**Project Requirement and Specification**

**On**

**WATER QUALITY MONITOR –(SAFE AND UNSAFE TO DRINK)**

**FOR DISEASE PREDICTION**

**(CSE V Semester Mini project)**

**2021-2022**

****

Submitted to: Submitted by:

**Department of Computer                Dhruv Singhal**

**application for the**

**degree of BTECH (CSE)                 Class rollno.-17**

**Roll. No.: 2014388**

**Sem-5th**

Sec- ‘H’

Btech (CSE)

**----------------------------Contents--------------------------**

* **Abstract**
* **Project Implementation**
* **Proposed system**
* **Implementation Details**
* **Result**
* **Conclusion**
* **References**

**Abstract**

Waterborne diseases are one of the leading causes behind human deaths across world. According to World Health Organization (WHO), waterborne Diarrheal diseases are accountable for 2 million deaths per year, ,annually about 37.7 million people are impacted by waterborne diseases, out of which 1.5 million kids die of Diarrheal diseases.The varieties of waterborne disease are caused by water contaminated with viruses, bacteria, metals, toxins and other chemical contaminants.

Machine Learning has been proved to be helpful in automation of disease diagnosis. Considering the harmful impact of waterborne diseases on human life, there is need for applying ML in early prediction of such diseases. Waterborne disease outbreaks are often reported in areas with poor water quality .Gastrointestinal diseases like cholera, typhoid, etc. arise due to microbial contaminants. On the other hand, water related diseases like diarrhea, alkalosis, fluorosis, bladder cancer and kidney diseases rise out of suspended metals and chemical contaminants in water. Changes in water pH levels can also significantly affect human health. Numerous efforts have been applied by analysts to create or utilize huge information investigation models and machine learning models for precise water quality evaluation .

In this regard, the main **motivation** in this study is to propose the method using machine learning algorithm for the efficient prediction of water related disease based on water quality parameters like pH, Temperature, Turbidity, Total Dissolved Solids (TDS), Dissolved Oxygen, etc. The system involves sensors like pH sensor, TDS sensor, Turbidity sensor, temperature sensor and conductivity sensor to continuously monitor water conditions.

**2. Project Implementation**

1. The first step is to collect water-related data with the help of sensors. Or if you don’t have experimental sensors then you can find pre-collected data also which contains attributes like temperature, TDS, Turbidity, pH and conductivity. The data is easily available on many websites.
2. After that collected data should be stored on the cloud. This is only for the monitoring and storing of the data. You can use a cloud service for the storage of the data.
3. Then you have to train and test the data, for that you need an accurate dataset. Water quality-related dataset can easily be available on any health-related website like WHO or **west Bengal pollution control board** or you can simply find it on Kaggle.
4. For training and testing of the data, we are using a gradient boosting classifier machine learning algorithm.
5. We will find that our given water sample is fit to drink or not by 0 and 1.
6. Scikit-learn is also known as learn which is a machine learning library developed in python language. It is used for classification, Support vector machine, regression, gradient boosting classifier and many more algorithms.
7. The algorithm helps for the prediction of the disease.
8. The next step is to show the result …

**3. Proposed System**

# A. Objectives

Following are the objectives of proposed system:

1. To produce an innovative, intelligent, system to predict waterborne diseases from water quality parameters, and to find whether the water is fit to drink or not.
2. To check water quality, predict disease based on it and inform people about possible disease so as to take appropriate preventive measures.
3. To check given water sample is fit to drink or not based on data set values.

**B. System Overview**

# Dataset

The system has a primary and a secondary data source. **Collection of data about variety of water related diseases** will require study and surveys to be conducted in different weather conditions and at least throughout a year. Since this is time consuming, we used secondary data source to train machine learning classifier model. The water parameter data was collected from the **West Bengal Pollution Control Board’s** Water Quality Information System . We preferred this data for machine learning task due to its authenticate and accurate nature.

**System Architecture:**

The proposed work put forth an embedded system for water quality monitoring and disease prediction. The system comprises of hardware components like ***sensors, microcontroller units (MCUs) and display, etc. interfaced with each other***.

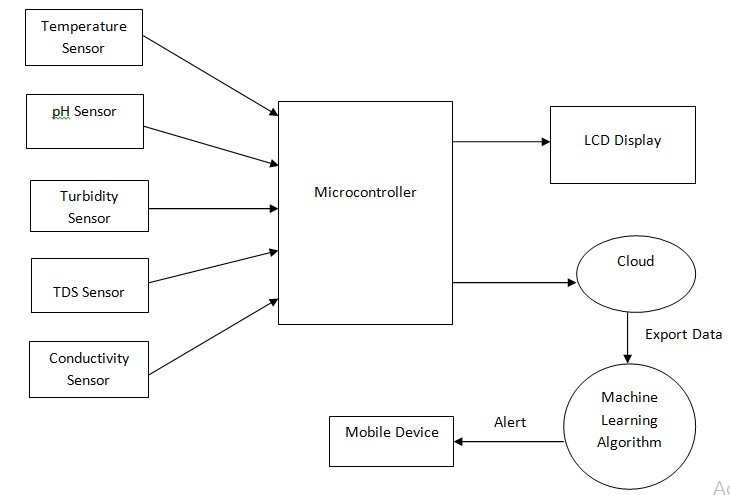


Fig 1: System Architecture

The proposed system for water quality monitoring for disease prediction involves ***Temperature sensor, pH sensor, Turbidity sensor, TDS sensor and conductivity sensor*** to continuously assess water quality. All the sensors are interfaced with central microcontroller ,which performs the task to **fetch sensed data from sensors**. Collected data is transferred to cloud server for storage. At the same time it is displayed on a scrolling display interfaced with system for visual cue. The data is exported from cloud storage and fed to **Machine Learning Classifier** for the disease prediction task. The Machine Learning classifier is previously trained using secondary data on water quality parameters. Finally we predict thw water sample is drinkable or not and disease causing or not.

**C. Technical Descriptions (IOT BASED SENSORS USING)**

# Temperature Sensor

The DS18B20 underwater temperature sensor delivers 9 - 12-bit temperature sense output which indicates the temperature of the water.

# E201-C-9 pH sensor

A pH Meter or a pH probe refers to a systematic tool that monitors the hydrogen-ion concentration of a solution, representing whether it is acidic or alkaline.

# Turbidity Sensor

Turbidity relates to cloudiness of water.

# TDS Sensor

Total Dissolved Solids ( TDS ) refers to the extent of soluble solids present in water.

# Conductivity Sensor

Conductivity is nothing but the capacity of the material to convey electric current. It can be considered as the reciprocal of the resistance. Conductivity is a significant factor of water quality. It can indicate the extent of electrolytes existing in water.

# DOT Matrix Display

A P10 dot-matrix display is an electric alphanumeric display tool that displays information requiring a modest alphanumeric (and/or graphic) display device of restricted resolution.

**4. Implementation Details**

Ten parameters chosen to train disease predictor include Ammonia, Chloride, Fluoride, Nitrate, pH, TDS, TSS, Turbidity, Zinc, and Total Coliform. These parameters are chosen depending on literature review, study of causes of various water related diseases and relevance of parameter with respect to disease outbreak [3]. Diseases included in study are Diarrhea Fluorosis, Convulsions, Bladder Cancer, Kidney Diseases, Metabolic Alkalosis, Methemoglobinemia, Dental Corrosion, and other Gastrointestinal Diseases like Cholera and Typhoid.

Table 1: Water Quality Parameters and associated Diseases

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SR. NO.** | **PARAMETER** | **UNIT** | **MAXIMUM**  **PERMISSIBLE LIMIT** | **ASSOCIATED DISEASE** |
| 1 | TOTAL COLIFORM | PPM | 0 | GASTROINTESTINAL DISEASES  LIKE CHOLERA, DIARRHEA, HEPATITIS |
| 2 | TURBIDITY | NTU | 3 | GASTROINTESTINAL DISEASES |
| 3 | AMMONIA | PPM | 0.5 | CONVULSIONS |
| 4 | CHLORIDES | PPM | 250 | BLADDER CANCER |
| 5 | FLUORIDES | PPM | 1.5 | FLUOROSIS |
| 6 | NITRATE | PPM | 1.5 | METHEMOGLOBINEMIA |
| 7 | TOTAL DISSOLVED SOLIDS (TDS) | PPM | 300 – 500 | KIDNEY DISEASES |
| 8 | TOTAL SUSPENDED SOLIDS (TSS) | PPM | 50 | KIDNEY DISEASES |
| 9 | PH   1. ACIDIC 2. ALKALINE | - | 6.5 – 7  7 – 8.5 | DENTAL CORROSION  METABOLIC ALKALOSIS |
| 10 | ZINC | PPM | 5 | DIARRHEA |

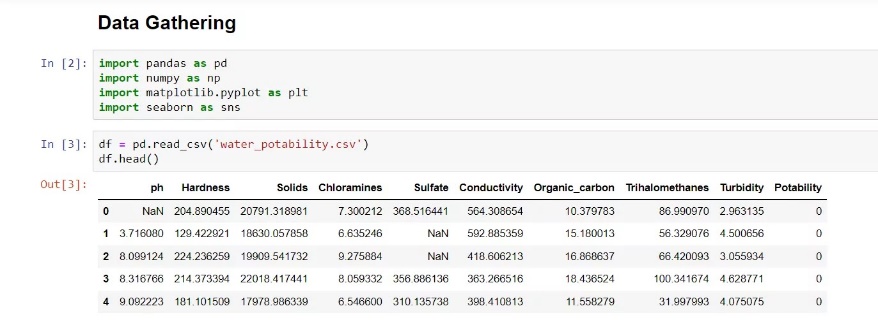
# B. Gradient Boosting Algorithm

Gradient Boosting is an ensemble based machine learning (ML) algorithm that involves predicting output on the basis of given sample inputs and combining a group of weak machine learning models in order to get a better and stronger learning model. It tells us of the disease type ,after comparison.

2 processes -

1. Train - training data is the data you use to train a machine learning algorithm or model
2. Test- prediction is done on the basis of test data

**SOME SCREENSHOTS –**



# 

# 

# 

# 

# 

# Results-

This Fig represents the classification report of trained machine learning+IOT model which informs about various performance metrics on individual prediction classes. For all disease classes, precision, recall and F1score was calculated based on test dataset. The average of these individual values gives an accuracy of 0.95 to the classifier.

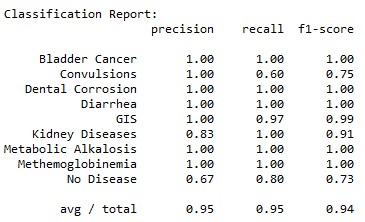
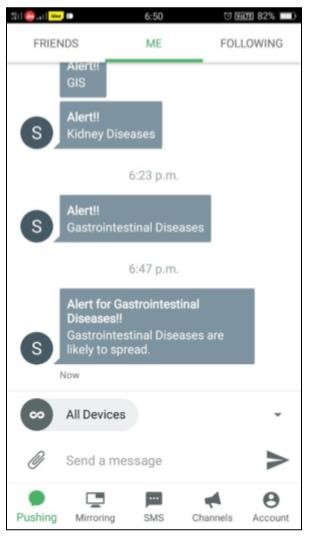


Fig : Classification Report of Gradient Boosting Classifier

***The model was used to predict possible waterborne and water related disease***. *The disease predicted was conveyed to mobile devices in the form of push notification*.



# CODE USED FOR IMPLEMENTATION ( DATA GATHERING and DATA ANALYSIS)

## **Data Gathering-**

//importing important lib to visualise data set

**import** pandas **as** pd // lib for data manipulation and analysis

**import** numpy **as** np // lib used for working with array

**import** matplotlib.pyplot **as** plt //to create animated and static visualisation

**import** seaborn **as** sns // used for data analysis

**df = pd.read\_csv('water\_potability.csv')** // loading data set info using

read.csv()(panda)

**Data Analysis-**

**df.shape** //to check the size of the data set(3276 rows,10 column)

**df.isnull().sum**() //to check null value is present or not in data set

**df.info()** // to check info of data set like type, column name,etc.

**df.describe()** //to describe the data set like mean,min,max,etc. of each column

**df.fillna(df.mean(), inplace=True)** //fill null value to every column using **df.isnull().sum()** mean to accept the new values for determination

**df.Potability.value\_counts()//**display the potability and check the value count

**sns.countplot(df['Potability'])** //to visualise the potability

**plt.show()**

**sns.distplot(df['ph'])** // to display the ph value calculated by sensor

**plt.show()**

**df.hist(figsize=(14,14))** // to display the entire data set (ph ,hardness **plt.show()** solids, chloramines, sulfate, turbidity, etc)to …………………………………………………………………………explore it using hist method .

**plt.figure(figsize=(13,8))**

**sns.heatmap(df.corr(),annot=True,cmap='terrain')**

**plt.show()** // to visualise the co-relation of all features using heatmap()

**df.boxplot(figsize=(14,7))** // to display outliers using boxplot()to find impurity of water(if solid is high ,water is unsafe and vice-versa)

**X = df.drop('Potability',axis=1)**

**Y= df['Potability']**

**from** sklearn.model\_selection **import** train\_test\_split

X\_train, X\_test, Y\_train, Y\_test **=** train\_test\_split(X,Y, test\_size**=** 0.2, random\_state**=**101,shuffle**=True**) **//to split data set into train and tree**

## ***Train -Tree Classifier and check accuracy***

**from** sklearn.tree **import** DecisionTreeClassifier

**from** sklearn.metrics **import** **accuracy\_score,confusion\_matrix,classification\_report**

**dt=DecisionTreeClassifier(criterion= 'gini', min\_samples\_split= 10, splitter= 'best')** //training the model using decision tree classifier

**dt.fit(X\_train,Y\_train**) // fit means learning from training data

**prediction=dt.predict(X\_test)** //checking performance of model using test data

**print(f"Accuracy Score = {accuracy\_score(Y\_test,prediction)\*100}")//**predicting accuracy

**print(f"Confusion Matrix =\n {confusion\_matrix(Y\_test,prediction)}")**

**print(f"Classification Report =\n {classification\_report(Y\_test,prediction)}")**

**res = dt.predict([[5.735724,158.318741,25363.016594,7.728601,377.543291,568.304671,13.626624,75.952337,4.732954]])[0**] //for predicting single row

res

output-1 (for result)

# 6. Conclusion

The water quality monitoring system for disease prediction is successfully designed and implemented.

# 

# 7. References

# http://www.wbpcb.gov.in/

* https://www.who.int/sustainabledevelopment/housing/healthrisks/waterborne-disease/en/