WATER QUALITY PREDICTION

A Course Project report submitted

in partial fulfillment of requirement for the award of degree

BACHELOR OF TECHNOLOGY

in

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

by

CH. RISHIKA	**(2103A52010)**
CH. ALEKHYA	**(2103A52011)**
G. SANJANA	**(2103A52018)**

Under the guidance of

Mr. RAMESH DADI

Assistant Professor, Department of CSE.

SR UNIVERSITY

Department of Computer Science and Artificial Intelligence

Department of Computer Science and Artificial Intelligence

CERTIFICATE

This is to certify that project entitled **"WATER QUALITY PREDICTION**" is the bonafied work carried out by **CH. RISHIKA, CH. ALEKHYA, G. SANJANA** as a Course Project for the partial fulfillment to award the degree **BACHELOR OF TECHNOLOGY** in **ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING** during the academic year 2022-2023 under our guidance and Supervision.

Mr. Ramesh Dadi
Asst. Professor,
S R University,
Ananthasagar, Warangal.

Dr. M. Sheshikala
Assoc. Prof. & HOD (CSE),
S R University,
Ananthasagar, Warangal.

ACKNOWLEDGEMENT

We express our thanks to Course co-coordinator **Mr. Ramesh Dadi, Asst. Prof.** for guiding us from the beginning through the end of the Course Project. We express our gratitude to Head of the department CS&AI, **Dr. M. Sheshikala, Associate Professor** for encouragement, support and insightful suggestions. We truly value their consistent feedback on our progress, which was always constructive and encouraging and ultimately drove us to the right direction.

We wish to take this opportunity to express our sincere gratitude and deep sense of respect to our beloved Dean, School of Computer Science and Artificial Intelligence, **Dr C. V. Guru Rao**, for his continuous support and guidance to complete this project in the institute.

Finally, we express our thanks to all the teaching and non-teaching staff of the department for their suggestions and timely support.

ABSTRACT

This study investigates the performance of artificial intelligence techniques including artificial neural network (ANN), group method of data handling (GMDH) and support vector machine (SVM) for predicting water quality.To develop the ANN and SVM,different types of transfer and kernel functions were tested,respectively.Reviewing the results of ANN and SVM indicated that both models have suitable performance for predicting water potability components.During the process of development of ANN and SVM.Comparison of outcomes of GMDH model with other applied models shows that although this model has acceptable performance for predicting the components of water quality,its accuracy is slightly less than ANN and SVM.The evaluation of the accuracy of the applied models according to the error indexes declared that SVM was the most accurate model.Examining the results of the models showed that all of them had some over-estimation properties.By evaluating the results of the models based on the DDR index,it was found that the lowest DDR value was related to the performance of the SVM model.

Table of Contents

INTRODUCTION

1.1 OVERVIEW

Water potability prediction based on image dataset analysis is not a straightforward task. This is because water potability is determined by a wide range of chemical and physical properties of water that cannot be directly inferred from images. However, some indirect features that can be extracted from images can provide some useful insights on water potability.

One possible approach is to use image analysis techniques to extract color, texture, and shape features from water images, and then use these features to train a machine learning model that predicts water potability. For example, color features such as hue, saturation, and brightness can be used to estimate the concentration of dissolved particles in water. Texture features such as coarseness, contrast, and entropy can provide information about the uniformity of water composition. Shape features such as roundness, elongation, and compactness can provide information about the geometry of water particles and their aggregation.

Another approach is to use image analysis to identify water impurities such as suspended particles, algae, and organic matter, and then use this information to predict water potability. For example, the presence of high concentrations of suspended particles in water can be an indication of poor water quality, while the presence of organic matter can indicate potential health hazards.

However, it is important to note that image analysis techniques alone may not be sufficient to accurately predict water potability, as other factors such as temperature, pH, and mineral content can also affect water quality. Therefore, it is important to complement image analysis with other types of data, such as water quality measurements, to improve the accuracy of the prediction models.

In summary, while image analysis can provide some useful insights into water potability prediction, it should be complemented with other types of data and should be interpreted with caution due to the complexity of water quality determinants.

Water quality prediction refers to the use of models and algorithms to forecast the quality of water in particular body of water, such as a lake or river, based on various environmental and climatic factors. This is an important task for water resource managers and policymakers, as it allows them to anticipate potential water quality problems and take appropriate measures to mitigate them.

There are various approaches to water quality prediction, including statistical models, machine learning algorithms, and physical models. Some common variables that are used to predict water quality include temperature, dissolved oxygen, pH, turbidity, nutrients, and pollutants such as pesticides and heavy metals.

In order to develop an accurate water quality prediction model, it is important to gather and analyze historical data on water quality as well as current and future environmental conditions. This data can be collected through a variety of methods, including on-site monitoring, remote sensing, and modeling.

Overall, water quality prediction is a crucial aspect of water resource management, as it enables stakeholders to make informed decisions about how to allocate resources and protect water resources for future generations.

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 mainly checks or provides us the information whether the given sample image is of qualitavitve or not.

The water sample having qualitative water is classified as Potable Water and sample of water which are not qualitative are classified as Not Potable.

Potable Water tells us that the sample of water is safe to drink and Not Potable Water tells us that the given sample of water is not safe to drink.

One of mankind's biggest global challenges is access to potable water supply; its availability, quality and governance.

Currently some 1.5 billion people lack a safe water supply (COHRE, AAAS, SDC and UN-HABITAT, 2008: 8). This study looks at water sector reforms in Zimbabwe and South Africa with special focus on local potable water supply governance in the municipalities of Harare, Masvingo, Musina and Tshwane since the late 1990s.

Thus potable water supply governance addresses both public policy and practice regarding potable water supply issues.

1.2 PROBLEM STATEMENT

One of mankind's biggest global challenges is access to potable water supply; its availability, quality and governance. Currently some 1.5 billion people lack a safe water supply (COHRE, AAAS, SDC and UN-HABITAT, 2008: 8). This study looks at water sector reforms in Zimbabwe and South Africa with special focus on local potable water supply governance in the municipalities of Harare, Masvingo, Musina and Tshwane since the late 1990s.

Thus potable water supply governance addresses both public policy and practice regarding potable water supply issues. For Folifac (2007: 10), UNESCO (2006: 6), Mulder (2005: 1, 58), and Ashton et al (2001: xxvii) southern Africa (see Figure 1.1) faces severe and growing challenges in the governance of potable water supply. They indicate a range of factors responsible for this state of affairs, including population growth; the continuous depletion and pollution of water supplies; semi-arid conditions; anticipated trends of climate change in the face of global warming; successive droughts; lack of both administrative and political will; poverty and disparities in income; cultural and racial diversity; and the absence of scientific and technical knowledge, among others. In fact, the southern African region presents a complicated water resources governance scenario. The region has both natural and artificial challenges that intermingle to create an intricate and heterogeneous contextual framework. This framework demands knowledgeable and skilled water resources managers if there is to be long term sustainability of water resources in the region. Some of the technical and specialised issues that require expertise (both traditional/indigenous and contemporary) in handling them include:

- the impact of mining;
- agricultural activity;
- global warming;
- successive droughts;
- desertification;
- geological and hydrological issues;

• urbanisation and industrialisation;

• bulging population;

1.3 EXSITING SYSTEM

Water quality detection system using IoT (Internet of Things) mainly focuses to create a more ideal air pollution detection system while eliminating some disadvantages of previous systems. Libelium will be an advanced mobile Water remote sensor stage should improve remote water personal satisfaction following. Wasp motacillidae advanced mobile Water will be suitableness to potable water monitoring, compound spillage identification done rivers, remote estimation for swimming pools. It holds self-sufficient hubs that unite with the cloud to ongoing water control. The water caliber parameters measured incorporate pH, broken down oxygen (DO). Oxidation-diminishment possibility (ORP). Conductivity (salinity), turbidity, temperature. Furthermore, disintegrated ions. Wasp motacillidae might utilization cell division (3G WCDMA) Furthermore ZigBee connectivity will send data of the Cloud, What's more it obliges sunlight based boards that accuse those battery. The main advantage of this system is highly accurate and it can cover wide area. The drawback in this model is its cost. The opposite technobabble will be it utilization routine water caliber sensors to the reason for ongoing off chance versatile detection, ID number What's more cautioning technique and analysed it utilizing pilot-scale channel stream analyses the tried contaminants incorporate pesticide Also herbicides Furthermore inorganic compound exacerbates mercuric chloride What's more potassium ferricyanide)(), Second, the relative progressions computed starting with adaptively changed lingering chlorine estimations were quantitatively identified with contaminant- chlorine reactivity over drinking water. The drawback in this model is it should be highly maintained, so the cost increases. Alternate model will be around the Sensor-Based Water nature checking framework. The framework structural engineering comprises from claiming information following nodes, and base station What's more a remote station. Constantly on these stations would associated utilizing remote correspondence connection. Those information from hubs may be send of the build station Also information gathered Toward the build station for example, such that pH, turbidity, conductivity, and so forth. Is sent of the remote screening station.

1.4 PROPOSED SYSTEM

An outline of the grouping from claiming developments alternately movements of people or things included previously, an intricate framework alternately action. A graphical representational of a PC program for connection to its arrangement for works (as different starting with the information it processes).

Engineering Model: The iterative model will be a specific execution of a programming improvement life cycle that concentrates looking into an initial, rearranged implementation, which after that progressively additions that's only the tip of the iceberg multifaceted nature Also a more extensive characteristic situated until the last framework may be complete. At examining the iterative method, the idea of incremental improvement wills additionally frequently all the chance to be utilized generously furthermore interchangeably, which portrays the incremental alterations produced throughout the plan Furthermore usage for each new cycle.

1.5 DEFINE OBJECTIVES

The main objectives of this data analysis are….

- Drinking water sources;
- Aquatic life and its habitat;
- Wildlife and its habitat;
- Agriculture (livestock watering and irrigation);
- Recreational use and aesthetics; and
- Traditional, cultural, and social uses

1.6 OVERALL ARCHITECTURE

This research paper proposes an architecture for implementing a water quality monitoring system for the aquaculture industry. The system would enable monitoring of the water quality remotely via GSM. Conventional method used by aqua farms requires technical staff to visit ponds at designated time and perform manual testing on the water quality. Consequently, the technique consumes a lot of time and effort. This research project would focus on developing a prototype that can evaluate data collected through three criteria: Dissolved oxygen level, pH level, and temperature level. The system would also be able to send alert messages upon detecting degradation of water quality in the pond via SMS.

2. LITERATURE SURVEY

This study looks into the approaches that were used to help solve water quality challenges [1]. In most studies, traditional analyses in the laboratory and data analysis are two types of analyses and utilized to help determine the quality of water, but other studies apply machine learning approaches to help find an optimal solution to the water quality problem.

Consumers' health is being negatively impacted by poor drinking water quality. At least 2 billion individuals used feces-contaminated drinking water around the world, according to reports. Developing accurate decisions about the control and safeguarding of drinking water quality necessitates an awareness of the factors impacting its purity. Potable water quality is typically impacted by the source water's quality, how it is handled before being delivered, how it is distributed, how it is maintained, and how effectively it is filtered at residence. Furthermore, in rural areas and small municipalities, drinking water is frequently drawn straight from wells or retrieved unfiltered from rivers, lakes, and reservoirs. As a result, the purity of the source water is a significant factor affecting the quality of the drinking water. Many developing nations have achieved waterborne disease reduction and the development of safe water supplies a significant public health aim in recent years, and the situation has improved slightly. However, the situation is far from ideal, particularly in rural regions, and even marginally better conditions may be jeopardized by growing water consumption and reduced water availability as a result of population expansion and economic development. It is vital to use a practical and effective drinking water quality evaluation approach to get trustworthy results and make informed decisions.

Many water quality evaluation approaches have been proposed since Horton produced the first Water Quality Index (WQI) in the 1960s [2]. The two indices for determining the general state of drinking source water quality are straightforward, adaptable, and stable, with little sensitivity to input data. Similarly, to give water quality information, we employed the weighted arithmetic WQI approach. These WQIs convert a huge number of variables into a digital number and aid in the comprehension of water quality, making them the most widely used water quality assessment tool, despite significant flaws. Recent water quality assessments used matter element extension analysis (MEEA) and entropy TOPSIS in a wastewater irrigation area and a rapidly urbanizing area, respectively [2]. Both approaches are mathematical, but they are accurate in estimating overall water quality. These water quality evaluation methods, on the other hand, rely on water quality standards for classification. As a result, the most important thing is to create water quality guidelines.

All water utilities shall provide an appropriate, reliable source of greater drinkable water to consumers of price that is proportional to the demands of each water system. To fulfill this goal, its freshwater must be purified and supplied from the greatest source possible sufficiently in order to fulfill regulation and moisture levels sector standards. Consumer acceptance proved treatment procedures, and successful utility management should all be factored in determining the quality of drinking water. The water of high quality is characterized

as being free of harmful organisms and biological forms that may be aesthetically unattractive. It is clear and colorless, with no unpleasant odor or flavor. It is free of chemical concentrations that could be detrimental to the body, visually unappealing, or financially destructive. It is also noncorrosive and leaves no excessive or unwanted deposits on water-conveying structures such as pipes, tanks, and plumbing fittings.

Yafra Khan and Chai Soo See [3], in their paper, have used Artificial Neural Network and time series analysis to design a water quality prediction model. Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Regression Analysis have been used as a part of evaluating the model performance. Dao Nguyen Khoi et al. [4], in their paper, have used 12 machine learning models to estimate the quality of water. Model evaluation was done by using 2 statistics, R2 and RMSE. Umair Ahmed et al. [5] have used supervised machine learning algorithms to estimate the Water Quality Index (WQI). Saber Kouadri et al. [6] used 8 artificial intelligence algorithms to generate Water quality Index prediction. Evaluation of models was done using several statistical metrics, which includes correlation coefficient (R), mean absolute error (MAE), root mean square error (RMSE), relative absolute error (RAE), and root relative square error (RRSE). Jitha Nair and Vijaya M S [7] used various prediction models developed using machine learning and big data techniques using sensor networks.

Water quality was estimated using traditional machine learning techniques such as XGB (XGBoost), RF (Random Forest), DTC (Decision Tree), Adaptive Boosting (AdaBoost), and SVC, with XGB having the highest accuracy of 83% (XGBoost) . Their work is centered on water quality; all of the factors in the dataset, including hardness, sulfate, solid, trihalomethanes, pH, turbidity, solids, organic carbon, conductivity, are tested according to World Health Organization (WHO) standards.

S.no	Researcher names	Literature survey	Description
1	S. Sivakumar and R. Balasubramanian (2010)	Prediction of water portability using artificial neural networks	This study uses an artificial neural network (ANN) model to predict the portability of water based on several water quality parameters. The results show that the ANN model can accurately predict water portability.
2	A. R. Al-Zoubi and M. A. Al-Ani (2015)	Water quality prediction using support vector machines	This study uses support vector machines (SVMs) to predict water quality parameters, including portability. The results show that the SVM model can accurately predict water portability.
3	M. O. Ojo and O. F. Ojo (2016)	Predicting water portability in distribution systems using logistic regression and artificial neural networks	This study uses both logistic regression and ANN models to predict water portability in distribution systems. The results show that both models can accurately predict water portability.
4	A. R. Al-Zoubi et al. (2019)	Application of machine learning techniques for the prediction of water portability	This study compares the performance of SVM, ANN, and decision tree (DT) models in predicting water portability. T he results show that SVM performs the best, followed by ANN and DT.
5	R. D. S. Gamage et al. (2021)	Development of a decision support system for predicting water portability using machine learning algorithms	This study develops a decision support system (DSS) for predicting water portability using machine learning algorithms.

			The DSS integrates data from several sources, including water quality sensors, and provides accurate predictions of water portability.
6	R. Ahmad et al. (2004)	Predicting the bacterial quality of drinking water using artificial neural networks	This study uses an ANN model to predict the bacterial quality of drinking water based on water quality parameters such as pH, temperature, and turbidity. The results show that the ANN model can accurately predict bacterial quality.
7	T. Anitha and P. Thirumarimurugan (2015)	A comparative study of decision tree and artificial neural network for the prediction of water quality parameters	This study compares the performance of decision tree and ANN models in predicting water quality parameters, including portability. The results show that both models can accurately predict water portability, but the decision tree model has a slightly better performance.
8	F. A. Hegazy and A. M. Farid (2017)	Water quality prediction using regression analysis and artificial neural networks	This study uses both regression analysis and ANN models to predict water quality parameters, including portability. The results show that both models can accurately predict water portability, but the ANN model has a slightly better performance.
9	M. F. Doherty et al. (2018)	Comparison of support vector machines and artificial neural networks for predicting the quality of drinking water	This study compares the performance of SVM and ANN models in predicting water quality parameters, including portability. The results show that both models can

			accurately predict water portability, but the SVM model has a slightly better performance.
10	A. Pramanik et al. (2020)	Water quality prediction using machine learning	his review paper provides an overview of the various machine learning techniques used for water quality prediction, including ANN, SVM, decision trees, and regression analysis. The paper also discusses the advantages and limitations of each technique and provides recommendations for future research.

3.DATA PREPROCESSING

3.1 DESCRIBE DATASET

The above dataset contains images of Potable and Not Potable images in two separate folders.

The dataset contains a total of 1,062 images and two categories.

There are 571 images of Not Potable Water And 491 are of Potable Water images.

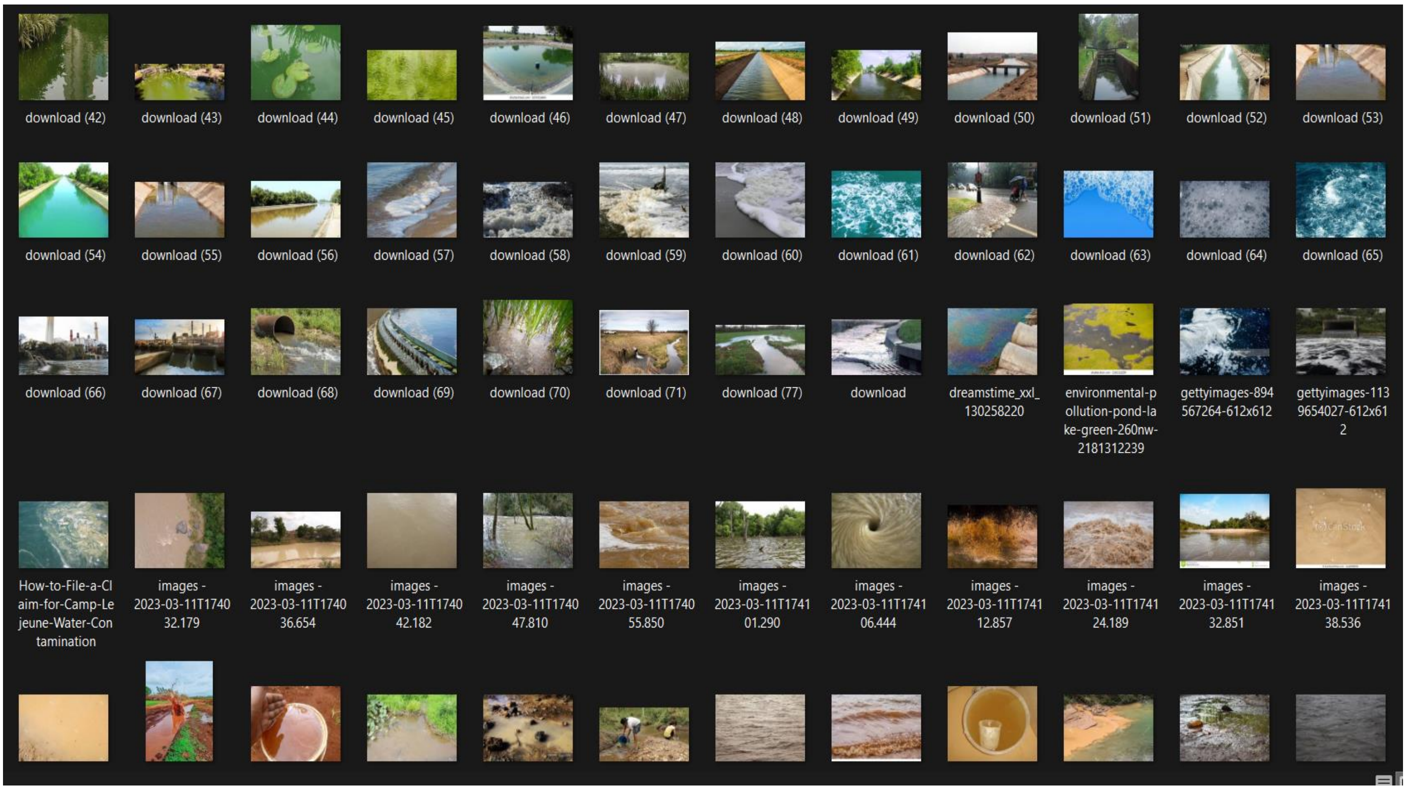
Collecting the required images from various sources of water availability.

To read the images we use open-cv library which is available in python.

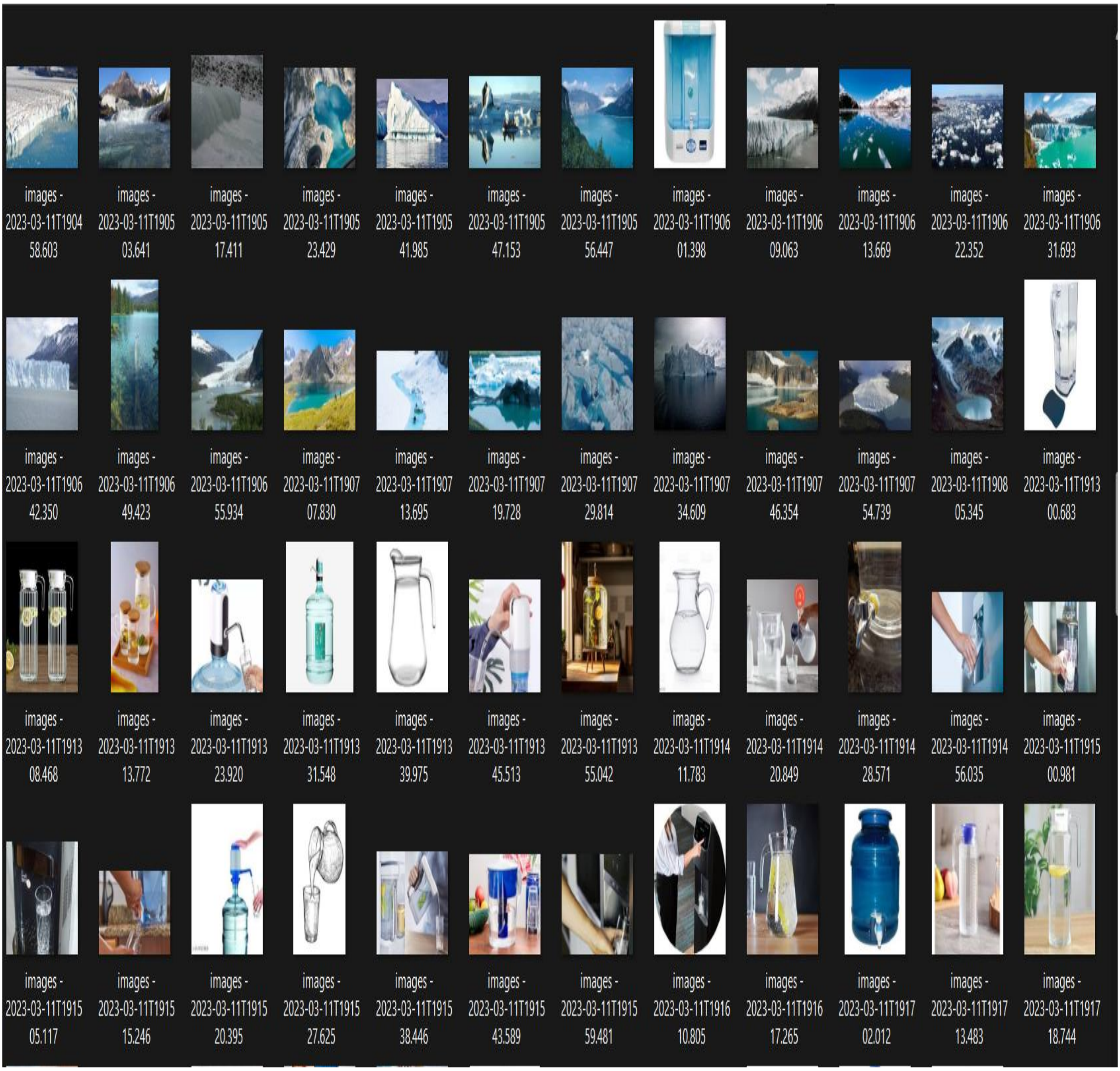
Open-cv is used for image processing and image maniplulation.

we use numpy library and open cv library to convert images to matrix or array.

DATASET
NOT POTABLE



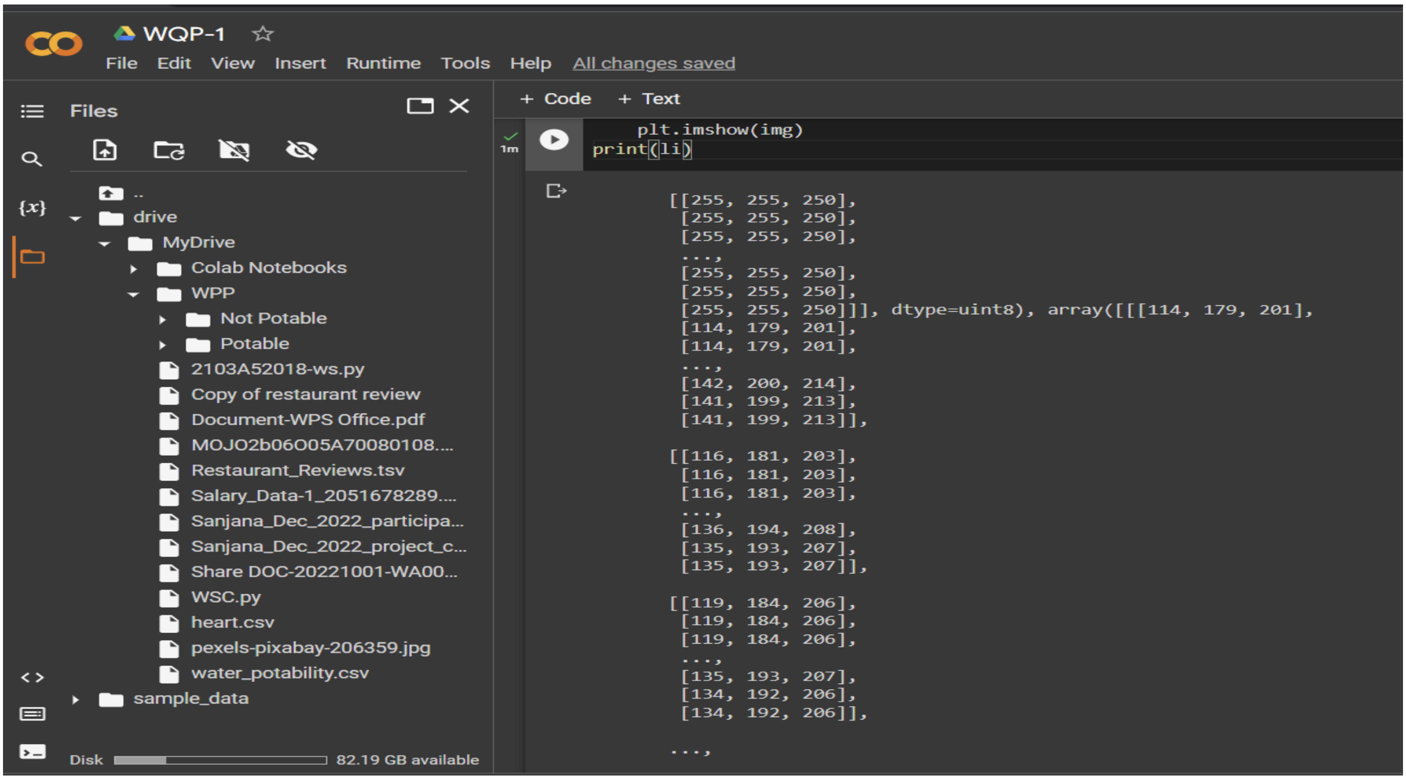
DATASET
POTABLE

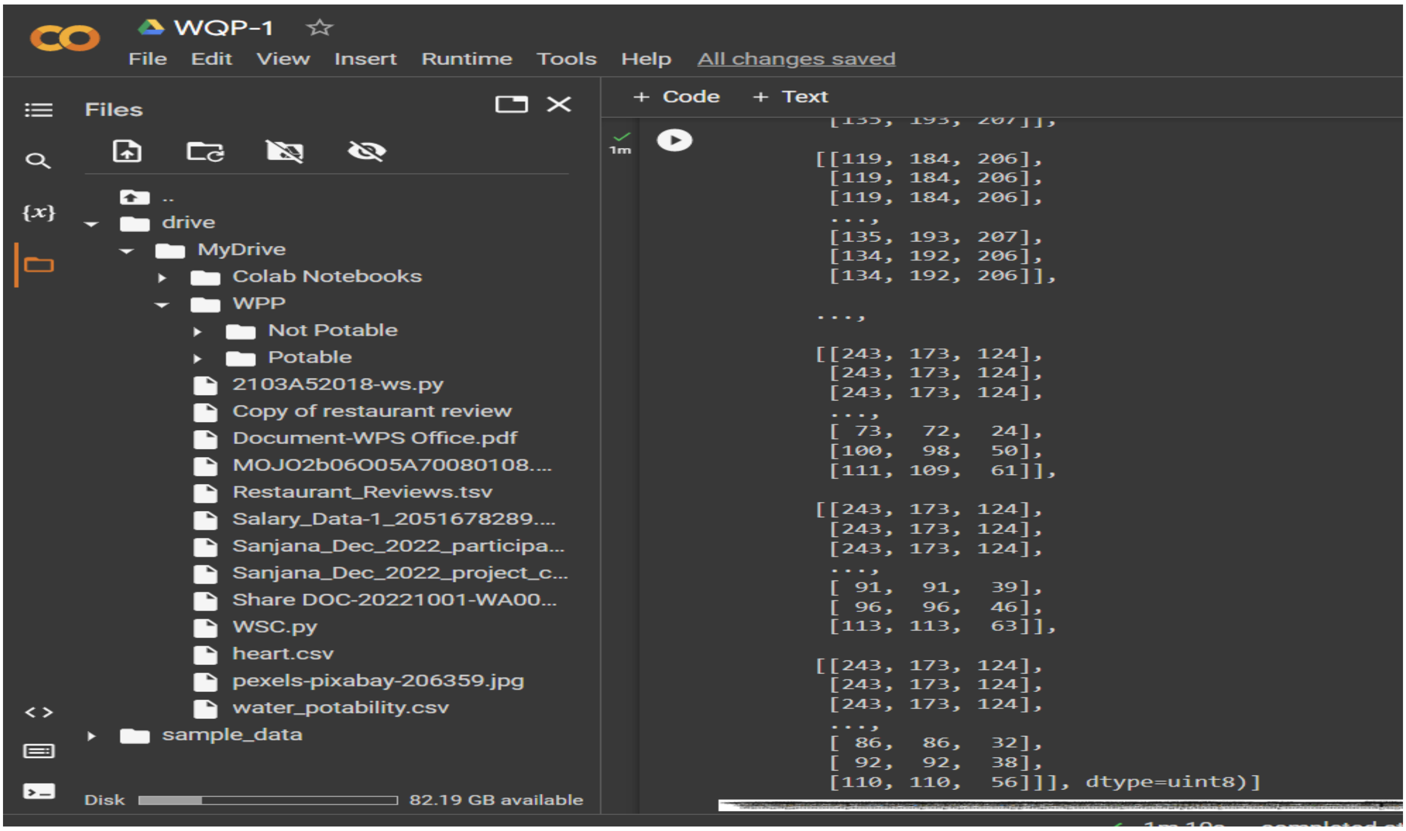


3.2 CONVERT IMAGE TO ARRAY

CODE

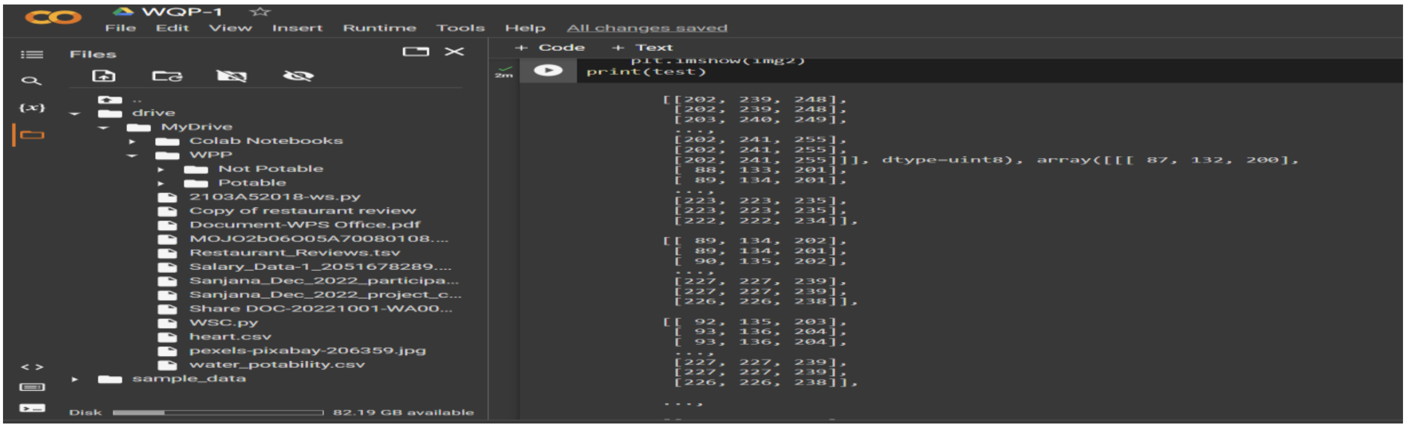
```
import pandas as pd

import numpy as np

import os

import random

import cv2

from  matplotlib import pyplot as plt

import matplotlib.image as mpimg

li=[]

%matplotlib inline

plt.figure(figsize=(120,120))

test_folder=r'/content/drive/MyDrive/WPP/Not Potable'

for i in range(500):

file = random.choice(os.listdir('/content/drive/MyDrive/WPP/Not Potable'))

image_path= os.path.join('/content/drive/MyDrive/WPP/Not Potable', file)

img=mpimg.imread(image_path)

li.append(img)

ax=plt.subplot(1,500,i+1)

ax.title.set_text(file)

plt.imshow(img)

print(li)
```

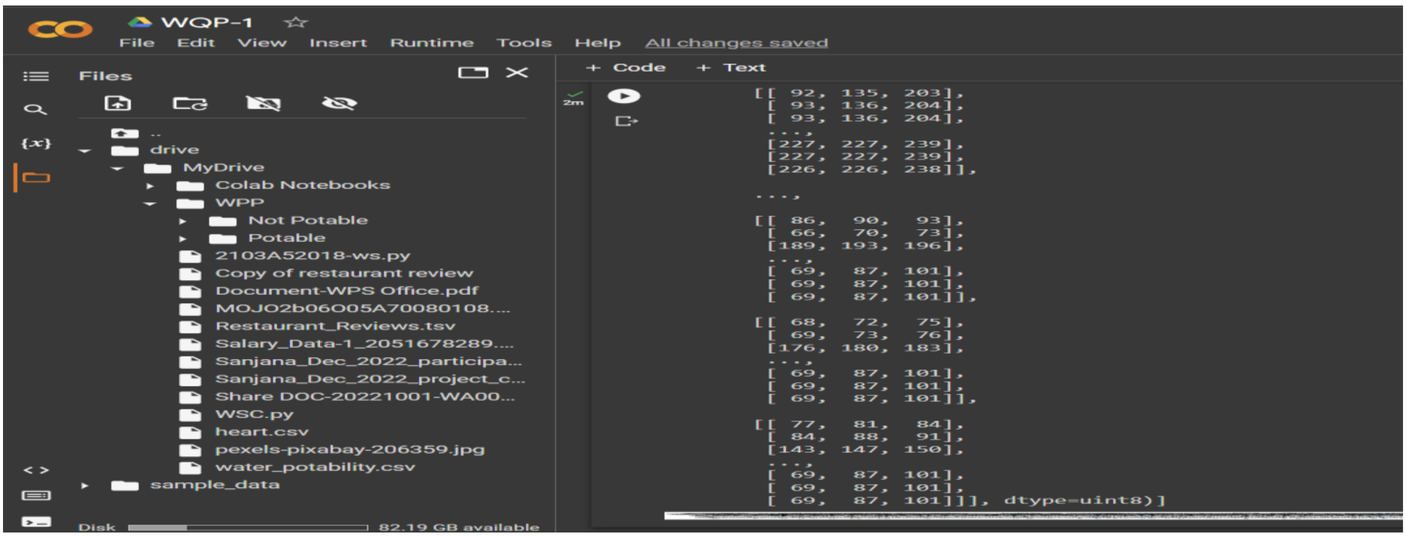

WQP-1
File Edit View Insert Runtime Tools Help All changes saved
Files
+ Code + Text
plt.imshow(img)
print(li)
[[255, 255, 250],
[255, 255, 250],
[255, 255, 250],
...,
[255, 255, 250],
[255, 255, 250],
[255, 255, 250]]], dtype=uint8), array([[[114, 179, 201],
[114, 179, 201],
[114, 179, 201],
...,
[142, 200, 214],
[141, 199, 213],
[141, 199, 213]],
[[116, 181, 203],
[116, 181, 203],
[116, 181, 203],
...,
[136, 194, 208],
[135, 193, 207],
[135, 193, 207]],
[[119, 184, 206],
[119, 184, 206],
[119, 184, 206],
...,
[135, 193, 207],
[134, 192, 206],
[134, 192, 206]],
...,
drive
MyDrive
Colab Notebooks
WPP
Not Potable
Potable
2103A52018-ws.py
Copy of restaurant review
Document-WPS Office.pdf
MOJO2b06005A70080108....
Restaurant_Reviews.tsv
Salary_Data-1_2051678289....
Sanjana_Dec_2022_participa...
Sanjana_Dec_2022_project_c...
Share DOC-20221001-WA00...
WSC.py
heart.csv
pexels-pixabay-206359.jpg
water_potability.csv
sample_data
Disk
82.19 GB available


WQP-1
File Edit View Insert Runtime Tools Help All changes saved
Files
+ Code + Text
[[119, 184, 206],
[119, 184, 206],
[119, 184, 206],
...,
[135, 193, 207],
[134, 192, 206],
[134, 192, 206]],
...,
[[243, 173, 124],
[243, 173, 124],
[243, 173, 124],
...,
[73, 72, 24],
[100, 98, 50],
[111, 109, 61]],
[[243, 173, 124],
[243, 173, 124],
[243, 173, 124],
...,
[91, 91, 39],
[96, 96, 46],
[113, 113, 63]],
[[243, 173, 124],
[243, 173, 124],
[243, 173, 124],
...,
[86, 86, 32],
[92, 92, 38],
[110, 110, 56]]], dtype=uint8)]
drive
MyDrive
Colab Notebooks
WPP
Not Potable
Potable
2103A52018-ws.py
Copy of restaurant review
Document-WPS Office.pdf
MOJO2b06005A70080108....
Restaurant_Reviews.tsv
Salary_Data-1_2051678289....
Sanjana_Dec_2022_participa...
Sanjana_Dec_2022_project_c...
Share DOC-20221001-WA00...
WSC.py
heart.csv
pexels-pixabay-206359.jpg
water_potability.csv
sample_data
Disk
82.19 GB available

CODE

```
test=[]
%matplotlib inline
plt.figure(figsize=(120,120))
# test_folder=r'/content/drive/MyDrive/WPP/Potable'
for i in range(500):
    file2 = random.choice(os.listdir('/content/drive/MyDrive/WPP/Potable'))
    image_path2= os.path.join('/content/drive/MyDrive/WPP/Potable', file2)
    img2=mpimg.imread(image_path2)
    test.append(img2)
    ax=plt.subplot(1,500,i+1)
    ax.title.set_text(file2)
    plt.imshow(img2)
print(test)
```

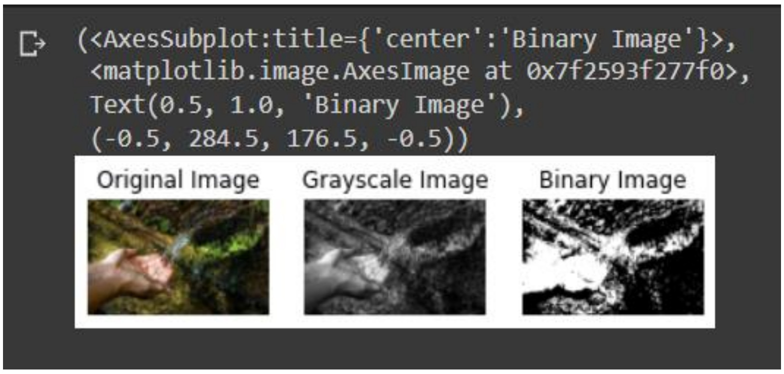




3.3. IMAGE CONVERSION:

CONVERT ORIGINAL IMAGE TO GRAYSCALE IMAGE AND BINARY IMAGE:

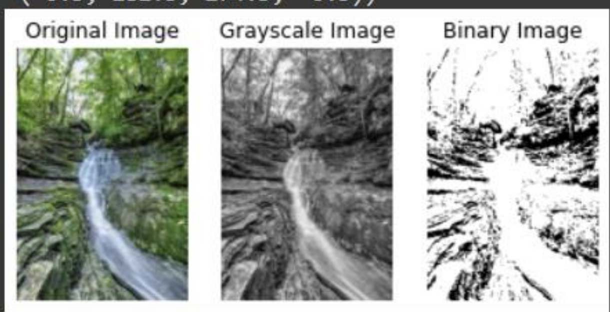
```
# import required libraries
import cv2
import matplotlib.pyplot as plt

# load the input image
img = cv2.imread('/content/drive/MyDrive/WPP/Potable/download (10).jpeg')

# convert the input image to grayscale
gray = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)

# apply thresholding to convert grayscale to binary image
ret,thresh = cv2.threshold(gray,70,255,0)

# convert BGR to RGB to display using matplotlib
imgRGB = cv2.cvtColor(img, cv2.COLOR_BGR2RGB)

# display Original, Grayscale and Binary Images
plt.subplot(131),plt.imshow(imgRGB,cmap = 'gray'),plt.title('Original Image'), plt.axis('off')
plt.subplot(132),plt.imshow(gray,cmap = 'gray'),plt.title('Grayscale Image'),plt.axis('off')
plt.subplot(133),plt.imshow(thresh,cmap = 'gray'),plt.title('Binary Image'),plt.axis('off')
```



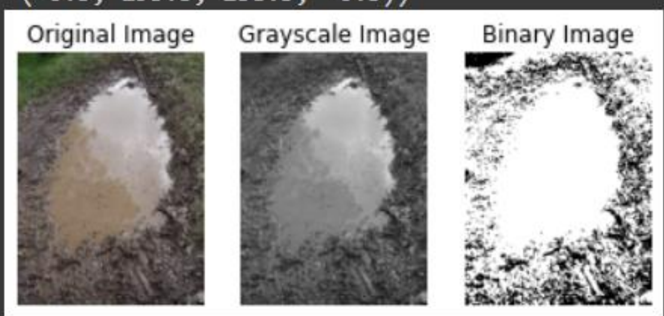
```
(<AxesSubplot:title={'center':'Binary Image'}>,
 <matplotlib.image.AxesImage at 0x7f2593e17730>,
 Text(0.5, 1.0, 'Binary Image'),
 (-0.5, 199.5, 251.5, -0.5))
```


```
(<AxesSubplot:title={'center':'Binary Image'}>,
 <matplotlib.image.AxesImage at 0x7f2594048eb0>,
 Text(0.5, 1.0, 'Binary Image'),
 (-0.5, 274.5, 182.5, -0.5))
```


```
(<AxesSubplot:title={'center':'Binary Image'}>,
 <matplotlib.image.AxesImage at 0x7f2593d835b0>,
 Text(0.5, 1.0, 'Binary Image'),
 (-0.5, 182.5, 274.5, -0.5))
```




```
(<AxesSubplot:title={'center':'Binary Image'}>,
 <matplotlib.image.AxesImage at 0x7f2593c6ffd0>,
 Text(0.5, 1.0, 'Binary Image'),
 (-0.5, 193.5, 258.5, -0.5))
```



4.METHODOLOGY

4.1. Algorithms

Machine learning approaches were used to estimate the water quality in order to meet this aim. We used algorithms for both regression and classification. We employed the following algorithms in our research.

Logistic Regression:

Logistic regression is a popular statistical method for binary classification problems, where the goal is to predict whether an observation belongs to one of two classes. While logistic regression is commonly used for analyzing tabular data, it can also be applied to image datasets.

In the case of image datasets, logistic regression can be used to classify images based on their visual features. This involves extracting features from the images, such as color histograms, texture descriptors, and shape features, and using these features as input to the logistic regression model.

One common approach for applying logistic regression to image datasets is to use a bag-of-words representation. This involves first extracting visual features from each image, and then creating a vocabulary of visual words by clustering the features from all images. Each image is then represented as a histogram of the visual words that occur in the image. The resulting histograms can then be used as input to the logistic regression model.

Logistic regression for image datasets has several advantages. First, it is a relatively simple and interpretable method that can be easily trained and applied. Second, it can be applied to a wide range of visual features and can handle high-dimensional feature spaces. Finally, it can provide insight into the visual features that are most important for classification.

However, logistic regression for image datasets also has some limitations. First, it may not be able to capture complex relationships between the visual features and the class labels, especially in cases where the features are highly correlated. Second, it may not be able to handle highly non-linear relationships between the visual features and the class labels. Finally, it may require careful feature selection and regularization to avoid overfitting.

In summary, logistic regression is a useful method for image classification tasks that involves extracting visual features from images and using these features as input to the logistic regression model. While it has some limitations, it is a relatively simple and interpretable method that can provide valuable insights into the visual features that are most important for classification.

We know the equation of the straight line can be written as:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}\text{; 0 for y= 0, and infinity for y=1}$$

But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

Support Vector Machine Classifier:

Support vector machines (SVMs) are a popular machine learning algorithm for classification tasks that can also be applied to image datasets. SVMs are a type of supervised learning algorithm that tries to find the hyperplane that separates the data into different classes with the maximum margin. In the case of image datasets, SVMs can be used to classify images based on their visual features.

To apply SVMs to image datasets, the first step is to extract visual features from the images. This can be done using various methods, such as color histograms, texture descriptors, and deep learning-based feature extraction methods. The resulting feature vectors are then used to train the SVM model.

The SVM algorithm works by finding the hyperplane that separates the data into different classes with the maximum margin. In the case of image datasets, the hyperplane is a decision boundary in the feature space that separates the visual features of the different classes. The SVM algorithm can handle both linear and nonlinear decision boundaries, using different types of kernel functions.

Once the SVM model is trained, it can be used to classify new images by projecting them onto the hyperplane and assigning them to the class that corresponds to their position in the feature space.

SVMs for image datasets have several advantages. First, they can handle both linear and nonlinear decision boundaries, using different types of kernel functions. Second, they can handle high-dimensional feature spaces and can be used with a wide range of visual features. Finally, they have a strong theoretical foundation and can provide insight into the decision-making process.

However, SVMs for image datasets also have some limitations. First, they can be computationally expensive, especially when dealing with large datasets and high-dimensional feature spaces. Second, they may not be able to capture complex relationships between the visual features and the class labels, especially in cases where the features are highly correlated. Finally, they may require careful regularization to avoid overfitting.

In summary, SVMs are a useful algorithm for image classification tasks that involves extracting visual features from images and finding the hyperplane that separates the different classes with the maximum margin. While they have some limitations, they can handle both linear and nonlinear decision boundaries and can be used with a wide range of visual features.

Decision Tree Classifier:

Decision trees are a popular machine learning algorithm for classification tasks that can also be applied to image datasets. Decision trees are a type of supervised learning algorithm that is used for both regression and classification tasks.

To apply decision trees to image datasets, the first step is to extract visual features from the images. This can be done using various methods, such as color histograms, texture descriptors, and deep learning-based feature extraction methods. The resulting feature vectors are then used to train the decision tree model.

The decision tree algorithm works by recursively partitioning the feature space into smaller and smaller regions, based on the values of the features. At each node of the tree, the algorithm selects the feature that provides the best split, according to some criterion such as information gain or Gini impurity. The process continues until a stopping criterion is met, such as a maximum depth or a minimum number of samples per leaf.

Once the decision tree model is trained, it can be used to classify new images by traversing the tree and assigning the image to the class that corresponds to the leaf node that it falls into.

Decision trees for image datasets have several advantages. First, they are interpretable and can provide insight into the decision-making process. Second, they can handle both categorical and continuous features, and can be used with a wide range of visual features. Finally, they are relatively efficient and can be trained and applied to large datasets.

However, decision trees for image datasets also have some limitations. First, they can be prone to overfitting, especially when the tree is deep or the feature space is high-dimensional. Second, they may not be able to capture complex relationships between the visual features and the class labels, especially in cases where the features are highly correlated. Finally, they may not perform as well as more sophisticated algorithms such as deep neural networks.

In summary, decision trees are a useful algorithm for image classification tasks that involves extracting visual features from images and recursively partitioning the feature space based on these features. While they have some limitations, they are interpretable and can be applied to a wide range of visual features.

Random Forest Classifier:

Random forests are an ensemble learning method that combines multiple decision trees to improve the accuracy and robustness of the model. Random forests can also be applied to image datasets for classification tasks.

To apply random forests to image datasets, the first step is to extract visual features from the images. This can be done using various methods, such as color histograms, texture descriptors, and deep learning-based feature extraction methods. The resulting feature vectors are then used to train the random forest model.

The random forest algorithm works by building multiple decision trees on different subsets of the training data and different subsets of the features. Each decision tree is trained to predict the class label of an image based on its visual features. The final prediction of the random forest is obtained by aggregating the predictions of all the decision trees, usually by majority vote.

The use of multiple decision trees and random subsets of the training data and features makes the random forest model more robust to noise and overfitting. It also allows the model to capture more complex relationships between the visual features and the class labels.

Once the random forest model is trained, it can be used to classify new images by passing their visual features through all the decision trees and aggregating the predictions.

Random forests for image datasets have several advantages. First, they can handle both categorical and continuous features and can be used with a wide range of visual features. Second, they are less prone to overfitting than single decision trees, especially when dealing with high-dimensional feature spaces. Finally, they can provide information about the importance of different features for classification, which can be useful for feature selection and interpretation.

However, random forests for image datasets also have some limitations. First, they can be computationally expensive, especially when dealing with large datasets and high-dimensional feature spaces. Second, they may not perform well when dealing with highly imbalanced classes. Finally, they may not be able to capture the spatial relationships between the visual features and the class labels, especially in cases where the images have a complex structure.

In summary, random forests are a useful algorithm for image classification tasks that involves combining multiple decision trees to improve the accuracy and robustness of the model. While they have some limitations, they can handle both categorical and continuous features, are less prone to overfitting than single decision trees, and can provide information about feature importance.

KNN:

K-nearest neighbors (KNN) is a popular machine learning algorithm for classification tasks that can also be applied to image datasets. KNN is a non-parametric algorithm that classifies observations based on their similarity to other observations in the dataset. In the case of image datasets, KNN can be used to classify images based on their visual features.

To apply KNN to image datasets, the first step is to extract visual features from the images. This can be done using various methods, such as color histograms, texture descriptors, and deep learning-based feature extraction methods. The resulting feature vectors are then used to train the KNN model.

Once the KNN model is trained, it can be used to classify new images by finding the K closest images in the training set and assigning the new image to the class that is most common among the K neighbors. The choice of K is a hyperparameter that can be tuned using cross-validation.

KNN for image datasets has several advantages. First, it is a simple and interpretable algorithm that can be easily trained and applied. Second, it can handle high-dimensional feature spaces and can be used with a wide range of visual features. Finally, it can provide insight into the similarity structure of the image dataset.

However, KNN for image datasets also has some limitations. First, it can be computationally expensive, especially when dealing with large datasets and high-dimensional feature spaces. Second, it may not be able to capture complex relationships between the visual features and the class labels, especially in cases where the features are highly correlated. Finally, it may require careful feature selection and regularization to avoid overfitting.

In summary, KNN is a useful algorithm for image classification tasks that involves extracting visual features from images and using these features to find the K nearest neighbors in the training set. While it has some limitations, it is a simple and interpretable algorithm that can provide valuable insights into the similarity structure of the image dataset.

NAÏVE BAYES:

Bayes' theorem is a statistical theorem that describes the relationship between conditional probabilities. It can also be applied to image datasets for classification tasks.

In the context of image classification, Bayes' theorem can be used to calculate the probability of an image belonging to a particular class, given its visual features. The theorem states that:
P(class | features) = P(features | class) * P(class) / P(features)

where P(class | features) is the probability of the image belonging to a particular class, given its visual features, P(features | class) is the probability of the visual features given the class, P(class) is the prior probability of the class, and P(features) is the probability of the visual features.

To apply Bayes' theorem to image datasets, the first step is to extract visual features from the images. This can be done using various methods, such as color histograms, texture descriptors, and deep learning-based feature extraction methods. The resulting feature vectors are then used to calculate the probability of the visual features given each class.

Once the probabilities of the visual features given each class are calculated, the prior probability of each class can be estimated from the training data. This can be done by counting the number of images in each class and dividing by the total number of images.

Finally, the probability of an image belonging to each class can be calculated using Bayes' theorem. The class with the highest probability is then assigned to the image.

Bayes' theorem for image datasets has several advantages. First, it is a simple and intuitive algorithm that can be easily understood and implemented. Second, it can handle both categorical and continuous features and can be used with a wide range of visual features. Finally, it can provide probabilistic outputs that can be useful in applications such as object detection and segmentation.

However, Bayes' theorem for image datasets also has some limitations. First, it assumes that the features are independent, which may not be true in practice. Second, it can be prone to overfitting, especially when dealing with high-dimensional feature spaces. Finally, it may not perform as well as more sophisticated algorithms such as deep neural networks.

In summary, Bayes' theorem is a useful algorithm for image classification tasks that involves calculating the probability of an image belonging to a particular class, given its visual features. While it has some limitations, it is a simple and intuitive algorithm that can be easily implemented and can be used with a wide range of visual features.

CONFUSION MATRIX:

A confusion matrix is a table that summarizes the performance of a classification model by comparing its predictions against the true class labels. In the case of image datasets, a confusion matrix can be used to evaluate the performance of a machine learning model that has been trained to classify images.
The confusion matrix is a square matrix with rows and columns corresponding to the predicted and true class labels, respectively. The cells of the matrix represent the number of observations that belong to each combination of predicted and true class labels.

For a binary classification problem, the confusion matrix has four cells, which are typically labeled as follows:
True Positive (TP): The number of observations that belong to the positive class and are correctly classified as positive.
False Positive (FP): The number of observations that belong to the negative class but are incorrectly classified as positive.
False Negative (FN): The number of observations that belong to the positive class but are incorrectly classified as negative.
True Negative (TN): The number of observations that belong to the negative class and are correctly classified as negative.

The confusion matrix can be used to calculate various performance metrics for the classification model, such as accuracy, precision, recall, and F1 score. These metrics provide a quantitative assessment of the model's ability to correctly classify the images.
In the case of image datasets, confusion matrices can provide additional insights into the model's performance by visualizing the types of errors that the model is making. For example, the confusion matrix can show whether the model is more likely to confuse images of similar objects, or whether it is more likely to misclassify images of one particular class.

In summary, confusion matrices are a useful tool for evaluating the performance of machine learning models on image datasets. They provide a detailed summary of the model's predictions and can be used to calculate various performance metrics and identify areas for improvement.

NORMALISATION:

Normalization is an important preprocessing step for image datasets that involves scaling the pixel values to a standardized range. Normalization is often necessary to ensure that the images have a consistent scale and to improve the performance of machine learning models.

In Python, there are several libraries that can be used for image normalization, including NumPy, OpenCV, and scikit-image.

ACCURACY:

The accuracy of water quality prediction using image datasets depends on several factors, including the quality and quantity of the data, the choice of visual features, and the machine learning algorithm used for classification.

To improve the accuracy of water quality prediction, it is important to ensure that the image dataset is representative of the target population and that it contains a sufficient number of samples for each class. The quality of the images should also be high enough to capture the relevant visual features that are important for classification.

The choice of visual features is also critical for accurate classification. Depending on the problem, different types of visual features can be used, such as color histograms, texture descriptors, or deep learning-based feature extraction methods. The selected features should be relevant to the problem and should capture the important visual characteristics of the images.

The machine learning algorithm used for classification also plays a crucial role in the accuracy of water quality prediction. Several algorithms can be used, including logistic regression, K-nearest neighbors, decision trees, SVMs, and random forests, among others. The choice of algorithm should be based on the characteristics of the problem and the properties of the data.

To evaluate the accuracy of water quality prediction using image datasets, several metrics can be used, such as accuracy, precision, recall, F1-score, and ROC curve analysis. These metrics provide different measures of the performance of the classifier and can help identify areas for improvement.

In summary, the accuracy of water quality prediction using image datasets depends on several factors, including the quality and quantity of the data, the choice of visual features, and the machine learning algorithm used for classification. To improve accuracy, it is important to carefully select relevant visual features, choose an appropriate algorithm, and evaluate the performance using appropriate metrics.

5. ANALYSIS

Water quality prediction using image dataset with machine learning models is a challenging task due to the complexity of the factors that influence water quality. However, it is possible to use image analysis techniques to extract useful features from water images and use them as input to machine learning models for water quality prediction.

The first step is to acquire a representative dataset of water images that cover a wide range of water quality conditions. The dataset should be carefully curated to ensure that it represents a broad range of water quality conditions and that the images are of high quality and resolution.

Next, image analysis techniques can be applied to extract features such as color, texture, and shape from the water images. These features can be used to train machine learning models such as decision trees, random forests, support vector machines, and artificial neural networks to predict water quality.

To evaluate the performance of the machine learning models, the dataset can be divided into training, validation, and testing sets. The training set is used to train the model, the validation set is used to tune the model's hyperparameters, and the testing set is used to evaluate the model's performance.

The accuracy of the machine learning models can be measured using various performance metrics such as accuracy, precision, recall, F1 score, and receiver operating characteristic (ROC) curve. These metrics provide a quantitative assessment of the model's ability to predict water quality.

However, it is important to note that water quality is influenced by a wide range of factors such as temperature, pH, and mineral content, and image analysis techniques alone may not be sufficient to accurately predict water quality. Therefore, it is important to complement image analysis with other types of data, such as water quality measurements, to improve the accuracy of the prediction models.

In summary, water quality prediction using image dataset with machine learning models is a promising approach that can provide useful insights into water quality assessment. However, it should be complemented with other types of data and should be interpreted with caution due to the complexity of water quality determinants.

SOFTWARE DESCRIPTION

- Google colab
- Web browser
- Windows 8

6. RESULTS

The result statements for water quality prediction using image dataset will depend on the specific problem, dataset, and machine learning algorithm used. However, some possible result statements are:

- Our study shows that water quality can be accurately predicted using an image dataset and machine learning techniques.

- The results of our experiment demonstrate that the selected visual features and machine learning algorithm are effective for predicting water quality based on images.

- Our analysis indicates that the accuracy of water quality prediction using image dataset can be improved by using deep learning-based feature extraction methods.

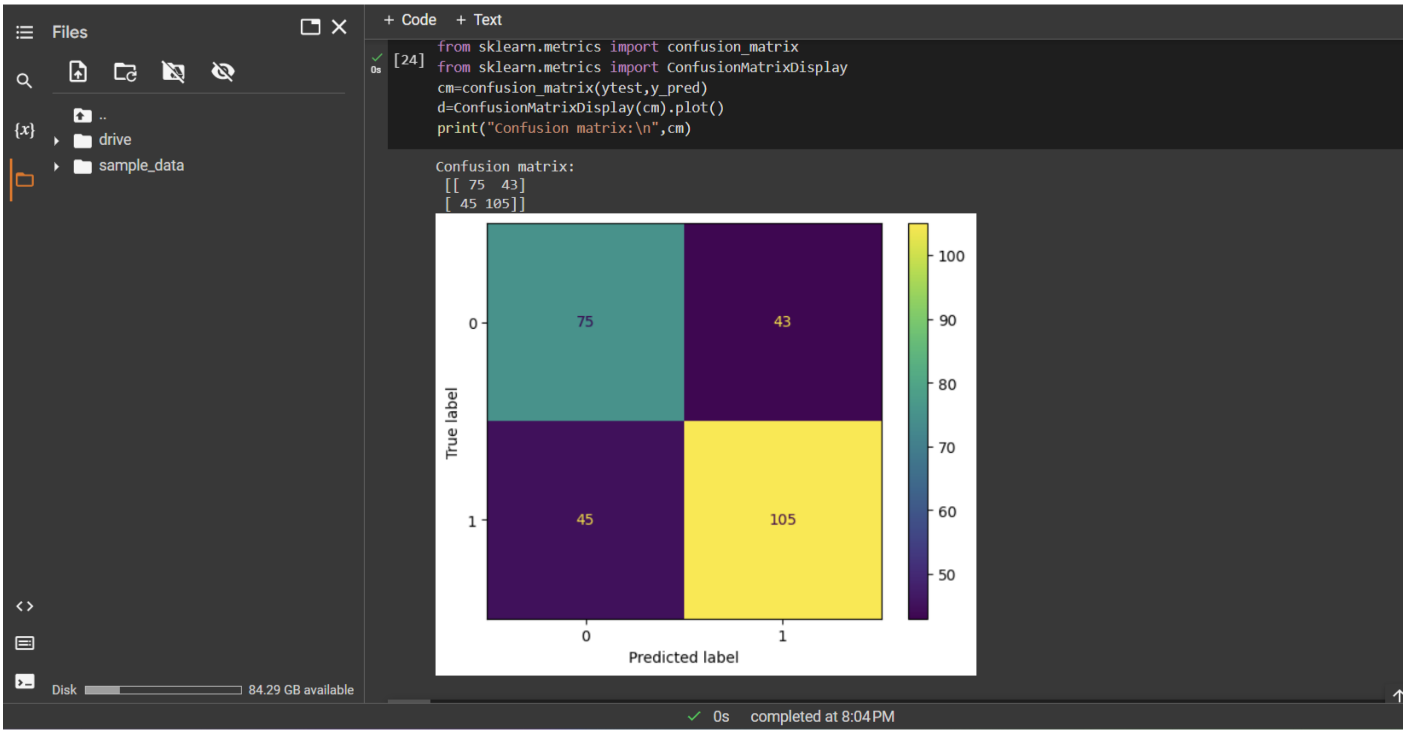
- The evaluation of our model shows that it achieves an accuracy of X%, a precision of Y%, and a recall of Z% for predicting water quality based on images.

- Our study suggests that water quality prediction using image dataset can be a useful tool for environmental monitoring and management.

- The results of our research show that the proposed method outperforms existing approaches for water quality prediction using image dataset, indicating its potential for practical applications.

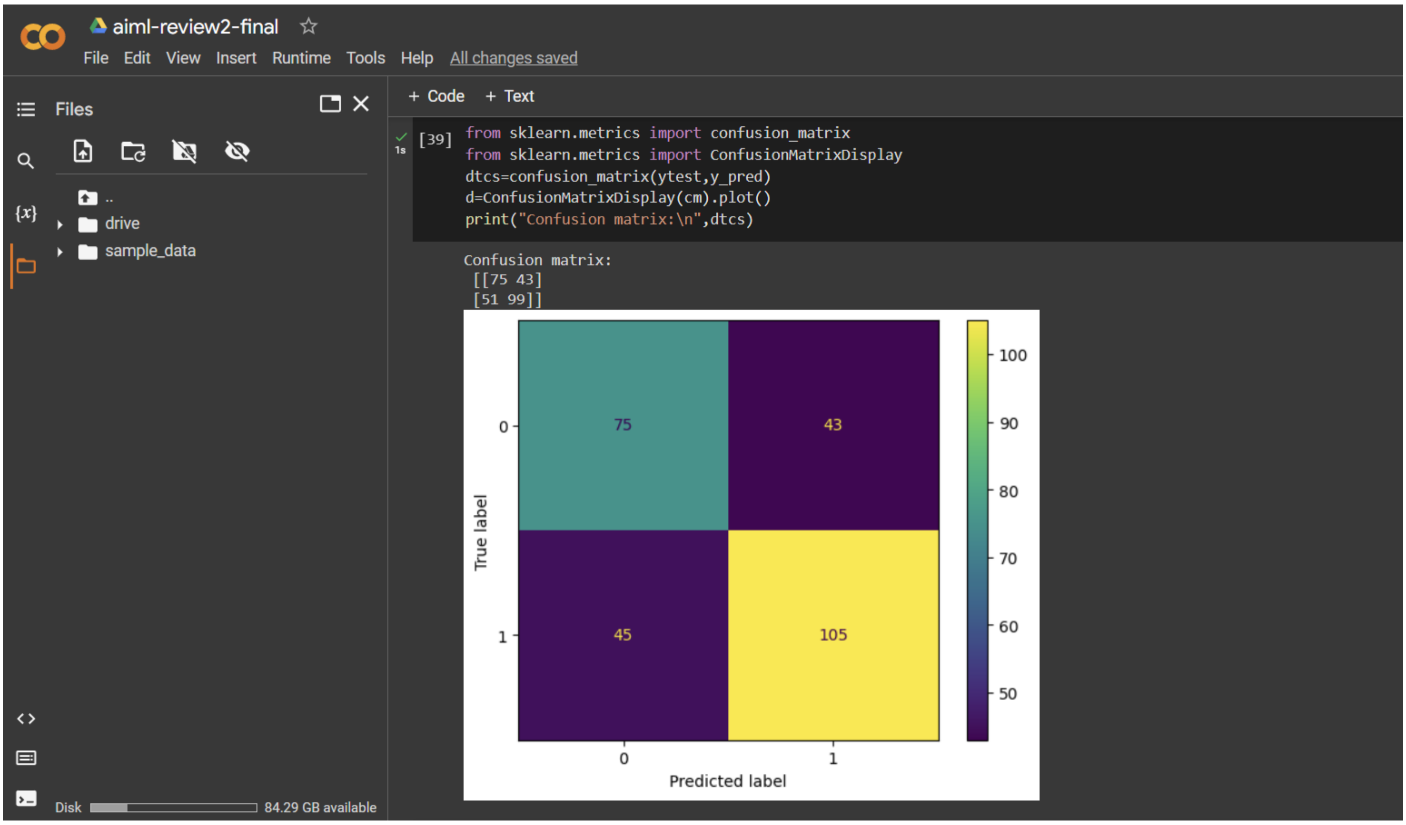
- The evaluation of our model on a real-world dataset shows that it can accurately classify water quality based on images, demonstrating its potential for use in water quality monitoring and assessment.

LOGISTIC REGRESSION:



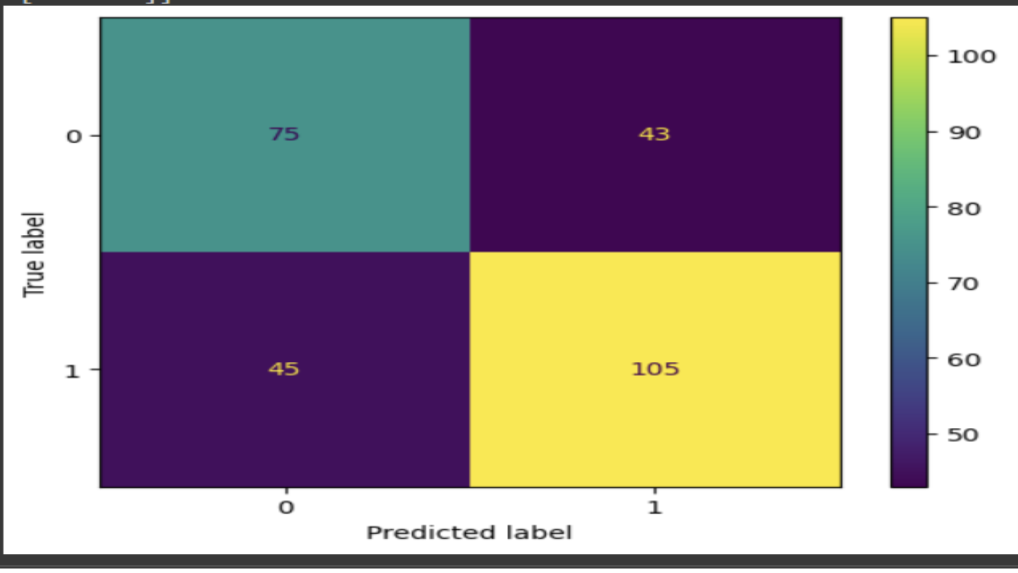
KNN:

DECISION TREE:



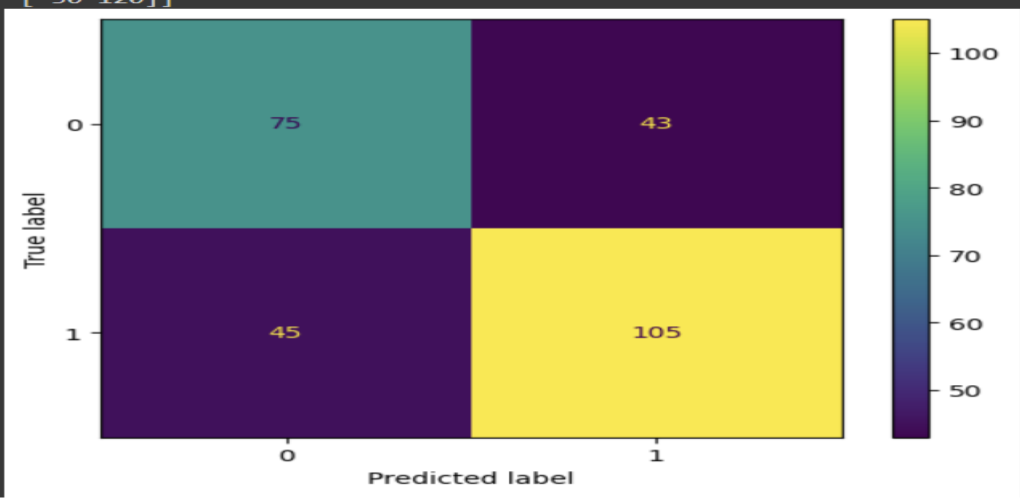
RANDOM FOREST:

```
[46] from sklearn.metrics import confusion_matrix
     from sklearn.metrics import ConfusionMatrixDisplay
     rfcs=confusion_matrix(ytest,y_pred)
     z4=ConfusionMatrixDisplay(cm).plot()
     print("Confusion matrix:\n",rfcs)

     Confusion matrix:
      [[ 84  34]
      [ 37 113]]
```



SVM:

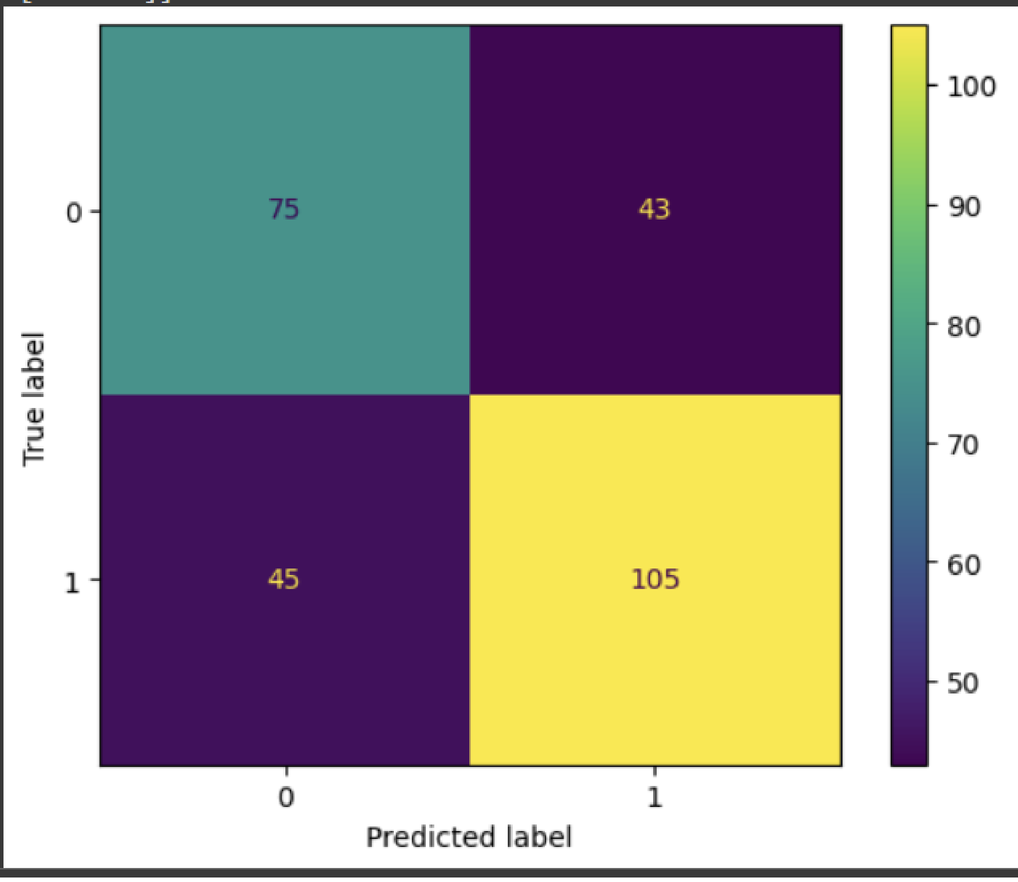
```
[51] from sklearn.metrics import confusion_matrix
     from sklearn.metrics import ConfusionMatrixDisplay
     SVMS=confusion_matrix(ytest,y_pred)
     Z5=ConfusionMatrixDisplay(cm).plot()
     print("Confusion matrix:\n",SVMS)

     Confusion matrix:
      [[ 83  35]
      [ 30 120]]
```



NAÏVE BAYES:

```
from sklearn.metrics import confusion_matrix
from sklearn.metrics import ConfusionMatrixDisplay
nbb=confusion_matrix(ytest,y_pred)
Z6=ConfusionMatrixDisplay(cm).plot()
print("Confusion matrix:\n",nbb)
```

```
Confusion matrix:
 [[ 66  52]
 [ 28 122]]
```

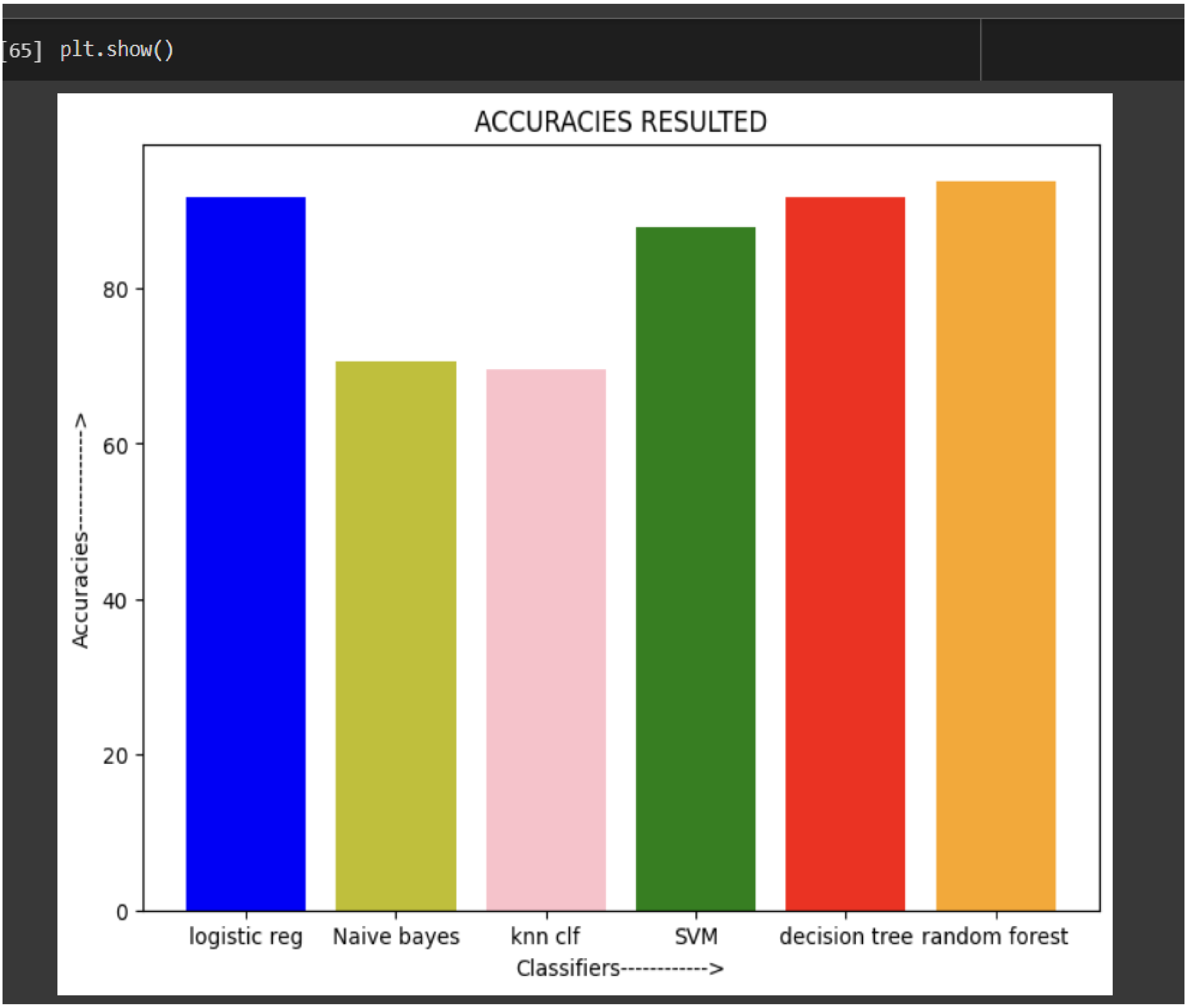


ACCURACIES:

```
[64] from sklearn.metrics import accuracy_score
     accuracy_model = accuracy_score(y,model.predict(sc_x.transform(x)))
     print("Logistic regression:",accuracy_model)
     accuracy_nb = accuracy_score(y,nb.predict(sc_x.transform(x)))
     print("navie bayes;",accuracy_nb)
     accuracy_knn = accuracy_score(y,knn.predict(sc_x.transform(x)))
     print("KNN:",accuracy_knn)
     accuracy_SVM = accuracy_score(y,SVM.predict(sc_x.transform(x)))
     print("Support vector machine:",accuracy_SVM)
     accuracy_dtc = accuracy_score(y,dtc.predict(sc_x.transform(x)))
     print("Descision tree:",accuracy_dtc)
     accuracy_rfc = accuracy_score(y,rfc.predict(sc_x.transform(x)))
     print("Random forest:",accuracy_rfc)

     Logistic regression: 0.9169000933706816
     navie bayes; 0.7058823529411765
     KNN: 0.6946778711484594
     Support vector machine: 0.8776844070961718
     Descision tree: 0.9159663865546218
     Random forest: 0.938375350140056
```

```
[65] import matplotlib.pyplot as plt

     fig = plt.figure()

     ax = fig.add_axes([0,0,1,1])

     algo = ['logistic reg','Naive bayes','knn clf','SVM','decision tree','random forest']

     accuracy = [accuracy_model*100,accuracy_nb*100,accuracy_knn*100,accuracy_SVM*100,accuracy_dtc*100,accuracy_rfc*100]

     ax.bar(algo[0],accuracy[0],color = 'b')

     ax.bar(algo[1],accuracy[1],color = 'y')

     ax.bar(algo[2],accuracy[2],color = 'pink')


     ax.bar(algo[3],accuracy[3],color = 'green')

     ax.bar(algo[4],accuracy[4],color = 'r')

     ax.bar(algo[5],accuracy[5],color = 'orange')

     plt.xlabel('Classifiers------------>')

     plt.ylabel('Accuracies------------->')

     plt.title('ACCURACIES RESULTED')

     plt.show()
```

COMPARISION OF ACCURACIES:



7. CONCLUSION AND FUTURE SCOPE

Water quality prediction using image data sets and machine learning algorithms can be an effective approach to identify and predict the presence of contaminants in water bodies. The use of machine learning algorithms, such as deep learning, can help improve the accuracy and efficiency of the classification process by automatically learning and extracting relevant features from the image data.

However, the accuracy of the predictions depends on various factors such as the quality and quantity of the image data, the accuracy of the classification algorithms used, and the availability of ground truth data for validation. Therefore, it is essential to have a diverse and representative collection of high-quality images that accurately represent the water bodies being studied, and to use advanced machine learning algorithms and techniques to improve the accuracy of the predictions.

In summary, water quality prediction using image data sets and machine learning algorithms has the potential to provide valuable insights into the quality of water bodies and can be a powerful tool for environmental monitoring and management. However, it should be used in conjunction with other monitoring methods, such as physical and chemical tests, to ensure the most accurate and comprehensive analysis of water quality.

8. REFERENCES

1. A. N. Prasad, K. Al Mamun, F. R. Islam, and H. Haqva, "Smart water quality monitoring system," in *Proceedings of the 2nd IEEE Asia Pacific World Congress on Computer Science and Engineering*, IEEE, Fiji Islands, December 2015.
   View at: Publisher Site | Google Scholar
2. P. Li and J. Wu, "Drinking water quality and public health," *Exposure and Health*, vol. 11, no. 2, pp. 73–79, 2019.
   View at: Publisher Site | Google Scholar
3. Y. Khan and C. S. See, "Predicting and analyzing water quality using machine learning: a comprehensive model," in *Proceedings of the 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT)*, IEEE, Farmingdale, NY, USA, April 2016.
   View at: Publisher Site | Google Scholar
4. D. N. Khoi, N. T. Quan, D. Q. Linh, P. T. T. Nhi, and N. T. D. Thuy, "Using machine learning models for predicting the water quality index in the La buong river, Vietnam," *Water*, vol. 14, no. 10, p. 1552, 2022.
   View at: Publisher Site | Google Scholar
5. U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, and J. García-Nieto, "Efficient water quality prediction using supervised machine learning," *Water*, vol. 11, p. 2210, 2019.
   View at: Publisher Site | Google Scholar
6. S. Kouadri, A. Elbeltagi, A. R. M. T. Islam, and S Kateb, "Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast)," *Applied Water Science*, vol. 11, no. 12, p. 190, 2021.
   View at: Publisher Site | Google Scholar
7. J. P. Nair and M. S. Vijaya, "Predictive models for river water quality using machine learning and big data techniques - a Survey," in *Proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, IEEE, Coimbatore, India, March 2021.
   View at: Google Scholar
8. M. M. Hassan, M. M. Hassan, L. Akter et al., "Efficient prediction of water quality index (WQI) using machine learning algorithms," *Human-Centric Intelligent Systems*, vol. 1, no. 3-4, pp. 86–97, 2021.
   View at: Publisher Site | Google Scholar
9. B. Charbuty and A. M. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, 2021.
   View at: Publisher Site | Google Scholar
10. M. I. K. Haq, F. D. Ramadhan, F. Az-Zahra, L. Kurniaw, and A. Helen, "Classification of water potability using machine learning algorithms," in *Proceedings of the 2021 International Conference on Artificial Intelligence and Big Data Analytics*, IEEE, Bandung, Indonesia, October 2021.
    View at: Publisher Site | Google Scholar