Enhancing Water Safety through Advanced Detection Techniques

1st Sajidul Islam Khandaker
Dept. of CSE
BRAC University
Dhaka, Bangladesh
sajidul.islam.khandaker@g.bracu.ac.bd
ID:20301190

2ndTahmina Talukdar
Dept. of CSE
BRAC University
Dhaka, Bangladesh
tahmina.talukdar@g.bracu.ac.bd
ID:20301412

3rd Prima Sarker
Dept. of CSE
BRAC University
Dhaka, Bangladesh
prima.sarker@g.bracu.ac.bd
ID:20301204

4thHumaion Kabir Mehedi
Dept. of CSE
BRAC University
Dhaka, Bangladesh
humaion.kabir.mehedi@g.bracu.ac.bd

5thAnnajiat Alim Rasel
Dept. of CSE
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—This project explores the creation of a complete "Water Safety Detection" system in the area of improving water safety. The main goal is to build a strong model that can recognize possible risks and guarantee safer aquatic ecosystems. The research seeks to train a multidimensional system that can reliably identify numerous safety risks in aquatic environments by utilizing modern machine learning approaches, such as Logistic Regression, Naive Bayes Classifier Algorithm, and K-Nearest Neighbors (KNN) Model. The initiative hopes to usher in a new era of water safety awareness and mitigation through the combination of these algorithms, considerably enhancing the preservation of human life and preventing accidents involving water.**

***Index Terms*—Water Safety Detection, Multidimensional System, Machine Learning, Logistic Regression, Naive Bayes, K-Nearest Neighbors**

I. Introduction

Water, the source of all life, serves as a fundamental base for the survival of every living being on Earth. The presence of functional, reliable, and safe drinking water sources is crucial in today's urban environment. Maintaining a fine balance between adequate water quantity and acceptable water quality across various sectors is an urgent concern brought on by the ever-increasing human population and the ensuing surge in water demand. The word "contamination," which is synonymous with water pollution as seen in the context of other natural elements, best describes this problem. When undesirable and harmful substances are present in water at levels that endanger both the environment and living things, this is referred to as contamination. Unchecked contamination can have disastrous effects, possibly causing serious illnesses and other health issues. Additionally, the resulting decline in water quality lowers the general standard of living for impacted communities.

This paper centers on the use of a meticulously curated dataset with the overarching goal of gauging water quality to address these urgent concerns. The dataset is used to determine the chemical makeup of water, concentrating on important components like arsenic, magnesium, potassium, iron, and more. The paper uses this dataset to try to determine whether water is safe to drink and appropriate for a variety of uses. Essentially, this analysis seeks to ascertain whether water satisfies the requirements for use, relying heavily on quantitative measurements of various chemical components present in it. In order to contribute to the larger conversation about protecting water resources and enhancing society as a whole, this paper sets out on a journey to investigate the complex relationship between water quality, contamination, and the measurement of elemental characteristics.

II. Literature Review

The ability to predict and maintain good water quality has become essential to protecting both the environment and human populations. Numerous investigations have been made to determine the best techniques for predicting water quality and locating potential contamination problems. This literature review synthesizes insights from three prominent papers that delve into distinct approaches for water quality prediction: "Groundwater Quality Prediction Using Logistic Regression Model for Garissa County," "Prediction of Water Quality Using Naive Bayesian Algorithm" by P. Varalakshmi and S. Vandhana, S. Vishali, and "Water Quality Prediction Using KNN Imputer and Multilayer Perceptron" by Afaq Juna, Muhammad Umer, and Hanen Karamti.

In their study titled "Groundwater Quality Prediction Using Logistic Regression Model for Garissa County," the authors address groundwater quality prediction by using a Logistic Regression Model. The study emphasises how important it is to accurately estimate groundwater quality, particularly in areas like Garissa County. The paper highlights the potential of the logistic regression approach to analyse and predict

groundwater quality attributes effectively. The study's findings provide important new information about how regression techniques can be used to address issues with water quality.

The implementation of the Naive Bayesian algorithm for predicting water quality is covered in "Prediction of Water Quality Using Naive Bayesian Algorithm" by P. Varalakshmi and S. Vandhana, S. Vishali. By taking into account various influencing factors, the study emphasises the algorithm's suitability for handling the complexities of water quality assessment. The authors demonstrate the Naive Bayesian algorithm's potency in forecasting water quality levels based on input parameters by utilising its probabilistic nature. This study emphasises how important it is to use cutting-edge computational methods when predicting water quality. The authors of the paper "Water Quality Prediction Using KNN Imputer and Multilayer Perceptron" take a different tack by fusing the K-Nearest Neighbours (KNN) imputer and Multilayer Perceptron. In order to improve the predictive accuracy for water quality assessment, the study acknowledges the contribution of machine learning algorithms. The integration of a Multilayer Perceptron model provides a strong predictive framework, while the use of KNN imputation fills in data gaps. This study demonstrates how machine learning ensembles can successfully handle tasks involving the prediction of water quality. In conclusion, the papers under review demonstrate the variety of methodologies used for predicting water quality. The combination of the KNN imputation and Multilayer Perceptron approach, the Logistic Regression Model, the Naive Bayesian algorithm, and the Naive Bayesian algorithm demonstrate the adaptability of techniques used to address the complex issues involved in assessing water quality. Collectively, these studies highlight the value of cutting-edge computational techniques in preserving water resources and advancing the conversation on water quality management.

III. METHODOLOGY

Data Collection:

The primary data source for this research project is a dataset called "Water quality Dataset for Water Quality Classification." It consists of artificially created urban water quality data that has been painstakingly organized to provide educational value, useful training, and a deeper comprehension of water quality assessment. This dataset has columns named with water quality attributes in addition to its rows and columns. As a teaching tool, it promotes practical experience and understanding of water quality analysis methods. This inventive dataset has the potential to give learners the knowledge and abilities they need for efficient methods for classifying the water quality of a body of water

Data Preprocessing:

The initial stages of data preprocessing reveal the core of our dataset. It serves as the analysis's canvas, having 7999 rows and 23 columns. 22 of the 23 columns depict distinct water quality characteristics, showing the presence of chemical

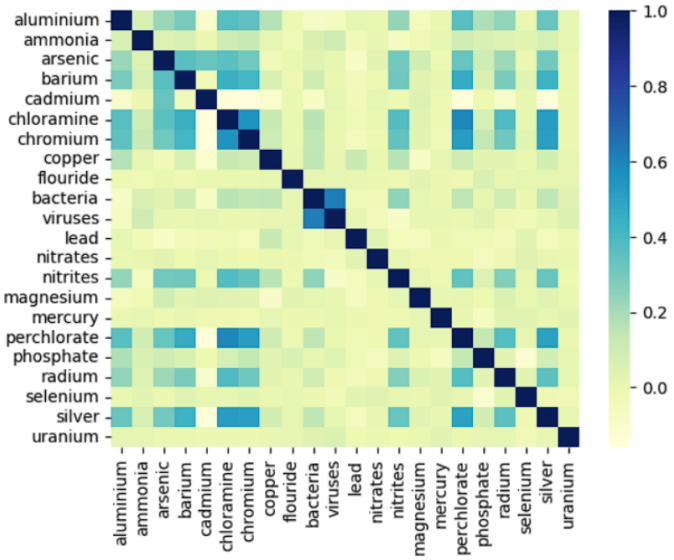

Fig. 1. An Example of heatmap

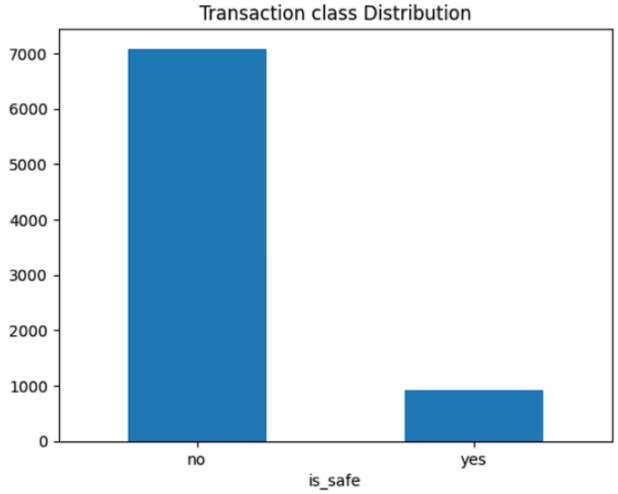

Fig. 2. An Example of imbalance distribution

components. The deciding label, 'is safe,' in the last column, denotes the safety of the water. This knowledge paves the way for enhancing our dataset and ensuring its suitability for the ensuing analytical journey.From the heatmap using seaborn library, we can see how one feature correlates with all features above

Again Here, an evident observation emerges: the ratio of safe to non-safe data is approximately 1:7, delineating an imbalanced distribution within our dataset

After carefully examining our dataset, we discovered important insights that inspired a series of calculated steps to maximise its usefulness and integrity. Removing Phosphate Column: The phosphate column in our dataset of 7999 data points has an astounding 7641 NULL values. Retaining the phosphate column would have limited analytical value given this significant data

absence. As a result, we wisely decided to completely remove the phosphate column, streamlining the dataset for insightful analysis.

Imputing for Magnesium Column: Although not as common as phosphate, our data showed that the magnesium column contains 1584 NULL values, which demanded attention despite not being as prevalent. We cleverly filled these gaps with the mean value of magnesium using data imputation, restoring the integrity of the column.

Feature Engineering through Binary Encoding: We set out to balance our dataset through under sampling after realising how unbalanced it was. Additionally, by using feature importance scoring, we identified the key characteristics guiding our analysis. With this information in hand, we saved all features that would be important for our investigation.

These carefully considered and sequentially performed actions led to the creation of a polished dataset. This dataset, which is balanced and free of NULL values, serves as the basis for the rest of our analytical work.

Evaluation:

The culmination of our efforts is encapsulated within meticulous evaluation framework that gauges the performanc of our models. Prior to this, our dataset was adeptly divide into training (70%) and testing (30%) subsets, ensurin unbiased assessments. This division was coupled wi thoughtful data scaling to bolster model efficacy. Thr distinct algorithms stood out during our evaluation proces Logistic Regression, Naive Bayes Classifier Algorithm, a K-Nearest Neighbours (KNN) Model. Through exhausti training and testing iterations, each algorithm's predictive ability was carefully examined, revealing its unique strengths and weaknesses. Accuracy, precision, recall, and F1-score served as guiding beacons in evaluating the performance metrics' real-world predictive abilities. Our models are put to the test during this evaluation phase, revealing insightful information about their potential contributions to the larger field of improving water safety.

Discussion: The main contributions and potential implications of our study are summarised in the discussion section. We analyse the subtle performance of Logistic Regression, Naive Bayes Classifier Algorithm, and K-Nearest Neighbours (KNN) Model using a foundation rooted in comprehensive model evaluation. We reveal each algorithm's predictive strength in the context of classifying water quality by tying together metrics like accuracy, precision, recall, and F1-score. These algorithmic discoveries connect theoretical expertise with practical applications, paving the way for well-informed choices regarding water safety initiatives. This discussion not only sheds light on the current findings but also projects its rays into the future, imagining avenues for improvement, ensemble techniques, and in-the-moment application that can increase the impact of our research on a wider range of water quality enhancement and public health issues.

IV. RESULT

In the context of our study, the Naive Bayes classifier showed 76% accuracy while the Logistic Regression model classified water quality with an accuracy of 77%. The accuracy of the K-Nearest Neighbours (KNN) model, which was exceptional at 83.39%, stood out. These results highlight the models' ability to predict attributes of water quality and offer insightful information that will help us in our mission to increase public awareness of water safety. By creating a visual classification report for the models, we furthered our analysis. In this report, the designations "0" and "1" designate safe and unsafe water conditions, respectively. These distinctions produce precision, recall, and f1-score values, giving each model's performance in detail.

```
print(classification_report(y_test,pred_kn))
```

	precision	recall	f1-score	support
0	0.93	0.73	0.82	277
1	0.77	0.94	0.85	271
accuracy			0.83	548
macro avg	0.85	0.84	0.83	548
weighted avg	0.85	0.83	0.83	548

Fig. 3. Example of a classification.

A bar graph displaying the 'Logistic Regression,' 'Naive Bayes,' and 'KNeighbours' models provides clear understanding. The 'KNeighbours' model, which boasts an impressive 83.3942% accuracy, stands out as the most accurate. In contrast, "Logistic Regression" and "Naive Bayes" both achieved 77.0073% and 76.6423%, respectively. The superior predictive accuracy of the KNeighbours model within our dataset is clearly illustrated by this graphic representation.

Our models show distinct complexities in the area of precision and recall analysis. The precision and recall values for logistic regression are consistent, reflecting a balanced performance. When switching to Naive Bayes, subtle distinctions become apparent: the '0' label has precision of 81% and recall of 70%, whereas the '1' label has precision of 73% and an impressive recall of 83Further distinctions are made by KNN: '0' label precision is 93% with a recall of 73%, and '1' label precision is 77% with an astounding 94% recall. These details highlight the models' ability to identify particular characteristics of water quality and provide helpful insights

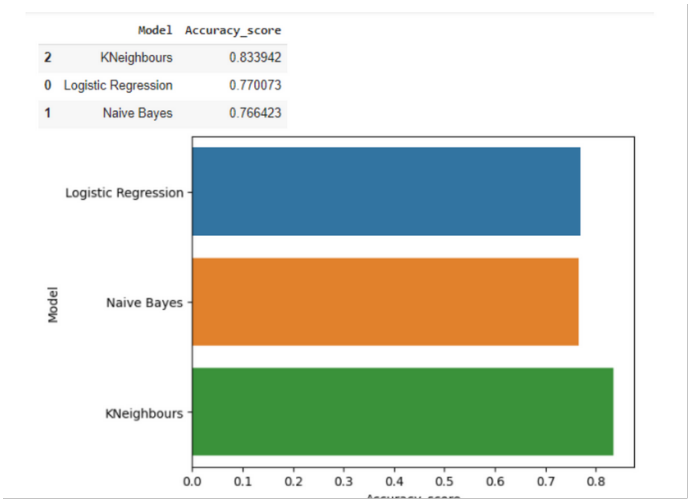

Fig. 4. An Example of barchart

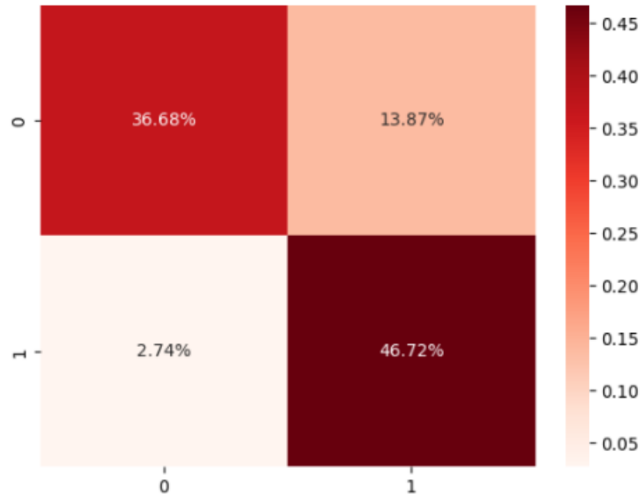

Fig: Confusion matrix for KNeighboures

Fig. 5. An Example of heatmap

into their classification prowess.

Confusion Matrix for each model:
Logistic Regression: The confusion matrix for Logistic Regression using the y test and pred kn arguments can be found here.The TN and FN percentages are now revealed to be 39.05% and 11.50%, while the TP and FP percentages are 37.96% and 11.50%.

Naive Bayes:The y test and pred kn arguments are used to create the confusion matrix for the Naive Bayes model. The TN and FN percentages are now revealed to be 35.58% and 14.96%, while the TP and FP percentages are 41.06% and 8.39%.

KNeighbourse: The confusion matrix for KNeighbourse is created with y test pred kn arguments. Here we get to know that the TN and FN percentages are 36.68% and 13.87% again TP and FP percentages are 46.72% and 2.74%

V. CONCLUSION

This study set out on a data-driven journey fueled by innovation and analysis to determine water safety. We set out on a careful investigation of water quality classification based on a Kaggle dataset. Our data preprocessing journey helped us to understand dataset complexities and address problems that might have jeopardised our analysis. Unbiased model evaluation was made possible by splitting our dataset into training and testing subsets, 70% and 30%, respectively. We utilised the predictive power of three distinct models—Logistic Regression, Naive Bayes, and K-Nearest Neighbours (KNN)—within this framework. Results showed that the KNeighbours model, with an accuracy of 83.3942%, was the most accurate, followed by Logistic Regression (77.0073%) and Naive Bayes (76.6423%). The confusion matrix revealed deeper insights that supported KNN's exceptional performance. Among the three, KNN's predictive abilities stole the show by demonstrating its ability to interpret subtle differences in water safety. As we draw to a close, the results of our study not only highlight the potential of predictive models in preserving public health but also point to KNN's ascension to the top of the water

REFERENCES

[1] Krhoda, G., Amimo, M. O. (2019b). Groundwater quality Prediction using logistic Regression Model for Garissa County. ResearchGate. https://www.researchgate.net/publication/331465889_Groundwater_Quality_Prediction_

[2] Prediction of water quality using Naive Bayesian algorithm. (2017, January 1). IEEE Conference Publication — IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/7951774

[3] Juna, A., Umer, M., Sadiq, S., Karamti, H., Eshmawi, A. A., Mohamed, A., Ashraf, I. (2022). Water quality prediction using KNN imputer and multilayer perceptron. Water, 14(17), 2592. https://doi.org/10.3390/w14172592