**DATA ANALYTICS WITH COGNOS -**

**GROUP 5**

**PROJECT: WATER QUALITY ANALYSIS**

**PHASE 3: DEVELOPMENT PART 1**

**SUBMITTED BY**

**JoslinPinu.P : 963321106050**

# WATER QUALITY ANALYSIS

Water quality is a critical concern for public health and the environment. In this analysis, we will explore a dataset containing various water quality parameters and predict the potability of water samples using Data Analytics.

Potability refers to the quality of water that makes it safe and suitable for human consumption, including drinking, cooking, and personal hygiene. Potable water is free from harmful contaminants and pathogens that can cause health issues when consumed. Ensuring the potability of water sources is of paramount importance for public health and well-being.

**Significance of Potability:**

**Public Health:** The primary significance of potability is safeguarding public health. Consuming contaminated water can lead to waterborne diseases such as cholera, dysentery, and giardiasis. Potable water reduces the risk of these diseases, especially in communities with limited access to clean water sources.

**Prevention of Waterborne Diseases:** Potable water is crucial in preventing waterborne diseases, particularly in developing regions where access to safe drinking water can be challenging. By providing potable water, governments and organizations can reduce the incidence of waterborne illnesses and improve overall health.

**Hygiene and Sanitation:** Potable water is essential for personal hygiene, including bathing and handwashing. Access to clean water supports good hygiene practices, which, in turn, reduces the spread of diseases.

**Environmental Impact:**

**Water Pollution:** Contamination of water sources, often due to industrial discharge, agricultural runoff, or improper waste disposal, can have severe environmental consequences. Polluted water can harm aquatic life, disrupt ecosystems, and reduce biodiversity.

**Analysing water quality data requires several steps, including data preprocessing and exploratory data analysis**

**(EDA)**.

**Data Collection:**

Start by gathering your water quality data. This data might include various parameters such as pH, turbidity, temperature, dissolved oxygen, and concentrations of different pollutants.

**Data Preprocessing:**

Data preprocessing is crucial to ensure the quality and consistency of your dataset.

1. **Data Cleaning:** Check for missing values: Identify and handle missing data, which can be filled in using techniques like mean, median, or interpolation.

**Remove duplicates:**

If you have duplicate records, remove them. Data type conversion: Ensure that data types are appropriate for analysis (e.g., numerical data is in numeric format).

1. **Feature Selection/Engineering:**

Assess the relevance of each feature. Remove irrelevant or redundant features. Create new features if they can provide valuable information (e.g., calculating the water quality index).

1. **Outlier Detection/Handling:**

Identify outliers using statistical methods and decide whether to remove or adjust them based on domain knowledge.

**Data Exploration:**

EDA helps you gain insights and a better understanding of the water quality data.

* 1. **Summary Statistics**:

Calculate basic statistics like mean, median, standard deviation, and quartiles for each feature.

* 1. **Data Visualization:** Use plots and graphs to visualize the data. Common plots include histograms, box plots, and scatter plots. Here are some examples: Histograms for each parameter to see their distributions. Box plots to identify outliers and compare parameter distributions. Time series plots to identify trends over time. Scatter plots for relationships between different parameters.
  2. **Correlation Analysis:**

Compute correlation coefficients (e.g., Pearson, Spearman) to understand relationships between parameters. Correlation matrices and heatmaps are useful for visualizing correlations. **Data Transformation:**

Depending on your analysis goals, you may need to transform the data.

Common transformations include:

* 1. **Normalization/Standardization:** Scale the data to have zero mean and unit variance for some machine learning algorithms.
  2. **Log Transformation:** If data is skewed, consider logtransforming it to make it more normally distributed.
  3. **Hypothesis Testing:**

Depending on your goals, you may need to perform hypothesis tests to determine if there are significant differences between water quality parameters in different groups or over time.

* 1. **Time Series Analysis:**

If your data includes time-related information, perform time series analysis to identify trends, seasonality, and any patterns over time.

**PREPROCESSING THE DATASET**

* + Data preprocessing is the process of cleaning, transforming, and Integrating data in order to make it ready for analysis.
  + This may involve removing errors and inconsistencies, handling Missing values, transforming the data into a consistent format, and scaling the data to a suitable range.

**PROGRAM**

import numpy as np *# linear algebra* import pandas as pd *# data processing, CSV file I/O (e.g. pd.read\_csv)* import seaborn as sns import matplotlib.pyplot as plt import plotly.express as px import missingno as msno

*# Input data files are available in the read-only "../input/" directory* *# For example, running this (by clicking run or pressing Shift+Enter) will l ist all files under the input directory*

import os for dirname, \_, filenames **in** os.walk('/kaggle/input'): for filename **in** filenames:

print(os.path.join(dirname, filename))

*# You can write up to 20GB to the current directory (/kaggle/working/) tha t gets preserved as output when you create a version using "Save & Run Al*

*l"*

*# You can also write temporary files to /kaggle/temp/, but they won't be sav ed outside of the current session*

*# ML* from sklearn.tree import DecisionTreeClassifier from sklearn.ensemble import RandomForestClassifier from sklearn.model\_selection import RandomizedSearchCV, RepeatedStra tifiedKFold, train\_test\_split from sklearn.metrics import precision\_score, confusion\_matrix

from sklearn import tree

/kaggle/input/water-potability/water\_potability.csv

df = pd.read\_csv("/kaggle/input/water-potability/water\_potability.csv")

In [3]: linkcode df.head()

## OUTPUT

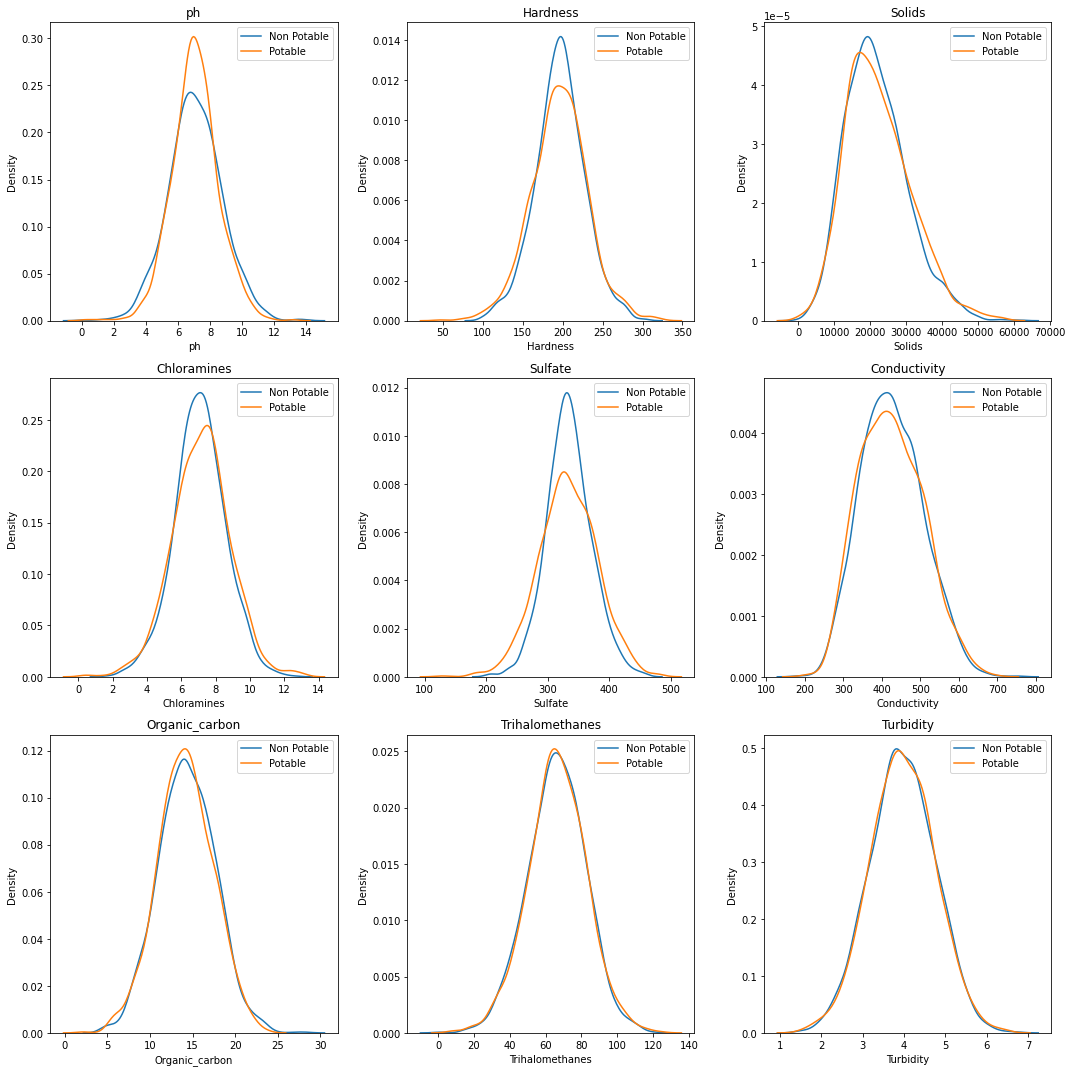
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ph | Hardness | Solids | Chlorami nes | Sulfate | Conductivity | Organic carbon | Trihalometh anes | Turbidity | Potabil  ity |
| 0 | NaN | 204.890455 | 20791.318981 | 7.300212 | 368.516441 | 564.308654 | 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 |

**Distribution of Features:**

|  |
| --- |
| non\_potable = df.query("Potability == 0") potable = df.query("Potability == 1") plt.figure(figsize = (15,15)) for ax, col **in** enumerate(df.columns[:9]):  plt.subplot(3,3, ax + 1) plt.title(col) |

sns.kdeplot(x = non\_potable[col], label = "Non Potable") sns.kdeplot(x = potable[col], label = "Potable") plt.legend() plt.tight\_layout()

### Output



**CONCLUSION:**

The largely fragmented approach that results has contributed to the overexploitation of water resources.