IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

WATER QUALITY AND POTABILITY USING MACHINE LEARNING

M Srikanth1 and K Anish Reddy2

[1] *Student, Sreenidhi Institute of Science and Technology*

[2]*Student, Sreenidhi Institute of Science and Technology*

Abstract Access to clean water to drink is crucial for good health, a fundamental human right, and an essential aspect of any health protection strategy. On a national, regional, and local level, this is significant as a development *:* and healthcare issue. As a result, monitoring water quality now heavily relies on modeling and forecasting water quality. In this work, Machine learning algorithms are developed namely Decision tree, Random Forest, and Naïve Bayes algorithms for the dataset to predict the safety for human consumption. The used dataset has 10 significant parameters and the developed model will be evaluated based on some statistical parameters. The goal of this algorithm is to create a model that predicts the value of a target variable (Decision tree, Random Forest, and Naïve Bayes score) and whether the water is potable ornot.

Index Terms - Naive Bayes, Machine Learning, Decision tree, Dataset, Random Forest

I. INTRODUCTION

All life on earth relies on water, making it one of our fundamental necessities. With a surface area of roughly 71% of the entire planet, it controls the majority of the available space. Water that is continuously extracted from the surface or the ground and used to such an extent to where it can no longer be used is sometimes referred to as water consumption and use. Water that has been contaminated by anthropogenic contaminants and is unfit for human consumption is referred to as contaminated water. Additionally, industries contribute to this pollution. Waterborne diseases in aquatic organisms can be brought on by pathogens in this polluted water. Several main causes of water contamination are population growth and advancements in technology. If the present state of affairs persists, life on Earth will be untenable since there will be a tremendous demand for water and a potential shortage. As a result, maintaining water potability now heavily relies on the prediction and modeling of water quality. In this study, decision trees, random forests, and simple Bayes algorithms are built for the dataset to predict food safety for human consumption. The created model will be assessed using some statistical variables, and the employed dataset has 10 significant parameters. With the help of decision trees, a random forest, and a naive Bayes score, this methodology tries to produce a model that can predict the value of the desired variable and whether or not the water is potable.

SCOPE OF THE PROJECT

II. This helps us to know whether water is potable (i.e. safe) for drinking water or not. Very good results for finding the results of water. More accuracy is defined and threats of diseases are eliminated. It can be applied very easily and less equipment is provided. The cost of implementation is low. It is essential and every human being has the right for drinking safe water.

Project Overview

Supervised machine learning requires training machines by providing vast datasets and parameters. This project uses various classification algorithms such as Random Forest, Decision Tree and Naive Bayes**.** These algorithms are used to find the patterns and recognize them and then using different patterns we find accuracy and better performance.

III. LITERATURE SURVEY

A. EXISTING SYSTEM

A water-based body's quality is assessed using a Water Quality Index (WQI), which is a monetary expression. To assess the total water quality in terms of the WQI, parameters associated with water quality were used. These measurements included the temperature, conductivity, pH, biochemical oxygen demand (BOD), oxygen in solution (DO) (% sat), and nitrates (NO3). To depict the water quality, the aforementioned variables are used as vector features.

B. PROPOSED SYSTEM

The proposed system is we use a machine learning algorithm to predict water quality. It consists of two phases one as training and another one as a testing phase.

Relatable work

We selected variables for water quality data and will consider some elements like:

1. *The pH level value: Assessing the balance of acids and bases of water requires consideration of this crucial trait. It also serves as a gauge for how acidic or alkaline the water is.*

2. *Hardness: Salts of the minerals magnesium and calcium are the main contributors to hardness. These salts are released by the geologic formations that water moves through. In its most basic form, the amount of time that water is exposed to a material that causes hardness defines how hard the water is. The original definition of hardness was the capacity of water to precipitate soap as a result of calcium and magnesium ions.*

3. *Solids (Total dissolved solids - TDS): Water can dissolve a wide variety of organic and inorganic minerals or salts, including potassium, calcium, and sodium as well as bicarbonates, chlorides, magnesium, calcium, and sulphates. The color and scent of the water were both subdued by these minerals. This is an important consideration while using water.*

4. *Chloramines: In public water systems, chlorine and chloramine are the two main disinfectants. The process of treating drinking water with chlorine and ammonia most frequently results in the chemical chloramine formation.*

◆ **Conductivity:** Pure water is a good insulator, not a good conductor of electricity. The electric conductivity of water is enhanced by an increase in ion concentration. Electrical conductivity is often based on the solids that have been dissolved content of the solution of water. water consumption..

IV. SYSTEM ANALYSIS

Functional Requirements

1. **Python** - Python is a high-level, interpretive programming language. It plays an enormous value on the understanding of the code.

2. **Scikit-learn** - **Scikit-learn** (formerly **scikits-**learn is a free software machine-learning toolkit for e-Python programming. It is also known as a program such as It includes different support vectors, random forests, gradient boosting regression, sorting, and clubs techniques.

Software Requirements

- OS: Windows
- Python Idle: 1. Python 2.7x and above

2. Pycharm IDE

• Installing Configuration Tools and pip for 3.6x or higher

Hardware Requirements

- RAM: 8GB
- Processor: Intel i3 7th Generation and Above
- Hard Disk: 500GB Minimum

Performance Requirements

Performance is measured against the output provided by the application. Requirements Specifications play an important role in the analysis of the system. Only if there are requirements and given the correct specifications, it is possible to design a system that fits a required environment. It is mainly left to the users of the existing system to give its requirements specifications because it is the person who ultimately uses the system. That is because the requirements need to be known in the early stages so that the system can be designed according to these requirements. It is very difficult to change the system once to design a system that is designed but does not justify it as user requirements are useless.

Feasible Study

Systems for making recommendations are created using machine learning techniques. Unattended machine learning and supervised machine learning are the two categories of machine learning. In supervised machine learning methods, the machine is provided with enormous data sets and conclusions in order to train it. To choose the best features for this project, various classification methods were employed as such as Naive Bayes, the Decision Tree, and the Random Forest approach.

V. SYSTEM DESIGN

A. SYSTEM ARCHITECTURE

The principles of software infrastructure specify how software is created and evolved. The program system's architecture specifies the organization and arrangement of the software system. Additionally, it discusses the links between the various software system components, levels of abstraction, and other features. Architecture can be used to specify the goals of a task or to direct the building of a fresh system.

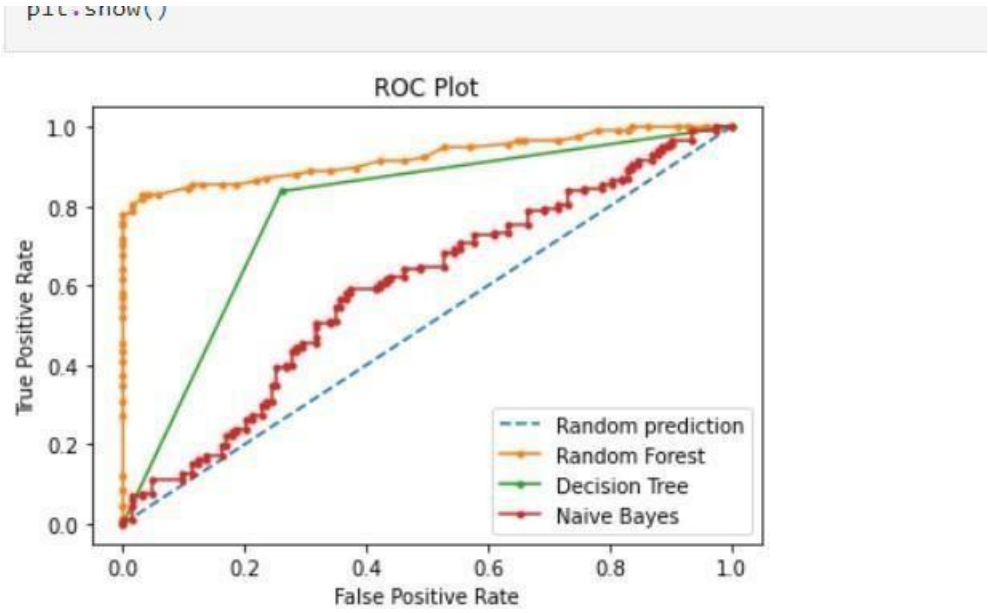

Fig1: architecture diagram

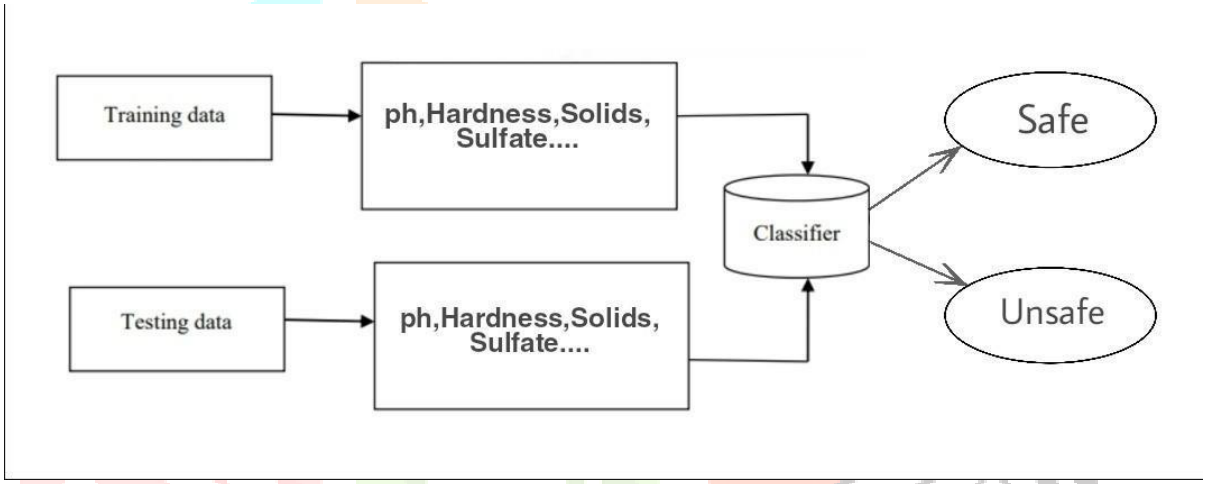

Fig2: data flow

B.UML DIAGRAM

Unified Modeling Language (UML) is a standard graphical language used to model software systems. It provides a set of diagrams that can be used to visualize, design, and document software systems.

UML diagrams can also be used to document and communicate the design of the system to other developers or stakeholders. This can help ensure that everyone involved in the development process has a clear understanding of the system's structure and behavior, and can help facilitate collaboration and feedback.

Use Case Diagram:- The ever-changing behavior of a system is represented by a use case diagram. It incorporates scenarios, individuals, and their interactions to encapsulate the functionality of the system. It simulates the duties, services, and operations performed by a system or application subsystem. It shows a system's high-level capability and also describes how a user interacts with a system.

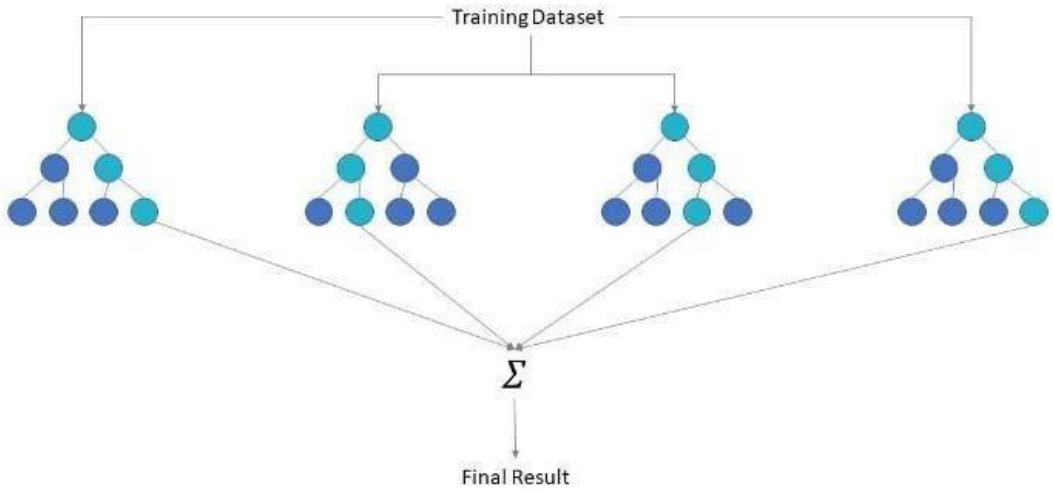

Fig3: use case diagram

Class Diagram:-

A Class diagram is a UML diagram that depicts the static structure of a software system, including the classes, their attributes and operations, and the relationships among them. This diagram serves as a visual representation of the system's components and their interconnections, enabling developers to better understand the system's architecture and design. By identifying the classes and their properties, developers can plan and implement the system's functionality effectively. The Class diagram typically consists of multiple classes, each with their own attributes and operations, which are interlinked to represent the relationships between the classes.

Fig4: class diagram

Sequence Diagram:-

A sequence diagram, which is also known as a scenario diagram, depicts the transmission of messages through the system. It aids in imagining a variety of dynamic events. It depicts communication between any two lifelines as a time-ordered series of events, implying that these lifelines participated in the run time. In UML, the message flow is represented by a straight dotted line that crosses the bottom of the page, whereas the lifeline is symbolized by a vertical bar. Both branches and iterations are incorporated. has and will frequently be accompanied by other kinds of diagrams as well. A circle or an ellipsis is used to denote each usage case. Stick figures of the actors are frequently used to depict them.

Fig5: sequence diagram

VI. RESULTS:

Fig6: comparative analysis

Fig7: opening value

Fig8: closing value

VII. CONCLUSION

- .In this paper, the performance of Machine learning including Decision trees, Random Fores,t, and Naïve Bayes was evaluated to predict the water quality of drinking water. To this end,d most dataset-related well-known components, such as pH, SO4, Na, Ca, Cl, Mg, HCO3, etc., were collected. Results indicated that the applied models have suitable performance for predicting water quality components, however, the best performance was related to the Random Forest. Results of Naïve bBayesindicated that its accuracy is acceptable for practical purposes. The lowest accuracy of models was related to Decision trees. The index of results of applied models shows that all three models slightly overestimate. Although the accuracy of model Naïve Bayes is less than that of model Random Forest, their DDR indices are close together. Furthermore, a comparison of the performance of applied models indicated that the outcomes of Random Forest and Naïve Bayes models were more reliable in comparison with the Decisiontree.

Further improvements: Machine learning (ML) has an array of applications, and in the near future, it will expand even further into a number of industries, including biometrics, banking, social media, facial and voice recognition, and online identification of fraudulent activity. Machine learning is the process of automatically deriving business value-generating insights from data. The application of algorithms for learning extends beyond the world of investments. Instead, it is growing in all industries, including banking and finance, IT, media & gaming, leisure, and automobile manufacturing. There are several industry sectors where academics are aiming to revolutionize the world for decades to c"Water quality assessment using machine learning techniques: A review" by F. Begum et al. in Water Science and Technology: Water Supply (2019). This paper provides an overview of various machine learning techniques used for water quality assessment and their applications"Prediction of drinking water quality parameters using machine learning algorithms" by M. Arslan et al. in Environmental Monitoring and Assessment (2018). The study explores the use of machine learning algorithms such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM) to predict the quality of drinking water.

REFERENCES

[1]Water quality assessment using machine learning techniques: A review" by F. Begum et al. in Water Science and Technology: Water
Supply (2019). This paper provides an overview of various machine learning techniques used for water quality assessment and their applications

[2] Prediction of drinking water quality parameters using machine learning algorithms" by M. Arslan et al. in Environmental Monitoring and Assessment (2018). The study explores the use of machine learning algorithms such as Artificial Neural Networks (ANN) and Support Vector Machines (SVM) to predict the quality of drinking water.

[3] Machine learning approach for water quality prediction in drinking water treatment plants" by M. Kaya et al. in Water Science and Technology: Water Supply (2020). The study uses machine learning techniques to predict the quality of water in a drinking water treatment plant.

[4] A comparative study of machine learning algorithms for predicting water quality parameters" by S. Singh et al. in Environmental
Monitoring and Assessment (2019). The study compares the performance of various machine learning algorithms in predicting water quality parameters.

[5] Machine learning approach for prediction of water quality index in a river basin" by S. Pal et al. in Water Science and Technology: Water Supply (2020). The study uses machine learning algorithms to predict the Water Quality Index (WQI) of a river basin.