KARNATAKA LAW SOCIETY’S

GOGTE INSTITUTE OF TECHNOLOGY

UDYAMBAG, BELAGAVI-590008

(An Autonomous Institution under Visvesvaraya Technological University, Belagavi)

**(APPROVED BY AICTE, NEW DELHI)**



*Lab Project Report On*

***Water Quality Analysis***

*Submitted in the partial fulfillment for the academic requirement of*

***6th Semester B.E. In***

***Computer Science Engineering***

***Submitted by***

|  |  |
| --- | --- |
| **NAME** | **USN** |
| CHIDAMBAR INAMDAR | 2GI19CS035 |
| DEEPAK KUMBLE | 2GI19CS038 |
| GOURAV SHANBHAG | 2GI19CS046 |
| IRFAN KAMATE | 2GI19CS052 |

#### Under the guidance of

**Prof. Gajendra C. Deshpande**

#### Dept. of CSE, KLS GIT

**Academic Year 2021-2022 (even semester)**

### Karnataka Law Society’s

**GOGTE INSTITUTE OF TECHNOLOGY**

### Udyambag, Belagavi -590008 Karnataka, India.

**Department of Computer Science and Engineering**



Certificate

This is to certify that the Lab Project Report on “Water Quality Analysis” carried out by students **Chidambar Inamdar, Deepak N Kumble, Gourav Shanbhag, Irfan Kamate** bearing **USNs: 2GI19CS035, 2GI19CS038, 2GI19CS046, 2GI19CS052** is submitted in partial fulfilment of the requirements for 6th semester B.E. in **COMPUTER SCIENCE AND ENGINEERING,** Visvesvaraya Technological University, Belagavi. It is certified that all corrections/ suggestions indicated have been incorporated in the report. The course project report has been approved as it satisfies the academic requirements prescribed for the said degree.

|  |  |
| --- | --- |
| Date: 02/06/2022 | Signature of Guide |
|  | Prof. Gajendra C Deshpande |

KLS Gogte Institute of Technology, Belagavi

Place: Belagavi

Name of the Examiners Signature of the Examiners

1. 1.

# TABLE OF CONTENTS

|  |  |  |
| --- | --- | --- |
| **S No.** | **Content** | **Page** |
| 1 | Problem Statement and Introduction | 4 |
| 2 | Factors affecting potability | 5-6 |
| 3 | Libraries used | 7 |
| 4 | Program & Contents of water\_potability.csv | 8-10 |
| 5 | Output | 11-15 |
| 6 | Conclusion and References | 16 |

# PROBLEM STATEMENT

Water Quality Analysis with the help of machine learning in Python. To design and implement water quality prediction using machine learning techniques. In this technique, our model predicts whether the water is safe to drink or not using some parameters like Ph value, conductivity, hardness, etc.

# INTRODUCTION

Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level. In some regions, it has been shown that investments in water supply and sanitation can yield a net economic benefit, since the reductions in adverse health effects and health care costs outweigh the costs of undertaking the interventions.

Water quality has a direct impact on public health and the environment. Water is used for various practices, such as drinking, agriculture, and industry. Recently, development of water sports and entertainment has greatly helped to attract tourists

One of the main areas of research in machine learning is the analysis of water quality. It is also known as water potability analysis because our task here is to understand all the factors that affect water potability and train a machine learning model that can classify whether a specific water sample is safe or unfit for consumption.

For the water quality analysis task, Kaggle dataset is used that contains data on all of the major factors that affect the potability of water. All of the factors that affect water quality are very important, so we need to briefly explore each feature of this dataset before training a machine learning model to predict whether a water sample is safe or unsuitable for consumption.

# FACTORS AFFECTING POTABILITY

The water\_potability.csv file contains water quality metrics for 3276 different water bodies. The various factors that affect the potability of water are listed below-

**1. pH value:**

PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status. WHO has recommended a maximum permissible limit of pH from 6.5 to 8.5. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards.

**2. Hardness:**

Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.

**3. Solids (Total dissolved solids - TDS):**

Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals produced an unwanted taste and diluted color in the appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized. The Desired limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which is prescribed for drinking purposes.

**4. Chloramines:**

Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.

**5. Sulfate:**

Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. The principal commercial use of sulfate is in the chemical industry. Sulfate concentration in seawater is about 2,700 milligrams per liter (mg/L). It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.

**6. Conductivity:**

Pure water is not a good conductor of electric current rather it's a good insulator. Increase in ions concentration enhances the electrical conductivity of water. Generally, the amount of dissolved solids in water determines the electrical conductivity. Electrical conductivity (EC) actually measures the ionic process of a solution that enables it to transmit current. According to WHO standards, EC value should not exceed 400 μS/cm.

**7. Organic carbon:**

Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to the US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is used for treatment.

**8. Trihalomethanes:**

THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water.

**9. Turbidity:**

The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for Wonda Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.

**10. Potability:**

Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable. (0) Water is not safe to drink and (1) Water is safe to drink.

# LIBRARIES USED

**NumPy-** NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed.

NumPy is a Python package. It stands for ‘Numerical Python’. It is a library consisting of multidimensional array objects and a collection of routines for processing of array.

**Pandas-** Pandas is defined as an open-source library that provides high-performance data manipulation in Python. The name of Pandas is derived from the word Panel Data, which means an Econometrics from Multidimensional data. It is used for data analysis in Python. Data analysis requires lots of processing, such as restructuring, cleaning or merging, etc. There are different tools are available for fast data processing, such as NumPy, SciPy, Cython, and Panda. But we prefer Pandas because working with Pandas is fast, simple and more expressive than other tools.

**Matplotlib -** Most of the Matplotlib utilities lies under the pyplot submodule.

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

* Create publication quality plots.
* Make interactive figures that can zoom, pan, update.
* Customize visual style and layout.
* Export to many file formats.
* Embed in JupyterLab and Graphical User Interfaces.
* Use a rich array of third-party packages built on Matplotlib.

**Seaborn -** An open-source Python library built on top of matplotlib. It is used for data visualization and exploratory data analysis. Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

**Plotly-** provides online graphing, analytics, and statistics tools for individuals and collaboration, as well as scientific graphing libraries for Python, R, MATLAB,Arduino, and REST. The plotly Python library is an interactive, open-source plotting library that supports over 40 unique chart types covering a wide range of statistical, financial, geographic, scientific, and 3-dimensional use-cases.

# SOURCE CODE

# import matplotlib.pyplot as plt

# import pandas as pd

# import seaborn as sns

# import numpy as np

# data = pd.read\_csv(r"C:\Users\Kumble\Workspace\water\_potability.csv")

# data.head()

# data = data.dropna()

# data.isnull().sum()

# plt.figure(figsize=(15, 10))

# sns.countplot(data.Potability)

# plt.title("Distribution of Unsafe and Safe Water")

# plt.show()

# import plotly.express as px

# data = data

# figure = px.histogram(data, x = "ph", color = "Potability", title= "Factors Affecting Water Quality: PH")

# figure.show()

# figure = px.histogram(data, x = "Hardness", color = "Potability", title= "Factors Affecting Water Quality: Hardness")

# figure.show()

# figure = px.histogram(data, x = "Solids", color = "Potability", title= "Factors Affecting Water Quality: Solids")

# figure.show()

# figure = px.histogram(data, x = "Chloramines", color = "Potability", title= "Factors Affecting Water Quality: Chloramines")

# figure. show()

# figure = px.histogram(data, x = "Sulfate", color = "Potability", title= "Factors Affecting Water Quality: Sulfate")

# figure.show()

# figure = px.histogram(data, x = "Organic\_carbon", color = "Potability", title= "Factors Affecting Water Quality: Organic Carbon")

# figure.show()

# figure = px.histogram(data, x = "Trihalomethanes", color = "Potability", title= "Factors Affecting Water Quality: Trihalomethanes")

# figure.show()

# figure = px.histogram(data, x = "Turbidity", color = "Potability", title= "Factors Affecting Water Quality: Turbidity")

# figure.show()

# CONTENTS OF WATER\_POTABILITY.CSV

# 

# OUTPUT

# water quality analysis: ph value

# hardness

# water quality analysis: solids

# chloramines

# water quality analysis: sulfate

# conductivity

# water quality analysis: organic carbon

# Trihalomethanes

# water quality analysis: Turbidity

# water quality dataset distribution

# Till now we explored all the features that affect water quality. Now, the next step is to train a machine learning model for the task of water quality analysis using Python. For this task, we have used PyCaret library in Python. It can be easily installed in the system as follows: pip install pycaret

# Before training a machine learning model, we see the correlation of all the features with respect to the Potability column in the dataset:

# correlation = data. corr()

# correlation["ph"]. sort\_values(ascending=False)

# 

# Now we have to check which machine learning algorithm is best for this dataset by using the PyCaret library in Python:

# from pycaret.classification import \*

# clf = setup(data, target = "Potability", silent = True, session\_id = 786)

# compare\_models()

# Model Selection for water quality analysis

# According to the above result, the random forecast classification algorithm is best for training a machine learning model for the task of water quality analysis. So, let’s train the model and examine its predictions:

# model = create\_model("rf")

# predict = predict\_model(model, data=data)

# predict.head()

# water quality analysis with machine learning

# CONCLUSION

# So, this is how you can analyze the quality of water and train a machine learning model to classify safe and unsafe water for drinking. Access to safe drinking water is one of the essential needs of all human beings. From a legal point of view, access to drinking water is one of the fundamental human rights. Many factors affect water quality, it is also one of the major research areas in machine learning.

# REFERENCES

# <https://www.kaggle.com/datasets/adityakadiwal/water-potability>

# <https://www.datascience2000.in/2021/10/water-quality-prediction-using-machine.html>

# <https://thecleverprogrammer.com/2021/08/19/water-quality-analysis/>