Water Quality Prediction Using Catboost Classifier Algorithm

A Yogeshwari [1]
Dept. of Information Technology
Sathyabama Institute of Science and Technology
Chennai 119,
India

J Anubama[2]
Dept. of Information Technology
Sathyabama Institute of Science and Technology
Chennai 119,
India

Mrs. J. Merlin Mary Jenitha[3]
Assistant Professor
Dept. of Information Technology
Sathyabama Institute of Science and Technology
Chennai 119,
India

Mrs. C. Geetha[4]
Assistant Professor
Dept. of Information Technology
Sathyabama Institute of Science and Technology
Chennai 119,
India

Mrs. D. Ramalakshmi[5]
Assistant Professor
Dept. of Information Technology
Sathyabama Institute of Science and Technology
Chennai 119,
India

B.Pavithra[6]
Assistant Professor
Department of Computer Science & Engg,
Sathyabama Institute of Science and Technology
Chennai 119,
India

Abstract— Human existence depends on water, so its purity must be assessed for improved consumption. The goal is to use a cat-boost algorithm to forecast the water quality findings as accurately as possible. It is anticipated that improving water quality will lessen health-related issues. By gathering a variety of data, including variable detection and analysis methods like univariate, bivariate, and multivariate analysis, the study of the water quality is carried out. The study compares and discusses how different algorithms function. Data gathering is the first step in the process, during which historical information about the water condition is gathered. With both dependent and independent factors, data analysis is performed. The main objective is to identify the physicochemical characteristics of water, including temperature, pH, EC, hardness, chlorides, alkalinity, phosphate, and sulfate in water samples taken from various monitoring sites. To forecast the trend of data and the quality of the water, the appropriate algorithm is used. The structure has logins for users and government officers, with the latter having access to submit complaints to officials. In order for the police to quickly locate the user, the user is also permitted to provide information about the water purity in addition to their residential data.

Keywords—water quality, Prediction, Accuracy, Data-Preprocessing, pH, Data Visualization, Cat boost Algorithm

I. INTRODUCTION

Water is the basic need of Human life. It has a direct impact to the health of human and to eco system. Water is the primary source in human activities. Water quality assessment is an important step in the development of agricultural endeavors in terms of trimming design, type of water system framework, and water refinement frameworks for business. Water quality is determined by various factors such as pH, temperature, nitrate, dissolved oxygen (DO), electrical conductivity, sediment, and cellular oxygen consumption are all variables that must be considered. (BOD). In many areas of the world, aquatic insects move and spread diseases like dengue fever, diarrhea, and schistosomiasis which is very harmful to the human health and also for the environment. water are not just affect the human health but also infect the society causing genetical and hereditary problems resulting a great impact in the human culture. In order to prevent these medical impacts and to safeguard human wellbeing the quality of water must be analyzed

II. MATERIALS ANDMETHODS

The proposed strategy entails creating a framework for the categorization of water quality and for a user-friendly report system. The procedure begins with data collection, where previous data connected to water quality

measures are gathered. Water purity is discovered beforehand, so it can reduce issues in human health. Machine learning is now being used to minimise manual error, which can improve the precision in forecasting the quality of water using different measures collected in the dataset. Data analysis is performed on the information, and appropriate variable identification is performed where both dependent and independent variables are discovered. Following that, the Cat boost Algorithm is used to forecast water purity.

III. PROPOSED METHOD

The proposed system has seven modules namely Data Pre-processing, Data Analysis, Data Visualization, Implementing Cat Boost Algorithm, Implementing Logistic Regression Algorithm, Implementing Random Forest Algorithm, Implementing Naïve Bayer's Algorithm and Deployment of the system.

[a] Data Pre -Processing: The process of preparing and assessing data for training. This program collects data from various sources and then removes null values, useless values, and incorrect data. This module processes the water in different areas such as pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic- carbon, Trihalomethanes, and Turbidity. The Water Quality for the specified categories is analyzed within a certain limit.

[b] Table I – water components and its portable range

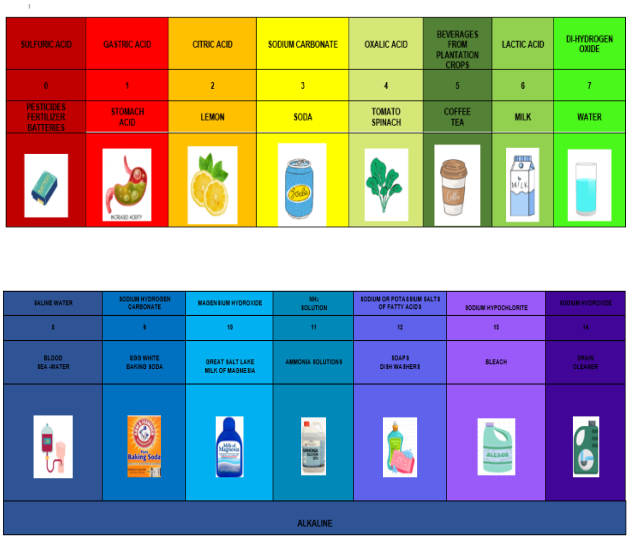
WATER COMPONENT	*ADEQUATE RANGE*
ph	6.5 to 8.5
Hardness	120 to 170 Mg
Solids	50 to 150
Chloramines	4 Mg/l or 4ppm
Sulfate	500 Mg/l
Organic carbon	Below0.500 Mg/l TOC
Trihalomethanes	100 ppb
Turbidity	Below 1 TOC
Conductivity	200 to 1000 mu/cm

*[1] pH :*The pH level of the water indicates whether the water is acidic or basic with the range of 0 to 14. The pH numbers less than 7 suggest acidity, while pH values greater than 7 indicate baseness. The pH of water is actually a determination of the amount of free hydrogen and hydroxyl molecules. Water that contains more free hydroxyl ions is basic, whereas water that contains more free hydrogen ions is acidic. Each figure represents a 10-fold variation in the acidity or basicity of the water. Water with a pH of five is ten times more corrosive than water with a pH of six.

Table II – pH of different water

water compound	*ph level*
Osmotic water	5 to 7
Faucet water	7.5
Alkaline water	8 to 9
Sea Water	>8
Acid	5 to 5.5

Figure I- pH level



[2] Hardness: The high concentration of calcium and magnesium refers to the hardness of water. The overall hardness (mg/L) is the aggregate of the calcium and magnesium concentrations, both given as calcium carbonate which is used to measure water hardness. The metal ions in hard water precipitate out of solution as insoluble precipitates, making them inaccessible to living creatures .

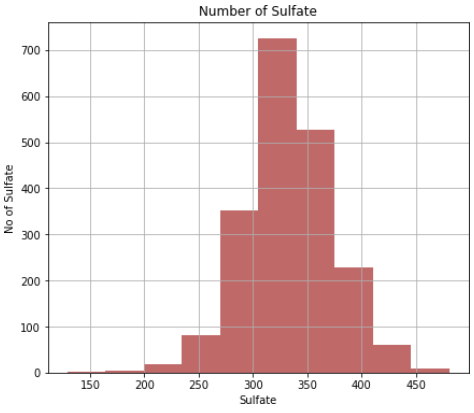
[3] Solids :Solids in water refers dissolved particles as well as floating and settleable particles. An organism swells when placed in water with few solids, such as distilled water is

particularly worrisome when pesticides are involved. Pesticide concentrations may increase considerably above those of the original application as irrigation water travels through irrigation channels in regions with a lot of solids. Increased solids ratios can block irrigation systems and cause watered plant roots to lose water instead of receiving it. Total solids are formed as a result of industrial pollution, sewage, fertilizers, fertiliser runoff from roads, and soil erosion.

[4] Chloramines:The method of disinfecting and eliminating pathogens from drinking water by adding chloramine. Monochloramine is a form of chloramine used to disinfect drinking water. Chloramine lasts longer in water pipelines and produces fewer disinfection residues. Some water providers are switching to chloramine to comply with EPA regulations that limit product disinfection.

*[5] Sulfate:*Water containing more than 250 mg/L sulphate may have a bitter or medicinal flavour. More corrosion-resistant conduit components, such as polyethylene lines, are commonly used in areas with high sulphate levels. Sulfate elements can be found in dirt and geological layers, and some of them dissolve into groundwater as it travels through. Epsom salt, sodium sulphate, and calcium sulphate are all elements that contain sulphate. (gypsum).Sulfate levels in groundwater in Minnesota are usually low—less than 250 mg/L. The southwest and the state's western border have greater sulphate amounts. Sulfate levels are also high in some wells in the state's northeastern and southeastern regions, though less frequently.

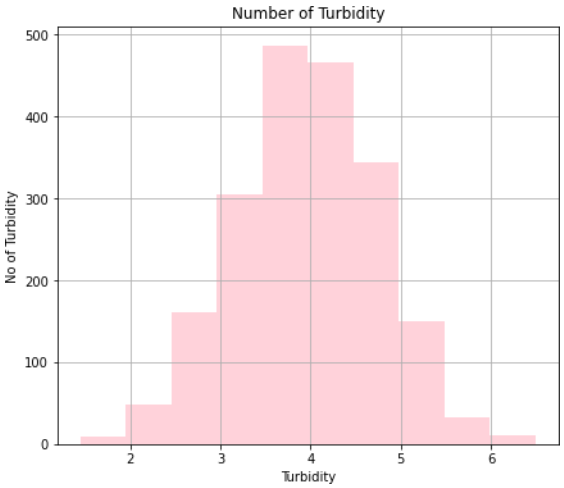
Figure II



.

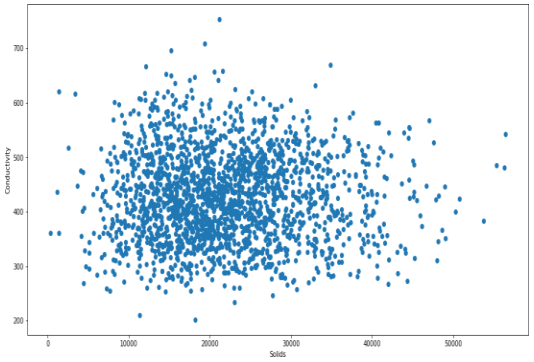
[6] Turbidity: Turbidity refers to the measure of the transparency of a liquid. When light is transmitted through a water sample, the quantity of light scattered by the components of water is measured. Turbid water looks hazy or murky due to particulates that are dissolved or suspended in it and disperse light. Particulate matter includes inorganic materials, soluble colored organic compounds, cyanobacteria, and other microscopic creatures.

Figure III



[7] Conductivity: The conductivity of water is a measurement of its ability to transport an electrical charge. Because electrical current can pass through dissolved salts and other inorganic compounds, conductivity increases with saltiness. Water conductivity is important because it can reveal the amounts of dissolved substances, chemicals, and elements in a fluid. The salt and conductivity of the water can be used to identify what is dissolved in it. A higher conductivity number in the SWMP data indicates that there are more compounds dissolved in the water. transmission is used to quantify electrical transmission in water.

Figure IV



[c] Data Visualization:Thevisual representation of data is known as data visualization. By utilizing graphic

components such as charts and graphs, data visualization tools make it simple to spot and assess trends, anomalies, and patterns in data.

Figure V

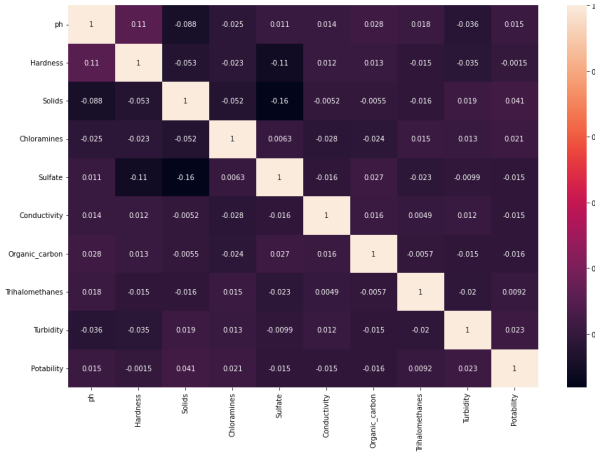
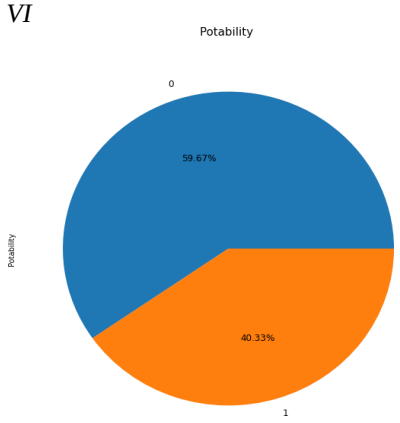


Figure VI



*[d] Random Forest Algorithm:*The Random Forest can do both classification and regression which can manage large databases with many variables to improves the model's precision and prevents overfitting.The Increase in large amount of data to find more accuracy might slow down the performance of the model.

[e] Naïve Baye's Algorithm: The Naive Bayes classifier is a quick, precise, and dependable technique that is used because it has high precision and speed on large datasets but the premise of these distinct predictor parameter, which implies that all of the parameter are independent of one another parameter

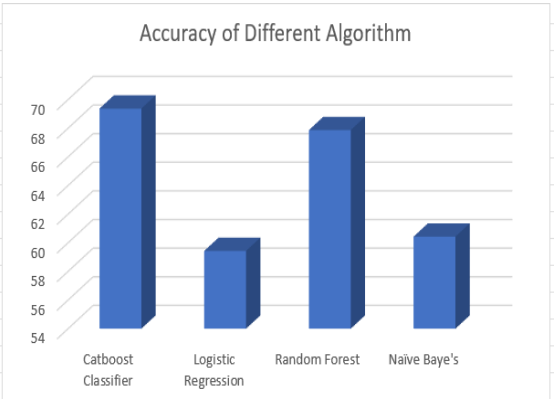
[f] Logistic Rsegression: Logistic regression is a technique for predicting the likelihood of an objective variable whichhas only two potential groups. The dependent variable is binary in character, with data recorded as either 1 (for success/yes) or 0 (for failure/no).

[g] Cat boost Algorithm: Cat Boost is a combination of the words "Category" and "Boosting" which is a categorical-based program. This implies the method of converting categorical values to numerical values by using various statistics on categorical feature pairings and categorical and numerical feature combinations.

Benefits of CatBoost:

The main advantageof using cat boost classifier algorithm is because of its efficiency to deal with categorical features. It's also rapid, simple to use, precise and incorporating categorical factors into your feature collection for better accuracy.

IV. RESULTS AND DISCUSSION



Water purity cannot be determined solely by one aspect. Because the catboost classifier deals with data classification and the data set contains numerous data factors such as ph, hardness, solidity, turbidity, chloramines, organic compounds, and sulfate, catboost is evaluated with the highest level of precision when compared to other algorithms.

V.CONCLUSION

The models investigate the feasibility of using Supervised Machine Learning techniques such as the Catboost Algorithm for predicting the quality of water and todeploy a framework for a water complaint system. The Catboost Algorithm is used in the implementation of a system to improve the precision of water purity, resulting in better health and cost reductions. The purity of water is expected

to reduce the number of problems that arise.

VI.REFERENCES

[1] A. N. Hasan and K. M. Alhammadi, "Quality Monitoring of Abu Dhabi Drinking Water Using Machine Learning Classifiers," 2021 14th International Conference on Developments in eSystems Engineering (DeSE),2021

[2] Ajayi 1, Antoine B. Bagula1 ,Hloniphani C. Maluleke 1 , ZaheedGaffoor 2 , Nebo Jovanovic2 , And Kevin Bellville, Cape Town 7535, South Africa 2Department of Earth Science, South Africa- WaterNet: A Network forMonitoringandAssessingWaterQualityfor Drinking andIrrigation.

Conference on Research in Intelligent and Computing in Engineering (RICE), 2018

[3] Dao Nguyen Khoi, NguyenTrong Quan, Do

[4] Di Wu;Hao Wang; Razak Seidu 2019 IEEEIntl Conf on Dependable, Autonomic and Secure Computing, - Tensor Model for Quality Analysis in Industrial Drinking Water Supply SystemIntl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing,(DASC/PiCom/CBDCom/Cyber SciTech) Year: 2019 | Publisher:IEEE

[5] G. Sebestyen, A. Hangan and Z. Czako, "Anomaly detection in water supply infrastructuresystems,"202123rdInternationa l Conference on Control Systems and Computer Science (CSCS), 2021

[6] GaganjotKang;JerryZeyu Gao; Gang Xie 2017- Data-Driven Water Quality Analysis and PredictionIEEEThirdInternationalConferenc e on Big Data Computing Service Applications (Big Data Service)Data Driven Water Quality AnalysisandPrediction:Year:2017|Publisher:

[7] Hao Wang; Hadi Mohammed; Razak Seidu-Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception Di Wu EEE Transactions on Sustainable Computing Year: 2020 Publisher:IEEE

[8] Huimin Jia; Xiaofeng Zhou – Water Quality Prediction Method Based on LSTM-BP - 2020 12th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC) Year: 2020 | Publisher:IEEE
IEEE 36

[9] Jian Zhou, Yuanyuan Wang, Fu Xiao, Yunyun Wang and Lijuan Sun - Water Quality Prediction Method Based on IGRA and LSTM which is proposed in the year 2018.Water quality prediction has great significance for water environmentprotection.

[10] Jitha P Nair; M S Vijaya 2021 International Conference on Artificial Intelligence andSmart Systems (ICAIS)Year: 2021 - Predictive Models for River Water Quality using Machine Learning and Big Data Techniques - A Survey Publisher:IEEE

[11] K. P. Rasheed Abdul Haq;V. P. Harigovindan Year:2022-WaterQualityPreditcionforSmart Aquaculture Using Hybrid Deep Learning Models

[12] Md. Saikat Islam Khan, Nazrul Islam, Jia Uddin ,Sifatul Islam, Mostofa Kamal Nasir in the year 2021 - A study on Water Quality PredcitionandClassificationbasedonPrincipal Component Regression and Gradient boosting Classifier approach

[13] N. Kumar Koditala and P. Shekar Pandey, "Water Quality Monitoring System Using IoT and Machine Learning," 2018International

[14] N.M.Ragi,R.HollaandG.Manju,"Predicting Water Quality Parameters Using Machine Learning," 2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), 2019

[15] N.RadhakrishnanandA.S.Pillai,"Comparison of Water Quality Classification Models using Machine Learning," 2020 5th International Conference on Communication and Electronics Systems (ICCES),2020

[16] P.Varalakshmi - Prediction of water quality using Naive Bayesian algorithm Year: 2017 | Publisher:IEEE

[17] Priskilla Angel Rani;R.Sujeethra;C. YesubaiRubavathi - Prediction of Water Quality using Artificial NeuralNetwork

Quang Linh, Pham Thi Thao Nhi, and Nguyen ThiDiemThuyintheyear2022– astudyusing Machine Learning Models for Predicting the Water Quality Index in the la Buong Rive r,Vietnam

[18] QuanXi Dong; YongZheLin;JingBi;Haitao Yuan International Conference on Systems, Man and Cybernetics (SMC)Year: 2019 -An Integrated Deep Neural Network Approach for Large-Scale Water Quality Time Series Prediction Publisher:IEEE

[19] Razak seidu-Quality Risk Analysis for Sustainable Smart Water Supply Using Data Perception IEEE Transactions on Sustainable Computing Year: 2020 | Publisher:IEEE

[20] S. J. Kale, M. A. Khandekar and S. D. Agashe, "PredictionofWaterFilterBedBackwashTime for Water Treatment Plant using Machine Learning

Algorithm," 2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA), 2021

[21] SanazImen, Ni-Bin Chang, Senior Member, IEEE, Y. Jeffrey Yang, and ArashGolchubian Developing a Model-Based Drinking Water Decision Support System Featuring Remote Sensing and Fast LearningTechniques

[22] Rongali Yang- Summary of Water Quality Prediction Models Based on