**DATA ANALYTICS WITH COGNOS - GROUP 5**

# PROJECT: WATER QUALITY ANALYSIS

**PHASE 5: PROJECT DOCUMENTATION &SUBMISSION**

**SUBMITTED BY**

# Joslin Pinu. P:963321106050

WATER QUALITY ANAYSIS

**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.

Some of the water quality parameters are,

* pH value
* Hardness
* Total Dissolved Solids
* Chloramines
* Sulfate
* Conductivity
* Organic carbon
* Trihalomethanes
* Turbidity
* Potability **1. Data Preparation:**

Import necessary libraries (e.g., pandas, numpy, matplotlib, scikit-learn).

Load your dataset.

Explore and preprocess your data. This includes handling missing values, encoding categorical variables, and scaling numerical features.

**2.Exploratory Data Analysis (EDA):**

Create visualizations to better understand your data. Common libraries for this are Matplotlib and Seaborn.

Examples of visualizations: histograms, scatter plots, box plots, etc., depending on your data type.

**3.Feature Engineering:**

If needed, create new features or transform existing ones to improve the performance of your predictive model.

**4.Splitting Data:**

Split your data into training and testing sets to evaluate your model.

**5.Building a Predictive Model:**

Select an appropriate algorithm for your problem (e.g., linear regression, decision tree, random forest, or neural network). Train your model on the training data.

Evaluate its performance using metrics like mean squared error (MSE), R-squared, etc.

**6.Predictions:**

Make predictions on your test data.

**7.Visualize Predictions:**

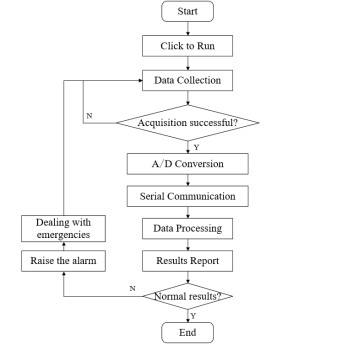
Create a bar chart or any other suitable visualization to display the predicted values alongside the actual values for comparison.

**Dataset Link:**

[**https://www.kaggle.com/datasets/adityakadiwal/waterpotability**](https://www.kaggle.com/datasets/adityakadiwal/water-potability)

**FLOWCHART:**

The flowchart for water quality analysis is as shown in the figure:



## OBJECTIVES

The water quality prediction problem is classified into five categories based on the size of a water quality dataset. The main objectives of this study are summarized as follows:

**Objective-1**: A first analysis was conducted on the available data to clean, normalize and perform feature selection on the water quality measures, and therefore, to obtain the minimum relevant subset that allows high precision with low cost. In this way, expensive and cumbersome lab analysis with specific sensors can be avoided in further similar analyses.

**Objective-2**: A series of representative supervised prediction (prediction, classification and regression) algorithms were tested on the dataset worked here. The complete methodology is proposed in the context of water quality numerical analysis.

## TECHNIQUES

The contribution is:

To carry out a systematic literature review in order to ascertain the current ML techniques used for the WQAD (Water Quality Anomaly Detection) problem.

To highlight the shortcomings and limitations of these current methods

To propose a hybrid DL-ELM framework in WQAD, which could be investigated further

To recommend future research directions T

## Project Content Remarks

Temperature ≥30℃ HighTemperature

18-30℃ Suitable temperature range

≤18℃ Temp too low

pH 7-8.5 Safe range of mariculture 6.5-8.5 Safe range of freshwater aquaculture

Turbidity ≥10 NTU Difficult to eat or breathe

## PROGRAM

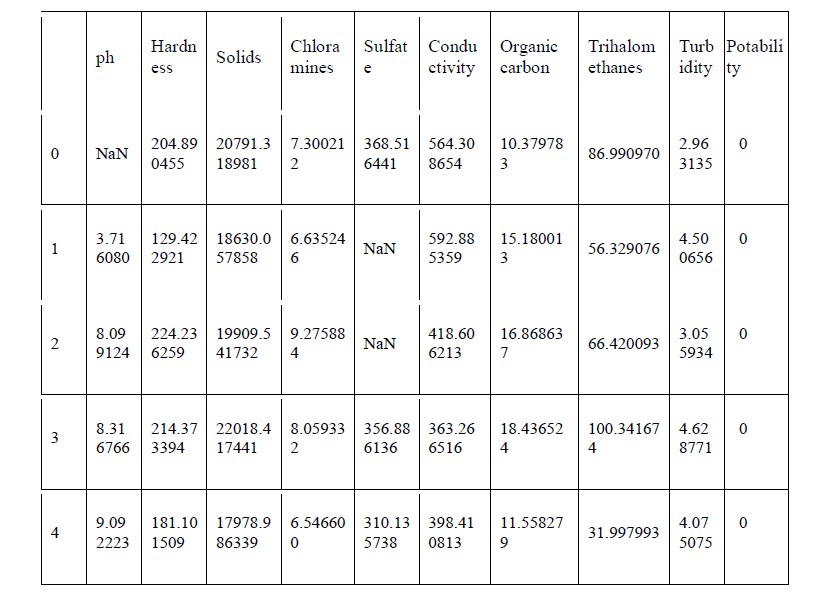
import numpy as np import pandas as pd import seaborn as sns; import matplotlib.pyplot as plt; import plotly.express as px; import missingno as msno;

from sklearn.tree import DecisionTreeClassifier; from sklearn.ensemble import RandomForestClassifier; from sklearn.model\_selection import RandomizedSearchCV,

RepeatedStratifiedKFold, train\_test\_split;

from sklearn.metrics import precision\_score, confusion\_matrix; from sklearn import tree; import os for dirname, \_, filenames **in** os.walk('/kaggle/input'): for filename **in** filenames: print(os.path.join(dirname, filename))

## OUTPUT



**Program:**

import matplotlib.pyplot as plt

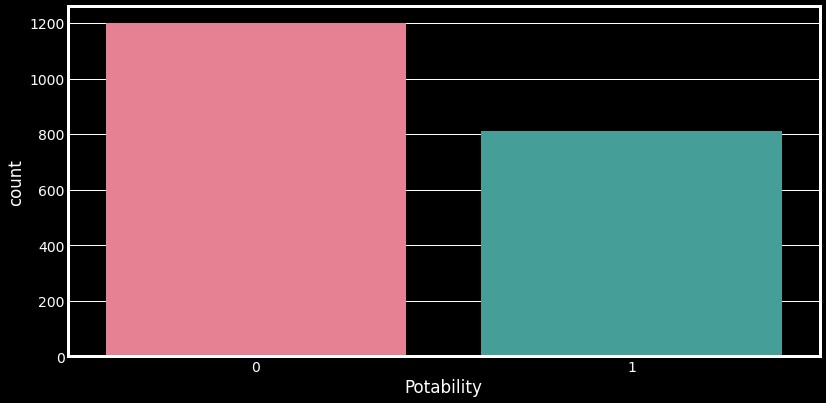
import matplotlib.pyplot as plt plt.style.use('fivethirtyeight') plt.style.use('dark\_background') import numpy as np import pandas as pd import seaborn as sns from matplotlib.colors import ListedColormap from scipy.stats import norm, boxcox

from sklearn.metrics import confusion\_matrix,

classification\_report, accuracy\_score from collections import Counter from scipy import stats from tqdm import tqdm\_notebook *## Importing LuciferML*  from luciferml.supervised.classification import Classification from luciferml.preprocessing import Preprocess as prep import warnings

warnings.simplefilter(action='ignore', category=**Warning**) plt.figure(figsize=(12, 6)) sns.countplot(x="Potability", data=dataset, palette='husl');

## OUTPUT



**Conclusion:**

Good data visualization should communicate a data set clearly and effectively by using graphics. The best visualizations make it easy to comprehend data at a glance.