Assessment of Machine Learning Algorithms Concerning Drinking Water Potability for Enhanced Sustainability

BHARATHA SOUMYA
COMPUTER SCIENCE OF ENGINEERING (AI-ML)
SR UNIVERSITY
Hasanparthy,Telangana
2203A52007@sru.edu.in

***Abstract*—Conceptually, this paper investigates water potability as the quality of water that is fit for drinking, yet both its determinants and implications for public health and sustainability are interpreted. Through the inclusion of perspectives from multiple disciplines and the consideration of latest technology we have pursued to supplement our knowledge about the intricate interplay between water pollution and access.**

There is a need to provide adequate and safe drinking water for it is basic to not only the survival of all life forms but also the public health worldwide. On the other hand, some places right now are facing the problems with water quality due to pollution, the lack of purification system, and environmental situations. Addressing these challenges are now being done by using the new technologies such as Artificial Intelligence (AI) and Machine Learning (ML) to facilitate for the effective management of water resources. The abstract is a detailed one regarding the construction of the reusable water quality assessment and management portability dataset which is a prerequisite for building AI and ML solutions. This dataset is a combination of different types of data, including those from water quality parameters, environmental factors, geographical aspects and socio-economic indicators. Researchers, policymakers, and practitioners are able to make use of this dataset to create predictive models, spot trends, and take the right actions in terms of water resource management. Via AI and ML approaches as regression, classification, clustering and anomaly detection meaningful data can be retrieved and so help to aware of water systems more profoundly. Finally, this data will create an evidence-based decision-making process cut across disciplines, hence promoting a speedy and timely effort of clean and provision of safe water for everyone.

***Index Terms*—Water Quality Parameters, Environmental Factors, Geographical Information, Socio-Economic Indicators, Data Quality and Governance, Sustainable Water Management Water Quality Parameters, Environmental Factors, Geographical Information, Socio-Economic Indicators, Data Quality and Governance, Sustainable Water Management A**

I. Introduction

THE critical demand for fresh safe drinking water has never been as significant like it is now. Two emerging entities, AI and machine learning, are taking the helm of this war with Water portability data at the centre of the development.

The issue of water portability, or the ability of water to be used for human consumption, can be confidently ranked as one of the most pressing water problems on a global scale. AI and machine learning are more and more essential for safe drinking water efficiency by means of the use of water portability sets of data. Usually these data sets covers information about various water quality parameters as pH, hardness, organic matter contents and microbiological presence. Besides, the four features are the necessary variable which shows whether water is drinkable. Availability of clean and safe water is borne on human sustain and public health everywhere across the world. Despite this, the problem of access to safe drinking water remains a firmly rooted issue in a lot of areas and is a result of many circumstances for example the pollution, poor infrastructure, and natural calamities. Overcoming this challenge needs to have the application of advanced technologies, which consider among AI and ML, to operate the water resources and assure water quality.

The theoretical bases of portability of water resource are developed upon the theoretical multidimensional framework that covers physical, chemical, biological, and socio-economic factors. The quality of sources of water is what lies at the heart of water portability, as these are exposed to natural processes as well as human activities and standards of regulations. Elements like climate, water, land use, and pollution sources are fundamentals, they determine the quality and availability of water.

Machine learning models are trained on these datasets so to detect patterns existing between water quality numbers and portability. This enables prediction of the water potability of new, unmonitored samples. Through the understanding of the key factors controlling water quality, researchers may develop more targeted and more efficient water treatment processes. The AI models can always analyse the sensor data from the water sources, hence facilitating real-time monitoring and detecting probable contamination cases even before they occur.

Even though AI and ML may have a considerable effect on water portability assessment, the data errors accuracy, how reliant the models are and the ethical supposed must strongly be taken into consideration. Violation of a human right, privacy, and algorithmic transparency are one of the few issues that must be solved if fair and just decision is to be ensured. Hence, AI technologies are found to be the befitting tools to bring about an era like never before in water management using equitable, transparent and Sustainability

principles.

In the event, of the competence on 'the provision of truthful and complete information' exists as the life-line for a smooth road of AI and ML integration in the broad space of WUA's. These datasets cover two broad range of specialized information which include the hydro-parameters, the environmental factor and the geographical elements. Indicators of socio-economic type are the especially those that are vital in constructing the prediction models, showing trends and taking correct decisions for the management and treatment of water.

II. LITERATURE REVIEW

[1] L. Poudel, D. Shrestha, et al. this study tries to compare the logistic regression, k-nearest neighbor, random forest, and artificial neural networks algorithms on a statistical imputated water potability dataset. It founds out the random forest algorithm as the best-performing algorithm. Incomplete datasets in features like pH and chloramine showed our system the need for more efficient methods for dealing with missing data, for example, median imputation, and focused future research in this area.

[2] Jinal Patel, Charmi Amipara, Tariq Ahamed Ahanger,et al. investigate various machine learning models namely SVM Decision Tree Random Forest Gradient Boost and Ada Boost The bias and model interpretability concern is being counteracted via implementing XAI techniques such as LIME feature importance analysis.

[3] S. Patel, K. Shah, S. Vaghela, et al. (2023) The abstract reveals that ML is a new approach to predicting water potability and calculating risks of waterborne diseases to decrease the illnesses related to water consumption. It describes XGBoost and Random Forest computations to get a high performance prediction for water quality.

[4] Heming Gao 1,†,Handong Lu 3,et al. The paper analyses water potability using statistical methods like binomial distribution and K-nearest neighbour algorithm on an Indian water dataset, emphasizing the independence of water features and the need for specific standards for potable water.

[5] Surjeet Dalal, Edeh Michael Onyema, et al. This research, emphasizing the algorithms based on machine learning should be taken as a clue which indicates the necessity of developing the model for SE Georgia in such a way that these parameters as quality and safety of water should be regulated. The parameters undertakes the higher level of accuracy in the forecasting at 96.4 level of percentage.

[6] Sanaa Kaddoura, robotics played a significant role in this work since machine learning approaches were used to predict the water quality by hampering the support vector machine and k-nearest neighbor as the best models. They were picked out depending on their F1-score and ROC AUC values. It stresses the drinking water cleanliness, the economic gains of water supply investments, and how pollution affects the water quality. Our recent work with a local food bank has been nothing short of transformative. Through our partnership, we have had the privilege of witnessing firsthand the positive impact our contributions have made on the lives of our community members in need.

[7] Afaq Juna 1,†, Muhammad Umer 1,et al. In this research, a nine-layer MLP and KNN imputer are adopted to effectively and accurately predict water quality achieving 99 percentage precision score using the high accuracy effluent and the effluent that is already supervised class for reference. It is to impute missing values in datasets which results to improved classifier performance and better results than seven other machine learning algorithms.

[8] Amir Hamzeh Haghiabi, and Abbas Parsaie the study evaluates artificial intelligence models like ANN, GMDH, and SVM for predicting water quality in Tireh River, Iran, finding SVM to be the most accurate. Transfer and kernel functions like tansig and RBF were optimal for ANN and SVM, with all models showing some overestimation properties.

[9] Umair Ahmed 1, Rafia Mumtaz 1,*, Hirra Anwar 1,et al. study of water quality prediction, the authors use supervised machine learning techniques in order to get the most efficient prediction taking into account the fact that the number of input parameters is minimal (temperature, turbidity, pH and total dissolved solids). Gradient boosting and polynomial regression provide with mean absolute errors of 1.9642 and 2.7273 while multi-layer perceptron classifies water quality measures consisting of 10 parameters with the accuracy of 85.07 percentage , demonstrating the effectiveness of such methodology in place of real-time water quality detection systems.

[10] Zaky Umar 1,*, Naswin Ahmad 2, et al. is assessing the capability of Decision Tree to predict water potability, which can achieve about 54.33 percent of the overall correct estimation. Despite the fact a Decision Tree is easily interpreted, one is recommended to make use of the complex models or ensemble method to achieve a better prediction accuracy in water quality assessment.

[11] Ivan Ivanov, Borislava Toleva* et al. (2023) the paper focuses on predicting water potability using a simple machine learning algorithm, offering quick insights into water quality in different regions. It addresses the deterioration of potable water sources, emphasizing the importance of water quality prediction for environmental and ecological studies.

[12] Afaq Juna 1,†, Muhammad Umer 1,†,et al. the paper introduces a method for water quality prediction using a nine-layer multilayer perceptron (MLP) with a K-nearest neighbour (KNN) imputer to address missing values, achieving high accuracy of 0.99.

III. PROPOSED APPROACH

A. Data

The dataset utilized in this study is indicated from Kaggle. The Kaggle data, which is reputable and can be accessed by anyone, is considered one of the simplest ways to get a dataset without much troubleshooting. The data on this study is located under the name of 'Water Quality'. The dataset is of 10 columns of which the instances are 935. One can pursue a fulfilling career. Its values can assume only two states, digits '0' and '1', where '0' implies that the water is not fit for consumption and the digit '1' signifies that the water is drinkable. The table 2 that describes as detailed as possible the used dataset is presented below.

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

B. Data Collection

The dataset used in the resulting method was collected from the Kaggle's Water Quality Dataset. The parameters used in the study were hardness, anionic form, total dissolved solids, tri-halomethanes, pH, turbidity, total solids, organic carbon, and conductivity. This TABLE-1 corresponds not only to all of the characteristics.

C. Data Preprocessing

Data Processing forms the basis of the Data Analysis to process the data for better quality. Dataprocessing is explained as "the process ofamassing and arranging data parts to gain valuable information".At this stage,the WQI was computed using the core of dataset.

TABLE I
MISSING VALUES FROM THE DATASET

Feature	Missing values	Percent of Missing values
pH	491	14.99
Hardness	0	0.00
Solids	0	0.00
Chloramines	0	0.00
Sulfate	781	23.84
Conductivity	0	0.00
Organic carbon	0	0.00
Trihalomethanes	162	4.95
Turbidity	0	0.00
Potability	0	0.00

D. Data visualization

Visually, data analysis which does the nitty-gritty determine hidden links between the variables is a very crucial activity. One of the purposes of Data Visualization is to represent vital statistics in an easy method to read that can enable the readers to obtain the needed information without any confusions. Data visualization could be carried out with GUI elements like a chart, map or infographic. Visualization of data directly explains to attendees what their own data means, so that they would not forget about digits or shapes, which might seem to be incomprehensible. Like any machine, it helps the analyst to discover the patterns and exaggerate the figures that show the opposite outcomes. What we have is the dataset that is composed of two classes, with the number of non-potable instances making up about 61 Percent of all elements and the amount of potable 39 Percent. The water-quality data collected during drinking water testing (bacteria faecal coliform test) is one of the bases for drinking water safety for humans.

The histogram of the nine features of the 9 features used for the training of several machine learning models is presented here (Fig. 2). Ultimately this is the eleventh component class it includes data from both the drinkable and non-drinkable

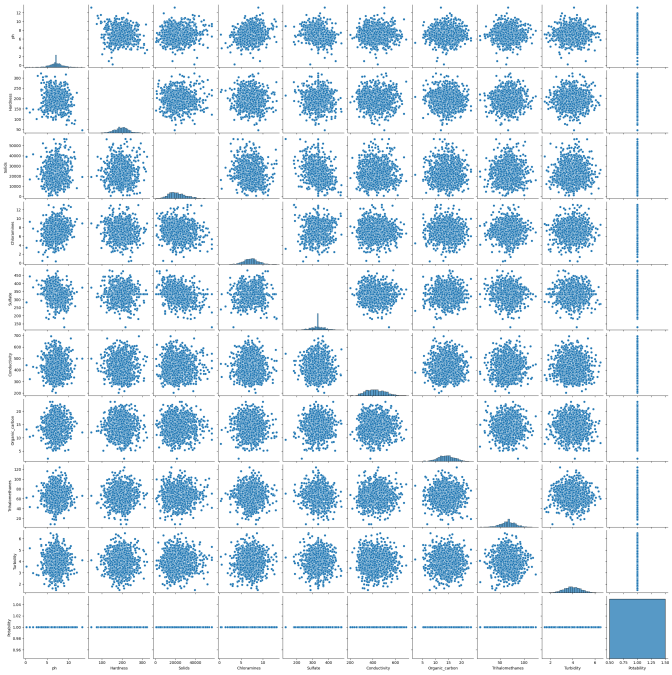

Fig. 1.

water. The histogram can be elaborated very well and by it, we can get an idea about the distribution of features of data. It shows how a feature/value holds, altogether, and rarely the overall frequency is not so fair.

As Fig.2 shows, the given features follow the normal curve and it seems to have little or no bias, skewed, or abnormal distribution. To add more, it is also skewed to one side with a bell shape. Actually, features have their range of values while in other cases they may be binary in nature, but the main thing is that by arranging cases by features one can determine the centroid for a particular feature. In this way, figure 2a demonstrates the pH values are spread almost evenly between 5.0 to 8.0.

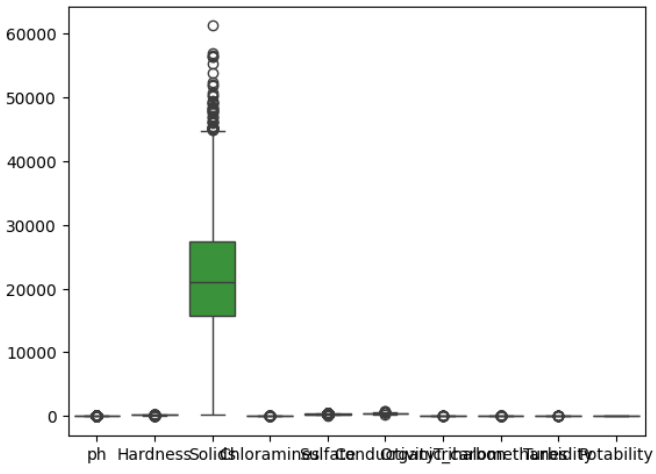

Fig. 2. Box plot

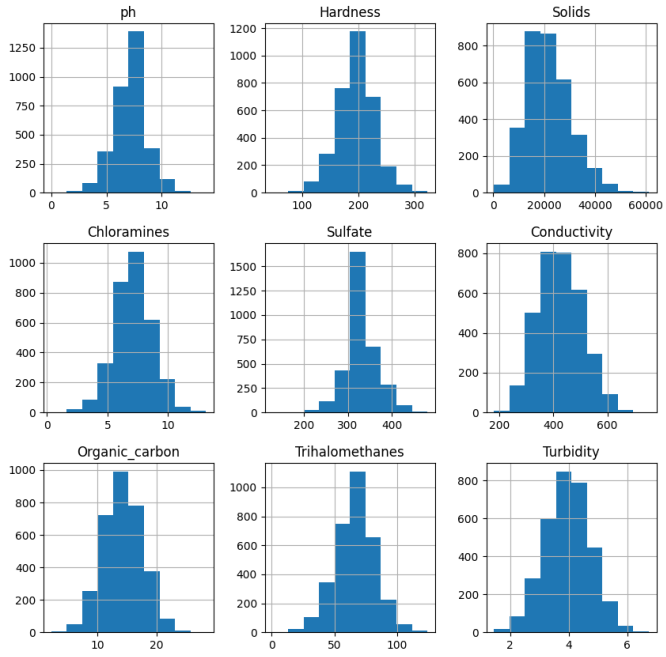

Fig. 3. Class-wise histogram representation of each feature.

E. Data Normalization Using Z-Score

The z-score, widely used for normalization, shows the number of standard errors, which makes the values converted to the mean equal zero. I think we must strive to keep it to an extent that it is at least -3 and not more than +3. By this is referring to the converting of different scales of values to the default scale which has been normalized. By the need of the z-score to commonize the statistics, we have in the beginning denotation of the variance. To do that, we took the mean (mu) from the original, rounded up value (x), added x squared, and finally divided the whole sum by the passenger length. Equation (1) is the variance.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{1}$$

Equation (2) represents the square root of variance.

$$\sigma^2 = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \tag{2}$$

This is followed by the calculation of the Z-score; we subtracted the mean value from an original value and divided it by the standard deviation, thus, the resulting score that is probably lies between 3 and + 3, showcasing how many standard deviations a point is above or below the mean that the equation, x stands for the original value, (mu) for the mean, and (sigma) for the The formula of the Equation (3) is to compute a Z-score.

$$z = \frac{x - \mu^2}{\sigma} \tag{3}$$

TABLE II
DATASET DESCRIPTION BEFORE AND AFTER OVERSAMPLING.

	Before oversampling	**After oversampling**
Not portable	1998	1998
Portable	1278	1998

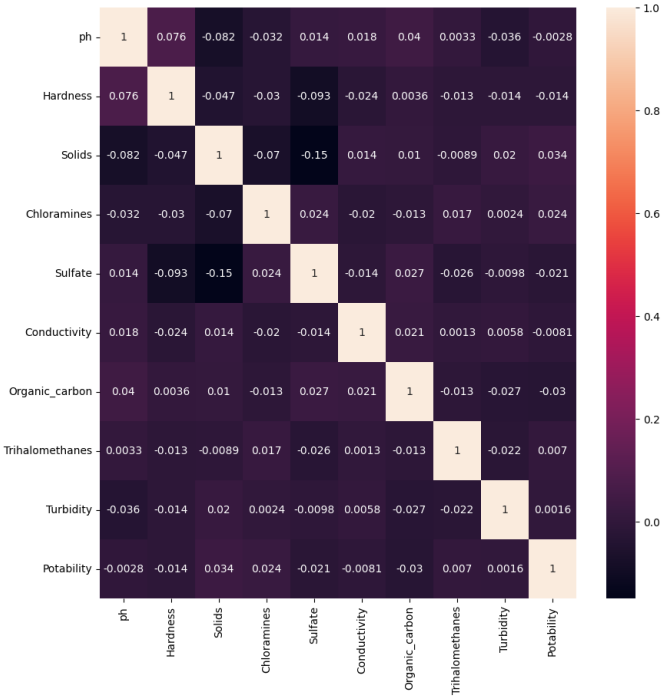

Fig. 4. Correlation Heatmap

IV. SIMULATION

A. ALGORITHMS

1) Logistic Regression: Logistic regression, a binary classifier, is a classification algorithm. It should be known that this model is a result of a logistic function or sigmoid function and the name is due to that. LR is the most ordinary algorithm used in any binary classification, and in our case, we chose multinomial logistic regression because the number of classes being considered was more than just two. We leveraged this as a 'penalize' solver in addition to the L2 penalty.

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Logistic regression	0.60	0.60	1.00	0.75
Support vector machine	0.67	0.66	0.92	0.77
KNN classifier	0.62	0.63	0.88	0.73
Decision tree	0.63	0.64	0.90	0.74
Random forest	0.65	0.65	0.89	0.75
Naive Bayes	0.62	0.63	0.88	0.73
stochastic gradient descent	0.59	0.60	0.97	0.74

TABLE III
OVERALL PERFORMANCE

2) Support vector machine: Firstly, Support Vector Machines (SVMs) are widely exploited in classification. Nevertheless, it is also possible to apply this approach to regression. In the picture, different data points on the plane are marked, and SVM attempts to draw a hyperplane, determining a margin

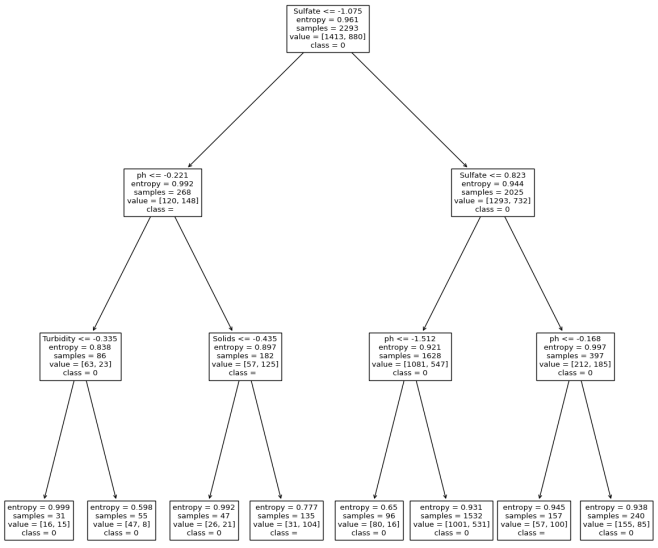

Fig. 5. Decision Tree

which expands to separate two classes as much as possible, thus reducing the number of close errors.

3) K Nearest Neighbor: Similar to this, the nearest K neighbor algorithm makes its predictions by looking at the given points closest to N neighbors and takes the majority classification of N neighbors and assigns that to the point. If the result of the model is not satisfactory, the method can be adjusted. For example, increase n or add bias towards one class in order to achieve conclusive performance. KNN is not meant for large datasets, as the calculation of nearest neighbors occurs only during the time of testing. Also, it goes through every sample data to compute the distance measurement every time. We used a five-person model for the submodel.

K nearest neighbor algorithm assigns classes by identifying for each point the neighbors' N that are closest and to assign it the class with the majority of n neighbors. Example of draw will be different ways to solve it. Increasing n, or make bias for one of the groups. K nearest neighbor algorithm lacks application to larger data sets because the whole testing is done by the process and every time a new nearest neighbors are computed by comparing the full training set data. We used the following n = 5 mode for our design.

4) Decision Tree: A decision tree is a trivial algorithm almost acting as a self-explanatory text for classification as well as regression. The decision tree as such has been trained to decide a set of inputs values. It is entitled to entropy to choose the main/independent variable and, afterward, goes for the other deviation ones. It is similar to the process mapping and that is the reason why it has all the decision parameters arranged as a tree and the different decision will be chosen based on various number of values.

5) Random Forest: Random forest which uses random sets of attributes for its learning is a model that combines many decision trees to generate better decisions. In random forest the base model is this decision tree and the decision tree have all the features of a decision tree plus you have the advantage of using the ideas of multiple models at the same time.

6) Naïve Bayes: Namely, Naive Bayes is a simple and quick calculating algorithm that is based on the Bayes theorem assumption that is unrelated to each other and the probabilities of the mere presence of another feature.

7) Stochastic gradient descent: This loop type function minimizes the loss function number by number till global optimum. With random gradient descent method, the samples are not selected systematically.

Analysis of Fig. 6, 7, 8, 9, 10, 11 yields the confusion matrices for all models LR and RF reaped the most rewarding of true positive and true negative points wherein appropriate steps were taken to handle patients with cancers based on the employed models. All the models display a maximum type I and type II errors with, 1413 false positive and no false negative whatsoever.

The Random forest demonstrated an outstanding confusion matrix with the number of false positives and false negatives its minimum value very high precision. The model of this algorithm is that the manner of highlighting the likelihood of a certain member in a class in the course of consideration across classes and as a result, it is able to pinpoint nearly all classes that have a high number of correct classes. Which means therefore it is highly reliable for the all binary classification that is needed and it can be applied for the classification task.

The Naive Bayes exhibited an exceptional confusion matrix that involved a minor number of false negatives.

• TP represents a hundred of numerous records that are predicted to be potable and clearly potable.

• FP stands for the number of records whose class is predicted to be not drinkable water, but they are acceptable.

• NF is the number of records which are predicted to have water quality but contain the water without water purity.

• TN is the calculation of amount of records being assumed are not safe and actually are not safe.

This depicts that if the algorithm is deployable it will extend the sensor lifetime, which will lead to cost cutting (machine learning algorithms for water quality either will be portable or not). We will measure the performance of the experiment using the metrics of accuracy and area under curve (AUC) in our case study. F-measure is computed by combining the two terms, whereas precision remains fixed, whereas the recall takes a variable value. The accuracy of machine learning (ML) classifier being the percentage of samples along the curve where the machine is said to have delivered a classification of the samples is the measure of model. It can be computed using the following expression: It is obtained from the computational formula listed below.

$$\text{Precision} = \frac{tp}{tp + fp}$$

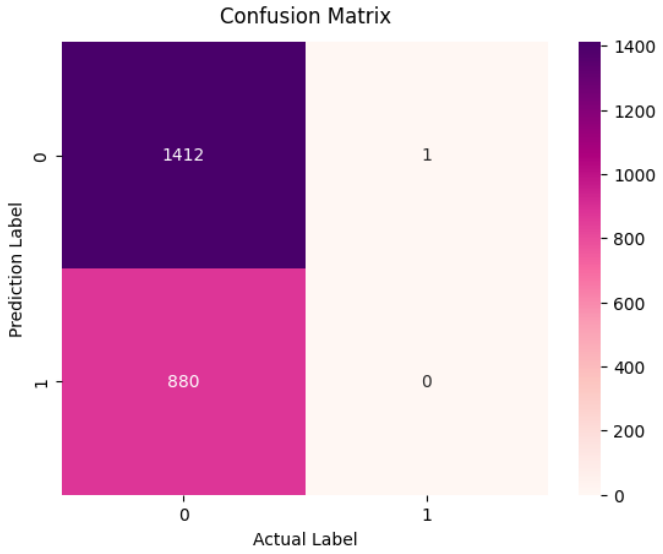

Fig. 6. Confusion matrix of Logistic Regression

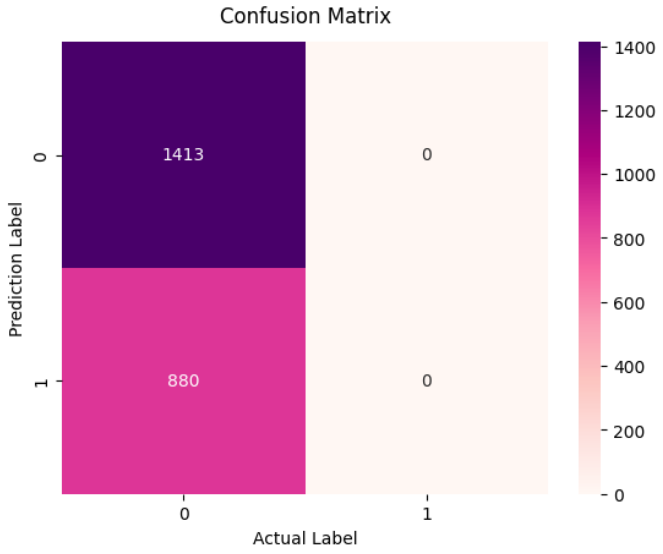

Fig. 7. Confusion matrix of SVM

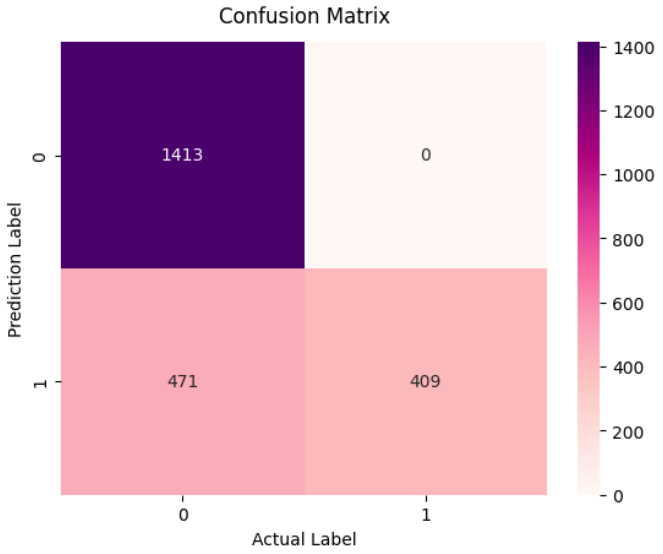

Fig. 8. Confusion matrix of KNN

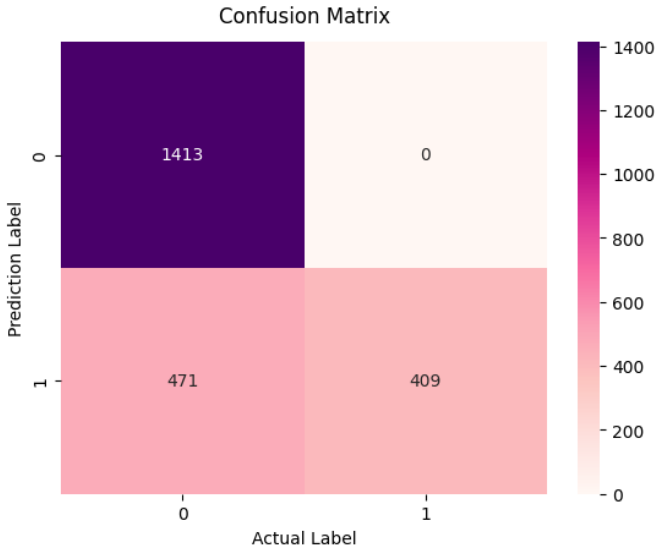

Fig. 9. Confusion matrix of Decision tree

where:

tp : True Positives

fp : False Positives

Although recall is a measure of the neglected samples that were successfully identified as portable among the portable samples that were detected by the model. It is distinct from accuracy because it keeps track of the amount of total correct samples identified as portable. It is calculated using the following formula: It is done by the applying the equation belows :

$$\text{Recall} = \frac{tp}{tp + fn}$$

where:

tp : True Positives

fn : False Negatives

The F1-score is obtained from the two values of recall and precision. Besides, it is a visualization of the relationship between Precision and Recall in a graphic manner.

$$\text{F1-score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

where:

Recall : Recall (Sensitivity)

Precision : Precision

Figure 12 is the performance evaluation chart of the algorithms namely Logistic regression, Support vector machine, K-nearest neighbors, Decision tree, Random forest classifier,

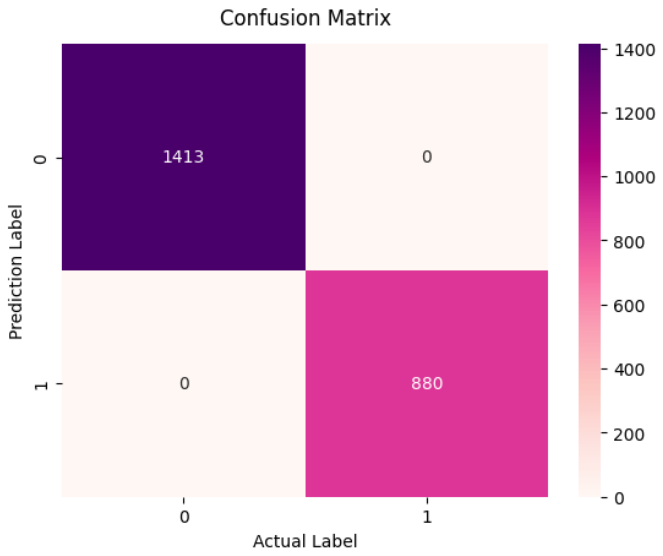

Fig. 10. Confusion matrix of Random Forest

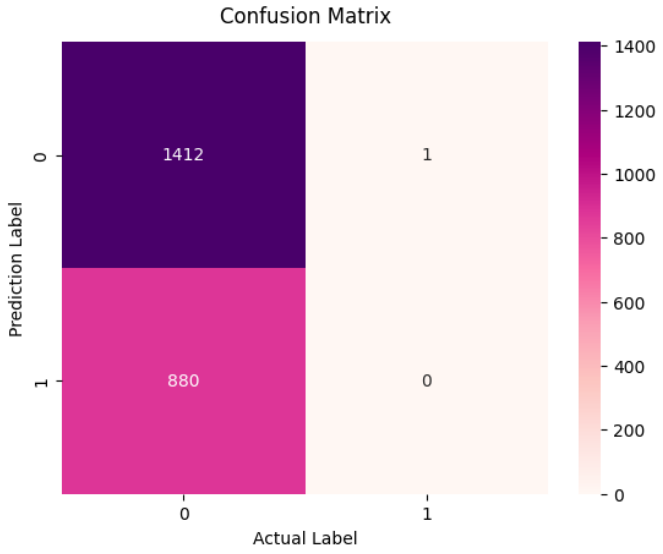

Fig. 11. Confusion matrix of Naive Bayes

Naive Bayes, and Stochastic gradient decent. with this conclusion, we therefore can claim that the random's accuracy forest has the highest accuracy among all the algorithms. The algorithms logistic regression, support vector machine, K-nearest neighbor, and decision tree show similarity regarding their accuracy. the algorithms of the Bayesian Naive, the Gradient of the Stochastic, and the accuracy the same.

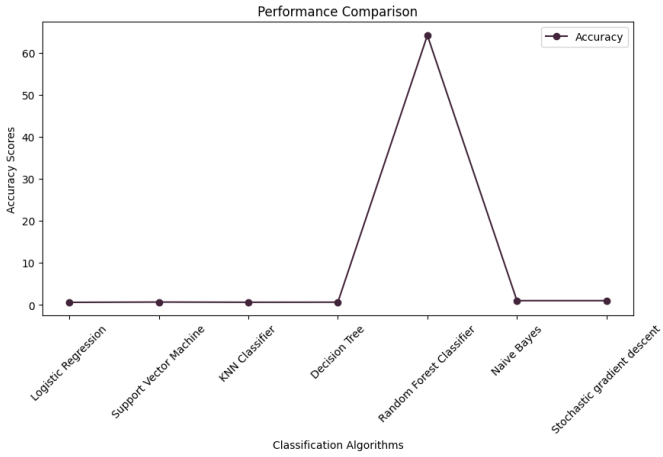

Fig. 12. Performance comparision

V. REFERENCES

[1] Comparison of machine learning algorithms in statistically imputed water potability dataset Diwash Poudela,, Dhadkan Shresthaa, Sulove Bhattaraia and Abhishek Ghimirea et al. (2022).

[2] A Machine Learning-Based Water Potability Prediction Model by Using Synthetic Minority Oversampling Technique and Explainable AI Jinal Patel,1 Charmi Amipara,1 Tariq Ahamed Ahanger, 2 Komal Ladhva,1 Rajeev Kumar Gupta,1 et al. (2022).

[3] Water Potability Prediction Using Machine Learning Samir Patel, Khushi Shah, Sakshi Vaghela et al. (2023).

[4] Machine learning-based forecasting of potability of drinking water through adaptive boosting model Surjeet Dalal, Edeh Michael Onyema*, Carlos Andrés Tavera Romero, Lauritta Chinazaekpere Ndufeiya-Kumasi, Didiugwu Chizoba Maryann, Ajima Judith Nnedimkpa, Tarandeep Kaur Bhatia (2022).

[5] Evaluation of Machine Learning Algorithm on Drinking Water Quality for Better Sustainability Sanaa Kaddoura (2022).

[6] Water Quality Prediction Using KNN Imputer and Multilayer Perceptron Afaq Juna 1,†, Muhammad Umer 1,† , Saima Sadiq 2,†, Hanen Karamti 3, Ala' Abdulmajid Eshmawi 4 , Abdullah Mohamed 5 and Imran Ashraf 6,* (2022).

[7] Water quality prediction using machine learning methods Amir Hamzeh Haghiabi, Ali Heidar Nasrolahi and Abbas Parsaie (2018).

[8] Efficient Water Quality Prediction Using Supervised Machine Learning Umair Ahmed 1, Rafia Mumtaz 1,* , Hirra Anwar 1, Asad A. Shah 1, Rabia Irfan 1 and José García-Nieto 2 (2019). [9] Performance Analysis of the Decision Tree classification Algorithm on the Water Quality and Potability Dataset Umar Zaky 1,*, Ahmad Naswin 2 , Sumiyatun 3, Aris Wahyu Murdiyanto 4 (2023) .

[10] Predicting the Water Potability Index Using Machine Learning Ivan Ivanov, Borislava Toleva* (2023).