DATA ANALYTICS WITH COGNOS - GROUP 5

PROJECT: WATER QUALITY ANALYSIS

PHASE 5: PROJECT DOCUMENTATION &SUBMISSION

SUBMITTED BY

MINI MOL P:963321106062

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:

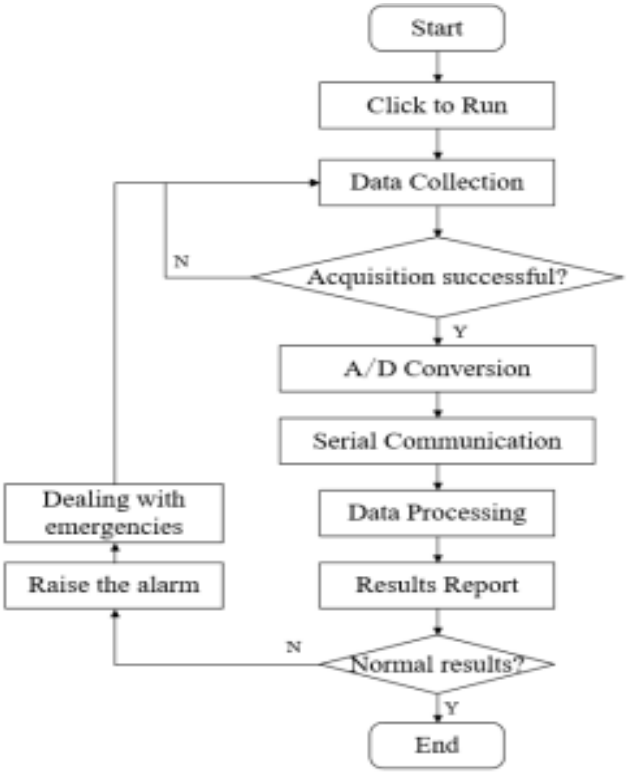
➢Create a bar chart or any other suitable visualization to display the predicted values alongside the actual values for comparison.

Dataset Link:

https://www.kaggle.com/datasets/adityakadiwal/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℃ 18-30℃ ≤18	HighTemperature Suitable temperature range Temp too low
pH	7-8.5 6.5-8.5	Safe range of mariculture 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

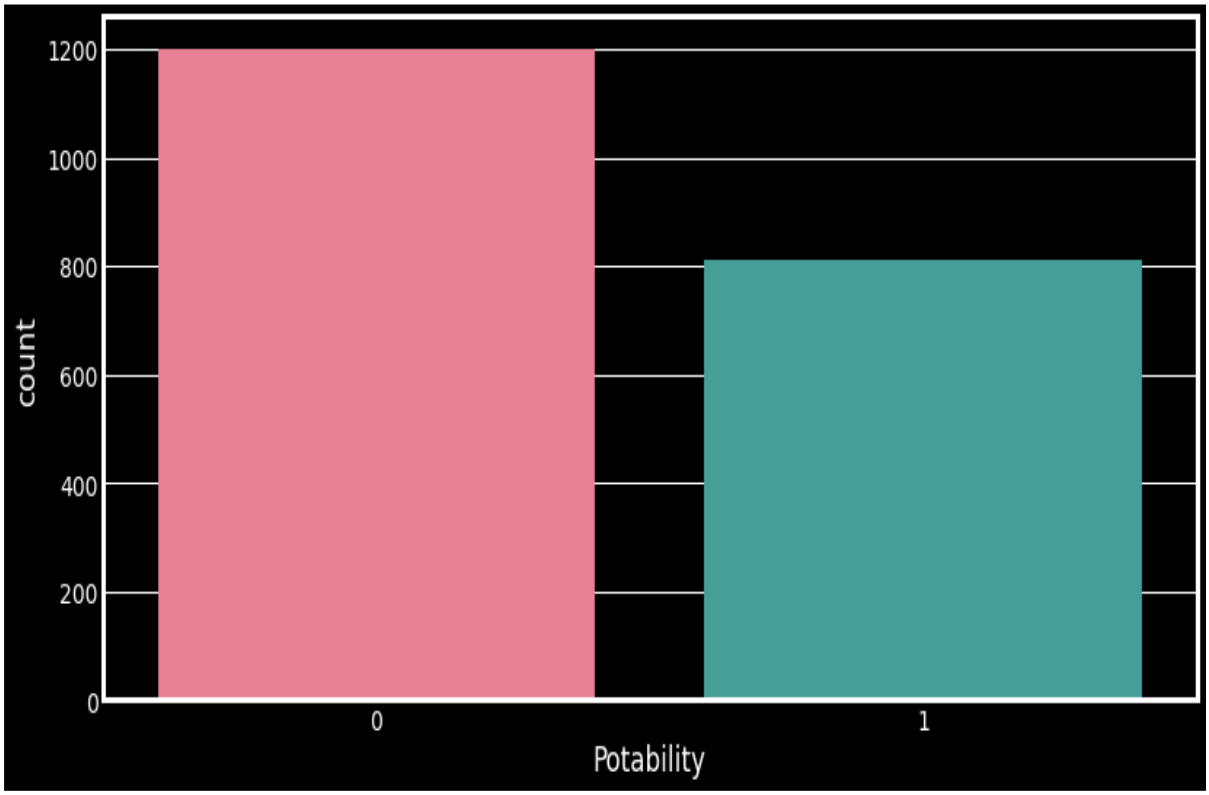
	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

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.