feature Distribution

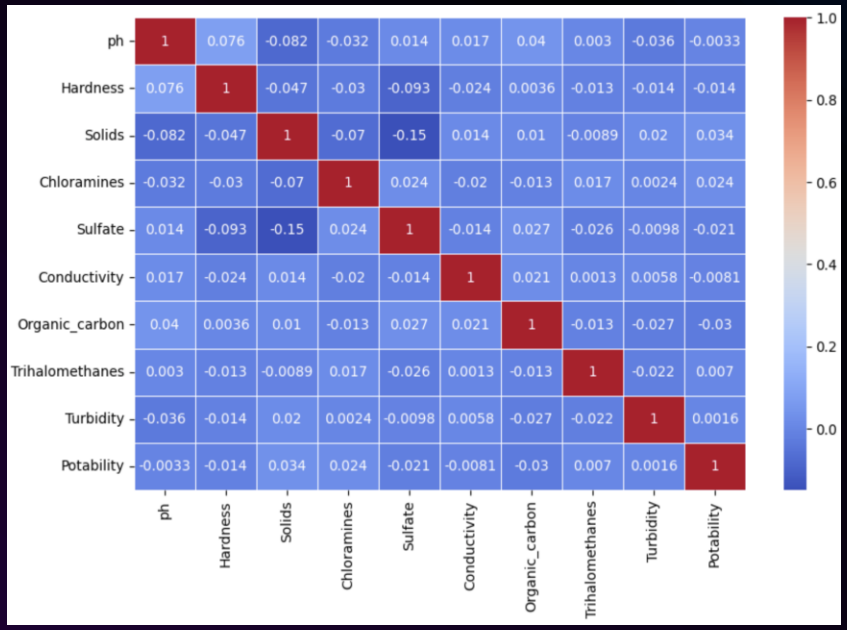
This project includes an analysis of ten different water quality parameters, and individual statistics are presented to provide context regarding the drinkable water standard. The analysis of histogram distributions reveals that Chloramines, Sulphate, pH, Trihalomethanes, Organic carbon, Hardness, and Turbidity exhibit a relatively uniform distribution pattern. Data distribution is represented in Figure. In contrast, Solids and Conductivity have a right-skewed distribution. Additionally, Potability is a binary variable, with only two possible values.



Distribution of all the water parameters

➢ Feature Engineering

Feature engineering in machine learning is the process of transforming raw data into features that better represent the underlying problem to the predictive models, resulting in improved model performance. It is a crucial step in the machine learning workflow, often considered more impactful than simply choosing a complex algorithm.The core idea is to extract or create meaningful information from the raw data that machine learning algorithms can effectively learn from.



Correlation Heatmap

Support Vector Classifier -

Support Vector Classifier (SVC) is a type of Support Vector Machine (SVM) used for classification tasks. SVM is a supervised learning model that finds the hyperplane which best separates the data points of different classes in a high-dimensional space. The main goal of SVM is to maximize the margin between the hyperplane and the nearest data points (support vectors) from any class.

Random Forest -

RF is a broadly used ML algorithm that is proficient in handling both regression and classification tasks. It falls under the category of ensemble learning, leveraging the collective predictive capabilities of multiple decision trees to improve prediction accuracy. RF is also popular due to its ability to handle high-dimensional data, eliminating the need for feature selection or dimensionality reduction techniques. As a result, it often outperforms other algorithms in various scenarios. Moreover, Random Forest exhibits robustness in the face of outliers and missing data, further contributing to its effectiveness in diverse data situations.

K-Nearest Neighbours -

The K-Nearest Neighbors (KNN) algorithm is a widely utilized and straightforward machine learning method employed in water potability prediction tasks, both for regression and classification purposes. KNN, being a non-parametric technique, stands out for its absence of assumptions about the intrinsic nature of the data. Commonly regarded as a "lazy learner" algorithm, KNN gradually learns from the training set through iterations. Instead of actively constructing a model, it stores the acquired information and subsequently applies it during the classification process. The basic principle of KNN involves predicting the classification of a new water sample by identifying the K nearest labeled instances in the training dataset and using their class labels to forecast the classification of the new instance

Decision Tree -

A decision tree in machine learning is a supervised learning algorithm used for both classification and regression tasks. It operates by creating a hierarchical, tree-like structure that models decisions and their potential consequences.

eXtreme Gradient Boosting -

XGBoost has emerged as a highly effective machine learning algorithm extensively utilized in water potability prediction. Its remarkable capability to handle extensive and intricate datasets, coupled with its ability to deliver accurate outcomes across diverse classification and regression tasks and has contributed to its widespread adoption in this field. As an ensemble learning method based on decision trees, XGBoost combines multiple DT into a model. One of the key advantages of XGBoost for water potability prediction is its ability to effectively handle missing values, allowing it to handle real-world water quality data without requiring extensive pre-processing. Additionally, XGBoost's inherent parallel processing capability enables the training of models on large water quality datasets within a reasonable timeframe.

MODEL SELECTION & TRAINING

The process of model selection plays a crucial function in the analysis. It includes selecting the optimal machine learning algorithm that suits the given dataset and problem. The following algorithms are employed in this project.The dataset was divided into 80:20 ratio of training and test data.

- Support Vector Classifier (SVC)
- Random Forest (RF)
- K-Nearest Neighbours (KNN)
- Decision Tree (DT)
- eXtreme Gradient Boosting (XGBoost)

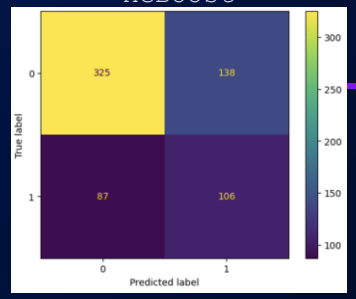
❖ Hyperparameter Tuning

Hyperparameter tuning is the process of selecting the optimal hyperparameters for a machine learning model to improve its performance. Hyperparameters are external configuration variables that are set by a data scientist before the training process begins and govern how the model learns. Unlike model parameters, which are learned from data, hyperparameters must be tuned through experimentation to find the best possible combination.we have used GridSearchCV.

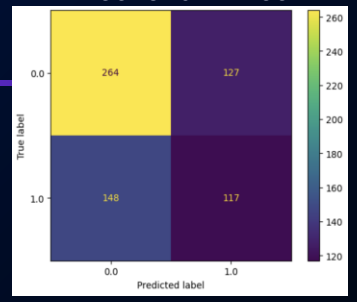
MODEL EVALUATION

While predicting water potability, evaluating the efficiency of machine learning models is also necessary. Various assessment metrics, including accuracy, F1-score, precision, and, recall are used for performance analysis.
The role of a confusion matrix is to compare the predicted class label of a data point with its actual class label,serving as a valuable tool in both binary and multi-class classification models. By providing essential evaluation metrics, including True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), the confusion matrix aids in assessing model performance.Accuracy quantifies the proportion of accurate predictions out of the total number of predictions. Precision gauges the ratio of correctly predicted positive observations to the overall predicted positive observations. Recall assesses the ratio of correctly predicted positive observations to the total number of actual positive observations.

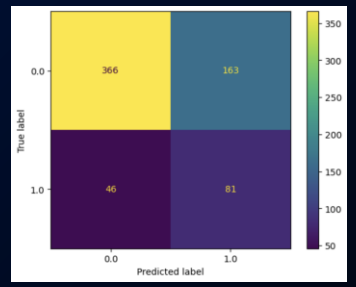
XGBoost



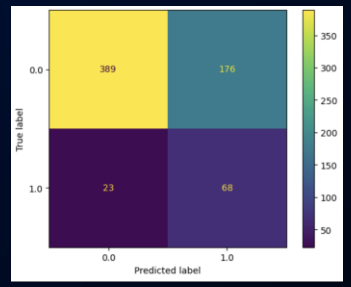
Decision Tree



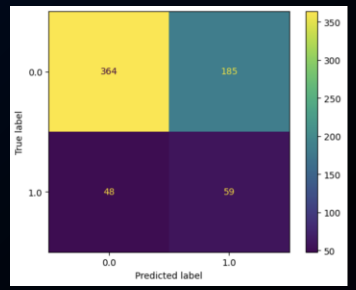
KNN



svc



Random Forest



The confusion matrix of all implemented models

Algorithm	Accuracy	Precision	Recall	F1-Score
Support Vector Classifier	0.69	0.94	0.69	0.80
Random Forest	0.68	0.88	0.70	0.78
K-Nearest Neighbours	0.64	0.88	0.66	0.76
Decision Tree	0.58	0.63	0.68	0.65
eXtreme Gradient Boosting	0.65	0.79	0.70	0.74

❖ RESULTS

For the proposed project, a dataset consisting of 3276 samples was employed. Each sample underwent analysis to determine nine specific water quality parameters: pH, Organic_carbon, Chloramines, Turbidity, Trihalomethanes, Sulphate, Hardness, Conductivity and Solids. A summary of these parameters is provided in Table 3. To facilitate the analysis, the dataset was divided into 80:20 ratio of training and test data.

Percentage of potable and non-potable water based on the dataset

Potable	Non Potable
39.01%	60.99%

This project aimed to evaluate the performance of various algorithms such as logistic regression, KNN, RF, XGBoost,DecisionTree, and SVC by employing multiple performance metrics via confusion matrix. Based on the findings, it was evident that the SVC algorithm outperformed the other algorithms, demonstrating remarkable results. It achieved an accuracy rate of 69%, precision as 0.94, and F1-score of 0.80, suggesting flawless classification performance.

But this model is not the best model. This is an average model with less accuracy.

I have done this project as per I am capable & tried but not succeeded in building the best model.

CONCLUSION

Ensuring the safety and purity of drinking water is of utmost importance to safeguard human health. Accurate prediction of water potability plays a crucial role in achieving this objective. Access to clean drinking water is a fundamental right for every individual, as it is vital for maintaining overall well-being and preventing waterborne diseases. However, the escalating global population and increasing pollution levels have raised significant concerns about the quality of water sources. Leveraging the power of machine learning techniques can greatly contribute to predicting water potability and implementing necessary measures to enhance water quality, thus ensuring the availability of safe drinking water for the population.

My project extensively explored various machine learning algorithms and their effectiveness in predicting water potability based on an extensive range of physicochemical factors. This project depicts the potential of all implemeted algorithms as valuable tools for monitoring and managing water quality, which has profound implications for both the water sector and public health.

However, it is important to acknowledge certain limitations of the study. The dataset utilized in the Project is relatively small, consisting of only 3276 observations. Consequently, it might be challenging to generalize the findings to larger populations. Additionally, the Project focused on a limited set of water quality parameters, and it is advisable for future investigation to consider other relevant factors that could influence the potability of water.

THANK YOU

Saurabh Sayaji Patil

8668654981

Srabhpatil@gmail.com