e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

WATER QUALITY PREDECTION

R. Sujatha[*1], Bale Swathi[*2], Bhukya Charan[*3], Badhavath Pooja[*4]

[*1]Assistant Professor, Department Of Computer Science And Engineering Malla Reddy College Of Engineering & Technology Hyderabad, India.

[*2,3,4]Final Year Student, Department Of Computer Science And Engineering Malla Reddy College Of Engineering & Technology Hyderabad, India.

ABSTRACT

The primary ideal of this exploration bid is to work machine literacy methodologies for the precise dimension of water quality, particularly fastening on drinkable — a vital metric for assessing the felicity of a body of water for colourful purposes. To gauge the overall water quality in terms of drinkable, a comprehensive set of water quality parameters is employed. These parameters encompass pH, hardness, solids, chloramines, sulcate, conductivity, organic carbon, trihalomethanes, and turbidity, inclusively forming a point vector to depict the water quality. In pursuit of estimating water quality classes, the study delves into the application of two distinct bracket algorithms Decision Tree(DT) and K- Nearest Neighbour (KNN). Experimental analyses are conducted exercising both real- world datasets sourced from different locales across Andhra Pradesh and synthetic datasets generated through arbitrary parameter configurations. The relative assessment of these classifiers reveals that the KNN algorithm surpasses its counterparts in terms of prophetic delicacy. The exploration findings emphasize the efficacity of machine literacy approaches in directly prognosticating water drinkable, therefore emphasizing the applicability and connection of data mining and bracket ways in water quality assessment. Crucial indicator terms similar as Potability, Water Quality Parameters, Data Mining, and Bracket are vital in delineating the compass and significance of this study. likewise, the keywords Machine literacy, Supervised Learning, K- Nearest Neighbour (KNN), Decision Tree, Hyper Parameter Tuning, and Python Programming emphasize the methodological and computational aspects bolstering the exploration, recapitulating the substance of the logical frame employed in this bid.

I. INTRODUCTION

Water quality analysis remains a complex and dynamic field, shaped by a multitude of influencing factors and intricately linked to diverse human and environmental needs. As water serves various purposes across different sectors, ranging from agriculture to municipal supply, the standards for its quality vary accordingly. The ongoing research efforts in water quality prediction underscore its paramount importance in ensuring sustainable water resource management.

Conventionally, water quality assessment revolves around a comprehensive set of physical and chemical parameters tailored to the intended use of the water. Establishing acceptable thresholds for each parameter is essential, delineating the boundary between suitability and unsuitability for specific applications. Water meeting the predefined criteria for a given purpose is deemed suitable, while deviations necessitate remedial measures before utilization.

However, the complexity of water quality evaluation lies in the interconnectedness of its constituent variables, rendering isolated analysis impractical for accurate spatial and temporal characterization. Hence, a holistic approach involves amalgamating multiple physical and chemical parameters into a consolidated metric. This often entails the formulation of quality value functions, typically linear, to encapsulate the relationship between each variable and its corresponding quality level. These functions are derived from direct measurements of substance concentrations or physical variables obtained through extensive water sample studies.

At the forefront of current research endeavours lies the exploration of machine learning algorithms for water quality prediction. Leveraging advanced computational techniques, such as machine learning, holds the promise of unravelling complex patterns and relationships within vast datasets, thereby enhancing our understanding of water quality dynamics. By harnessing the predictive capabilities of machine learning, this research aims to empower decision-makers with actionable insights for proactive water quality management and preservation.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

II. METHODOLOGY

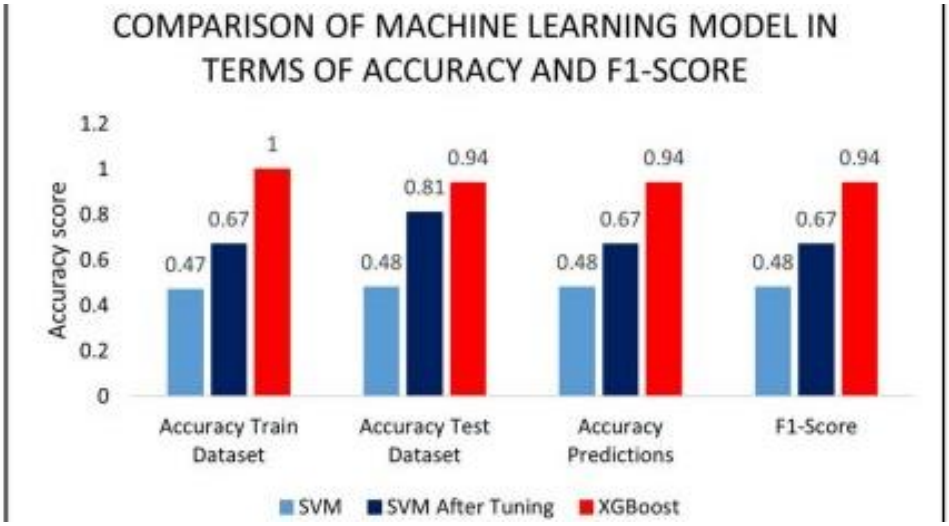
The proposed system aims to determine water potability through a two-phase process involving training and testing. In both phases, the following procedures are carried out. The dataset selection process involves identifying essential parameters influencing water quality, determining the number of data samples, and defining class labels for each data sample—a prerequisite for model construction.

In this study, ten crucial parameters, including pH, hardness, solids, chloramines, sulcate, conductivity, organic carbon, trihalomethanes, turbidity, and potability, constitute the dataset. However, the proposed approach remains flexible regarding the number and selection of parameters. To establish a robust learning and testing framework, a k-fold cross-validation technique is employed, dividing the dataset into k-disjointed sets with similar class distributions. Each subset serves as a test set in turn, while the remaining subsets act as the training set.

Decision Tree (DT) and K-Nearest Neighbour (KNN) methods are employed for classification, each approaching the underlying relational structure between indicator parameters and class labels differently. Consequently, the performance of each technique may vary for the same dataset. To validate classifier performance on unknown datasets, various metrics provided by data mining techniques are utilized.

The classification process involves estimating river water quality class using DT and KNN methods, both parametric and nonparametric classifiers, respectively. Parametric classifiers make assumptions about the form of the mapping function, while nonparametric classifiers do not, resulting in potentially higher accuracy. DT relies on learning techniques, while KNN operates based on similarity principles, making it advantageous for small datasets with complete domain expertise.

Comparing the operation modes of different classifiers is essential to determine the most suitable for approximating the underlying function for water quality datasets. By evaluating their performance on training and testing data, insights can be gleaned into their efficacy in predicting water potability.



III. MODELING AND ANALYSIS

Data Collection and Generation

Data mining techniques rely heavily on domain knowledge to generate accurate predictions. For applications related to water quality, a comprehensive understanding of how various parameters impact water quality is essential. This knowledge can be acquired from domain experts or historical data collections. In the context of forecasting, two types of datasets were employed: a meticulously crafted large synthetic dataset and an existing real dataset. The key similarity between these datasets lies in their examination of an equal number of indicator parameters, although the number of samples differs between them. The real dataset has a limited number of observations due to the scarcity of large authentic datasets, necessitating the creation of synthetic data.

The synthetic dataset, carefully designed to mirror real-world scenarios, preserves identical relational structures and distributions of water quality parameters. Ten essential water quality parameters, including pH and Hardness, were utilized to assess overall water quality in terms of potability for each dataset. These

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

parameters were chosen based on their common monitoring and critical significance, aligned with well-defined water quality standards. However, the predictive modelling described in this paper remains adaptable to any number of parameters.

Synthetic Dataset Creation

The utilization of data mining methods necessitates a target dataset with sufficient volume to identify patterns effectively. To fulfil this requirement, a synthetic dataset was generated, offering a realistic approach to obtain a large dataset. This synthetic dataset was meticulously crafted, considering potential ranges of water quality parameters. These concentration ranges were established after meticulous review of water quality standards set by various national and international organizations such as the European Union (EU), the World Health Organization (WHO), and the Central Pollution Control Board (CPCB), among others.

Each sample in the synthetic dataset reflects a combination of concentration values for the 10 parameters under investigation. To facilitate the development of a predictive model using classification techniques, the dataset was supervised. This involved assigning a label to each instance to predict the water contamination level. Subsequently, potability was determined for each instance based on the concentration values of the selected parameters.

IV. RESULT AND ANALYSIS

Performance Metrics Results Performance criteria play a pivotal part in assessing the effectiveness of bracket algorithms. In this environment, the following performance measures are employed True Cons(TP) Cases where the model rightly predicts the positive class. True Negatives(TN) Components of a confusion matrix indicating cases rightly prognosticated as negative class. False Cons(FP) Cases inaptly prognosticated as positive class by the model. False Negatives(FN) Instances inaptly prognosticated as negative class by the model. delicacy, the most abecedarian performance metric, is calculated as the rate of rightly prognosticated compliances to the total number of compliances Accuracy = Accuracy = TP TN/ TP FP FN TN The delicacy scores, perfection, recall, and f1- Score for two bracket algorithms, Decision Tree and K- Nearest Neighbour, are presented in Table 1 for comparison. Table 1. Comparison of Algorithms SN. Algorithm Type delicacy Score Precision Recall f1- Score 1 Decision Tree58.50.420.380.40 2 K- Nearest Neighbour61.70.430.120.18 These criteria give perceptivity into the performance of each algorithm in rightly prognosticating drinkable. While Decision Tree exhibits a advanced delicacy score compared to K- Nearest Neighbour, KNN demonstrates advanced perfection but lower recall and f1- Score. Each metric contributes to understanding the strengths and sins of the bracket algorithms in the environment of water quality vaticination.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Id	STATION C	LOCATION	STATE	Temp	D.O. (mg/l	PH	CONDUCT	B.O.D. (mg	NITRATEN	FECAL COL	TOTAL CO	year
2	1	1393	DAMANGA	DAMAN &	30.6	6.7	7.5	203	NAN	0.1	11	27	2014
3	2	1399	ZUARI AT I	GOA	29.8	5.7	7.2	189	2	0.2	4953	8391	2014
4	3	1475	ZUARI AT I	GOA	29.5	6.3	6.9	179	1.7	0.1	3243	5330	2014
5	4	3181	RIVER ZUA	GOA	29.7	5.8	6.9	64	3.8	0.5	5382	8443	2014
6	5	3182	RIVER ZUA	GOA	29.5	5.8	7.3	83	1.9	0.4	3428	5500	2014
7	6	1400	MANDOVI	GOA	30	5.5	7.4	81	1.5	0.1	2853	4049	2014
8	7	1476	MANDOVI	GOA	29.2	6.1	6.7	308	1.4	0.3	3355	5672	2014
9	8	3185	RIVER MAI	GOA	29.6	6.4	6.7	414	1	0.2	6073	9423	2014
10	9	3186	RIVER MAI	GOA	30	6.4	7.6	305	2.2	0.1	3478	4990	2014
11	10	3187	RIVER MAI	GOA	30.1	6.3	7.6	77	2.3	0.1	2606	4301	2014
12	11	1543	RIVER KAL	GOA	27.8	7.1	7.1	176	1.2	0.1	4573	7817	2014
13	12	1548	RIVER ASS	GOA	27.9	6.7	6.4	93	1.4	0.1	2147	3433	2014
14	13	2276	RIVER BIC	GOA	29.3	7.4	6.8	121	1.7	0.4	11633	18125	2014
15	14	2275	RIVER CHA	GOA	29.2	6.9	7	620	1.1	0.1	3500	6300	2014
16	15	3189	RIVER CHA	GOA	30	6	7.5	72	1.6	0.2	4995	9517	2014
17	16	1546	RIVER KHA	GOA	29	7.3	7	247	1.5	0.2	1095	2453	2014
18	17	2270	RIVER KHA	GOA	29.1	7.3	7	188	1	0.1	1286	3048	2014
19	18	2272	RIVER KUS	GOA	28.7	7	6.9	224	1.2	0.3	3896	6742	2014
20	19	1545	RIVER MAI	GOA	28.7	7.3	6.7	144	1.5	0.1	1940	3052	2014
21	20	2274	RIVER MAI	GOA	29.5	5.3	6.8	319	1.8	0.3	6458	10250	2014
22	21	2271	RIVER SAL	GOA	29	6.3	6.4	79	1.6	1.4	7592	12842	2014
23	22	2273	RIVER SAL	GOA	29.4	5.4	7.6	39	1.4	0.1	3176	6367	2014
24	23	3183	RIVER SAL	GOA	28.3	2.2	6.5	322	4.7	1.2	11210	14920	2014
25	24	3184	RIVER SAL	GOA	30.1	5.2	7.1	192	2.6	0.3	5073	8925	2014
26	25	3190	RIVER SIN	GOA	30.3	5.6	7.5	282	1.8	0.1	3205	5082	2014
27	26	3191	RIVER SIN	GOA	30.5	5.5	7.4	275	1.5	0.1	4698	8625	2014

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

V. CONCLUSION

Water potability serves as a crucial determinant of water quality, a fundamental resource essential for sustenance. Traditionally, assessing water quality entailed costly and time-intensive laboratory analyses. However, this study delves into an innovative approach employing machine learning methods to predict water quality using simplified criteria.

By harnessing a curated set of representative supervised machine learning algorithms, this approach aims to identify water of inferior quality prior to consumption, enabling timely intervention by relevant authorities. The overarching objective is to curtail the consumption of subpar water, thereby mitigating the risk of waterborne diseases such as typhoid and diarrhoea.

VI. FUTURE SCOPE

Furthermore, the adoption of prescriptive analysis, leveraging projected values derived from machine learning models, holds promise in enhancing future capabilities to support decision-making and policy formulation. By furnishing insights into potential water quality issues, this approach empowers decision-makers and policymakers to implement proactive measures ensuring the safety and sustainability of water resources.

VII. REFERENCES

[1] The importance of monitoring water quality is underscored by various studies and reports, such as the National Water Quality Monitoring Programme's Fifth Monitoring Report by the Pakistan Council of Research in Water Resources (PCRWR). This report sheds light on the state of water quality, providing valuable insights into areas requiring attention and improvement.

[2] Similarly, research efforts by Kangana et al. (2017) have contributed to the development of a water quality index (WQI) specifically tailored for assessing the Loka Lake in India. This index serves as a comprehensive tool for evaluating multiple parameters and their collective impact on water quality.

[3] Thukral et al. (2005) discuss the significance of water quality indices (WQIs) as effective tools for simplifying complex water quality data into easily interpretable metrics. These indices play a crucial role in facilitating informed decision-making regarding water resource management and conservation.

[4] Srivastava and Kumar (2013) explore the challenges associated with calculating water quality indices in scenarios where certain parameters are missing or unavailable. Their research provides insights into methodologies for addressing such gaps in data, ensuring the accuracy and reliability of water quality assessments.

[5] Furthermore, the Environmental Protection Agency (EPA) emphasizes the importance of understanding and monitoring various parameters of water quality to safeguard human health and the environment. By comprehensively evaluating parameters such as pH, turbidity, and dissolved oxygen levels, among others, authorities can effectively identify and address potential threats to water quality.

[6] Collectively, these studies and resources underscore the critical importance of continuous monitoring and assessment of water quality, highlighting the need for robust methodologies and tools to ensure the sustainability and safety of water resources.