WATER QUALITY ANALYSIS

Phase 5: project Documentation and submission

Problem statement

The problem involves analyzing the water quality data represented by the given columns (pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes, Turbidity, Potability) to assess the suitability of the water for specific purposes such as drinking. The objective is to identify any potential issues or deviations from regulatory standards, and determine the potability of the water based on these parameters

- Analysis objectives: Define specific objectives for analyzing water quality including potability,identifying deviation from standards and understanding parameters relationships

- Data collection: Gather the provided water quality data containing parameters like
- Visualization strategy : plan how to visualize parameters distribution,correlation and potability using suitable tools
- Predictive Modeling: Decide on the machine learning algorithms and features to use for Predicting water potablilit

s.no	Objectives	Code link
1	Design thinking	https://github.com/AJAY1813/water-quality-analysis/blob/main/DAC_phase%201.pdf
2	Anomaly detection	https://github.com/AJAY1813/water-quality-analysis/blob/main/pca%202.ipynb
3	Preprocessing & feature engineering	https://github.com/AJAY1813/water-quality-analysis/blob/main/phase%20code%203.ipynb
4	Model building	https://github.com/AJAY1813/water-quality-analysis/blob/main/DAC_phase4.ipynb

GITHUB Project Repository Link

https://github.com/AJAY1813/water-quality-analysis.git

Design Thinking Document: Assessing Water Quality for Potability

1. **Empathize:**

Understand the problem: The problem is to assess the suitability of water for specific purposes, particularly drinking, using various water quality parameters. Identify stakeholders: Stakeholders may include water quality analysts, regulatory authorities, and the general public.

2. **Define**:

   - Problem Statement: Determine the specific objectives and constraints for assessing water quality
   - Objective: Assess water potability
   - Constraints: Compliance with regulatory standards

3. **Ideate:**

- Data Analysis: Use statistical analysis to identify patterns and outliers in the water quality data.
- Visualization: Create visual representations (charts, graphs) of the data to make it
- more understandable.
- Machine Learning: Explore the use of predictive models to assess water potability
- Expert Consultation: Seek advice from water quality experts or regulatory authorities.

4. Prototype:

- Develop a data analysis pipeline that cleans, preprocesses, and analyzes the water quality data.
- Implement data visualization tools to generate informative charts and graphs.
- If applicable, build a machine learning model to predict water potability.

5. Test:

- Apply the data analysis pipeline to the provided water quality dataset.
- Examine visualizations to identify any anomalies or deviations from regulatory standards.
- Evaluate the accuracy of the machine learning model (if used).

6. Feedback:

- Water quality analysts: Assess the effectiveness of data analysis and visualization tools.
- Regulatory authorities: Verify if the solution aligns with regulatory standards.
- General public (if applicable): Gather input on the transparency and comprehensibility of the water quality assessment.

7. Iterate:

- Adjust data preprocessing and analysis techniques as needed
- Improve data visualization for better communication of results.
- Enhance machine learning models for more accurate predictions.

8. Implement:

Implement the data analysis pipeline in a production environment.
Ensure regular data updates and monitoring for ongoing assessment of water quality.

9. Evaluate:

Monitor water quality data over time to detect trends and changes.
Review regulatory compliance regularly.
Seek feedback from stakeholders for continuous improvement.

10. Share:

- ✓ Publish water quality reports to inform the public.
- ✓ Collaborate with regulatory authorities to ensure compliance.
- ✓ Share insights and best practices with the broader water quality community.

Analysis objectives:

Define specific objectives for analyzing water quality including potability,identifying deviation from standards and understanding parameters relationships

Anomaly Detection

dataset information

0	ph	2785	non-null	float64
1	Hardness	3276	non-null	float64
2	Solids	3276	non-null	float64
3	Chloramines	3276	non-null	float64
4	Sulfate	2495	non-null	float64
5	Conductivity	3276	non-null	float64
6	Organic_carbon	3276	non-null	float64
7	Trihalomethanes	3114	non-null	float64
8	Turbidity	3276	non-null	float64
9	Potability	3276	non-null	int64

Doing Anomaly Detection

We want to find outliers in the dataset through anomaly detection. When solving the water portability problem, it seems right to remove outliers.

Pycaret

PyCaret is a versatile library that can be used not only for traditional supervised machine learning tasks but also for anomaly detection. Anomaly detection involves identifying rare and unusual data points that deviate significantly from the norm. PyCaret makes it relatively easy to apply anomaly detection techniques to your datasets

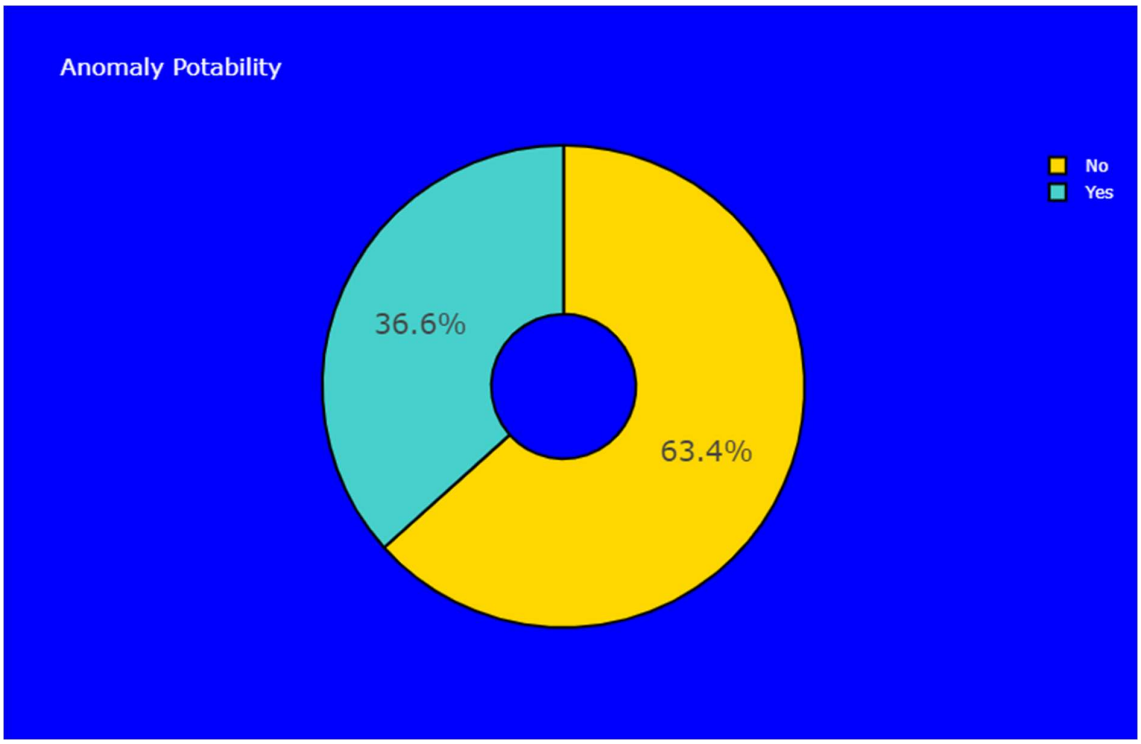
PCA

Principal Component Analysis (PCA) is a widely used dimensionality reduction technique and, when applied to anomaly detection, it becomes a powerful tool for identifying unusual patterns or outliers in datasets. In anomaly detection using PCA, the primary goal is to transform the data into a lower-dimensional space while retaining as much variance as possible. Anomalies can then be detected by identifying data points that deviate significantly from the expected distribution in this reduced space.

`the size of anomaly = 101` # detection anomaly in dataset

Observation:

- ✓ There are 110 anomaly data.
- ✓ If you look at the Top 10 anomaly dates, there are many data judged to have potability.
- ✓ Looking at the target of data judged as anomaly, there are more cases judged as potability.
- ✓ What can be predicted from this is that there are many cases in this dataset where undrinkable
- ✓ water is judged to be drinkable.

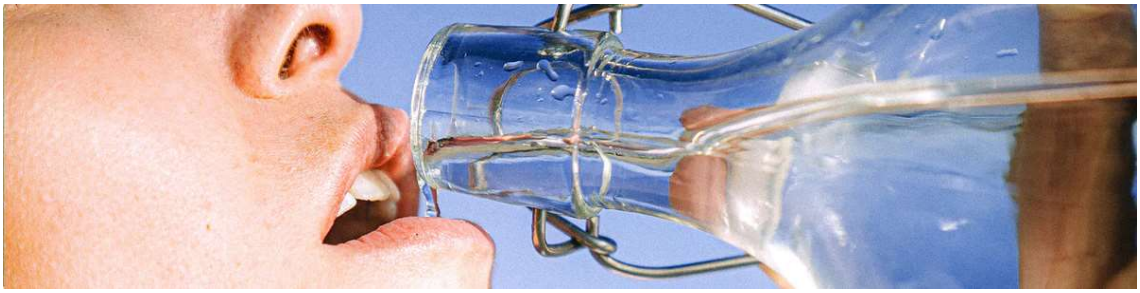


INSIGHTS

The analysis of dataset used histogram based technique to apply on each parameter and PCA algorithm is applied overall dataset determined anomaly of potability

Finally 36.6% yes and 63.4% No show the potability of water. 63.4% is large anomaly Of dataset so we need used hyperparameter and standardisation to enhance model Predict the potability accurately .

Visualization strategy : plan how to visualize parameters distribution,correlation and potability using suitable tools



Checking Missing Values

Observation:

- There are missing values for ph, sulfate, and trihalomethanes.
- ph 491
- Hardness 0 Solids 0 Chloramines 0 Sulfate 781 Conductivity 0 Organic_carbon 0 Trihalomethanes 162 Turbidity 0 Potability 0 is_ph_ok 0 is_Hardness_ok 0 is_Solids_ok 0 is_Chloramines_ok 0 is_Sulfate_ok 0 is_Conductivity_ok 0 is_Organic_carbon_ok 0 is_Trihalomethanes_ok 0 is_Turbidity_ok 0

Handling Missing Values

The method of handling missing values will differ depending on the dataset and the nature of the problem to be solved. This is a matter of determining drinkable water. Filling in missing values with certain predicted values can be a very risky decision.

For example, suppose that the missing value in the 'ph' feature is filled with a certain value. Suppose the actual value of ph is 0, but for some reason it is treated as a missing value. If the ph is 0, it is the same ph as the battery, and people should never drink this kind of wate

After Imputation of missing value

```
ph 0 Hardness 0 Solids 0 Chloramines 0 Sulfate 0 Conductivity 0
Organic_carbon 0 Trihalomethanes 0 Turbidity 0 Potability 0 is_ph_ok 0
is_Hardness_ok 0 is_Solids_ok 0 is_Chloramines_ok 0 is_Sulfate_ok 0
is_Conductivity_ok 0 is_Organic_carbon_ok 0 is_Trihalomethanes_ok 0
is_Turbidity_ok 0
```

Doing Anomaly Detection

We want to find outliers in the dataset through anomaly detection. When solving the water portability problem, it seems right to remove outliers. This is because there should not be outliers in the dataset related to life.

```
the size of anomaly =  101
```

Observation:

- There are 110 anomaly data.
- If you look at the Top 10 anomaly dates, there are many data judged to have potability.

Observation:

- Looking at the target of data judged as anomaly, there are more cases judged as potability.
- What can be predicted from this is that there are many cases in this dataset where undrinkable water is judged to be drinkable. That is, we can predict that recall may be low.

Exploratory Data Analysis

General Data Analysis

Observation:

We see that we have some degree of unbalancedness in our data; we will not apply any upsampling/downsampling methodology as the proportions are more close to equal than to be extremely balanced (cases like 90% / 10% where upsampling is crucial). Also, the more significant label ("Not potable") is the one with more samples; logically, we would prefer a model that will have more false negatives rather than a model that has more false positives.

It appears that there is no linear/ranked correlation between our output label and our features, mostly due to the fact that we have a binary label and continuous features, traditional linear correlation coefficients won't tell us the true underlying story about the relationships between our features and the target variable

Looking at the distribution of all our features divided by our target label, we see that some of them have some difference, a key point that can help us select the features with which we will train our models. To better understand the differences between the features with respect to the target label, a more robust analysis is required to confirm any hypothesis we may have at this point just from looking at the distribution plots.

Statistical Difference Analysis

Explanation:

In order to test for any significant difference between "potable" and "non-potable" water samples, we will treat both labels as two separate populations from which we sampled 'n' and 'k' samples (n = the number of "potable" samples, 'k' = the number of "non-potable" samples). We will perform a two-tailed t-test to check if there is any significant difference between the two sample means, considering the sample size differences and unequal variance. We expect to see low p-values for the features that indeed are significantly different between the labels. We will set our significance level alpha to be equal to or less than 0.1

.

After performing the two-tailed t-test, we see that only "Solids" and "Organic carbon" have p-values below our pre-defined alpha value, even though there are two more features closer to our alpha level than the other 4. When we get to the modeling stage, the 4 features we will use will be all the features we see in the above plot with p-values below 0.18 (first 4 features in the plot)

As an additional metric for consideration, we use "Mutal Information" to test and see if there is any similarity between the probability distribution of or continuous features with the Bernoulli distribution that represent our target. We see that some of the worst scoring features in our t-test have the highest mutual information with our target label, conceptually meaning that knowing something about "Ph" decreases my uncertainty in assuming about "Potability," unfortunately, mutual information doesn't tell me exactly to what assumption does "Ph" contribute. Still, none the less it is an indicator of relationship and a strong what in the matter, so we will indeed include it as well in our modeling section.

Checking Feature Importance

Here we check feature importance in various ways.

Knowing which features are important when judging heart disease from different features will help you make a decision. It will also be very helpful when explaining to people who have been diagnosed.

Observation:

- The correlation coefficient of all features is low.
- Correlation coefficients of newly created derived features are relatively high.

Feature importance with partial dependence

Partial dependence plots (PDP) show the dependence between the target response and a set of input features of interest, marginalizing over the values of all other input features (the‘complement’ features). Intuitively, we can interpret the partial dependence as the expected target response as a function of the input features of interest

Observation:

- The ph feature has a large partial dependence at values between 6 and 8.5.
- For Hardness, the value of partial dependence rapidly increases around 210.

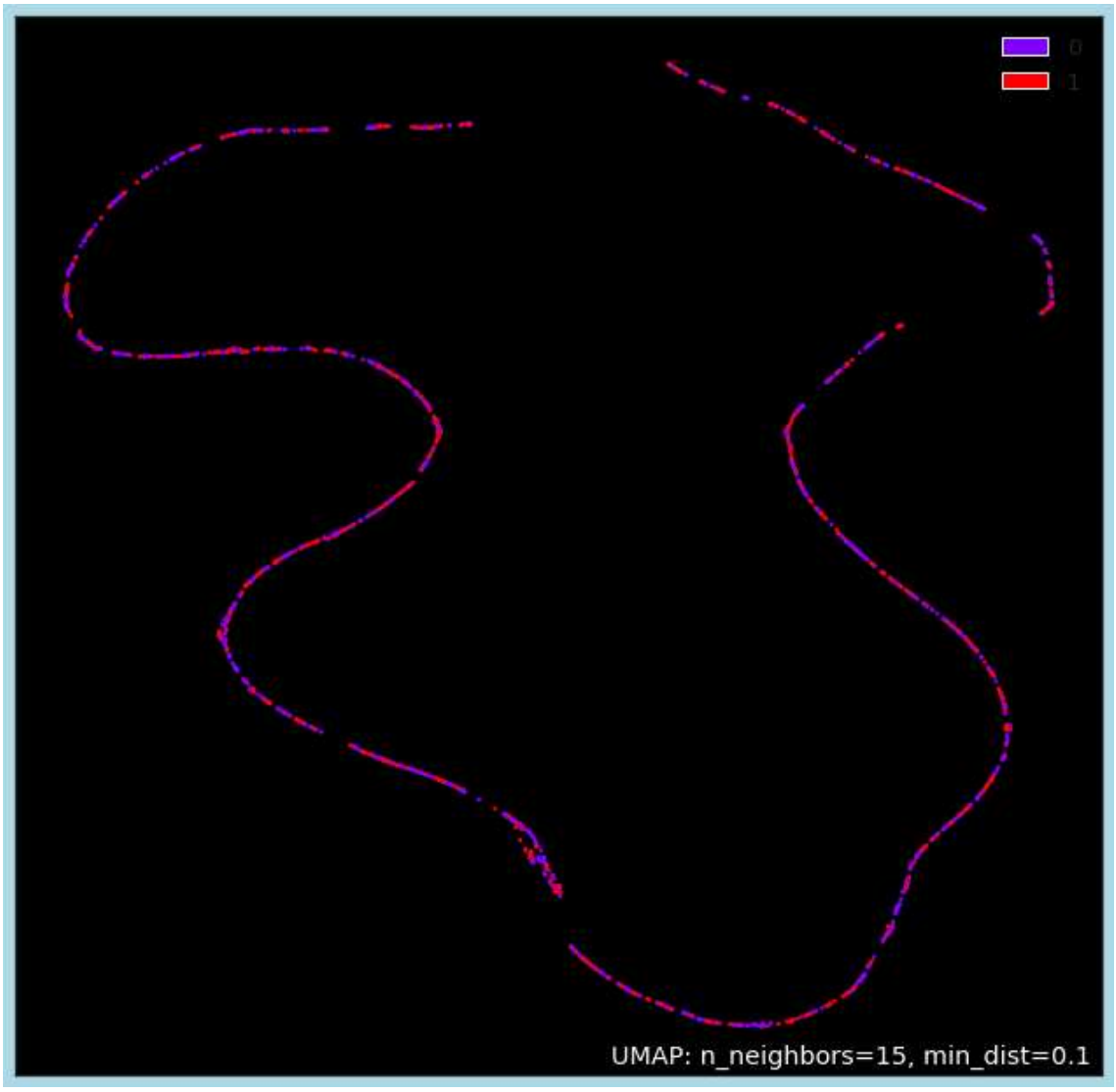
Feature importance based on mean decrease in impurity

After calculating the sum of the decrease of impurity at the point of splitting based on the corresponding feature in each tree, this Mean decrease Gini is the average of all tree values. This value wi

Observation:

- ph and sulfate features were judged to be important features.

Visualizing Training Dataset after Dimension Reduction



Observation:

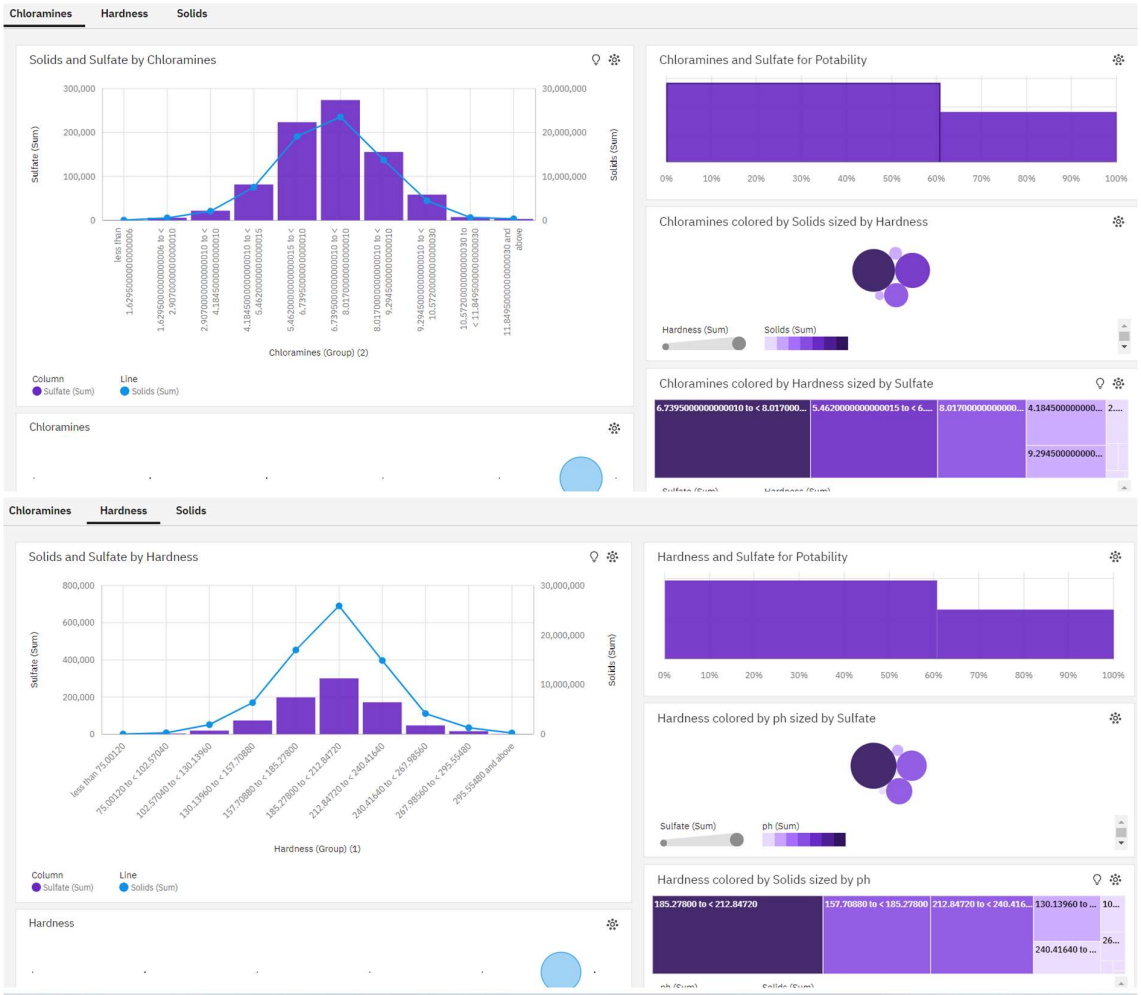
- A specific pattern or boundary is not visible.

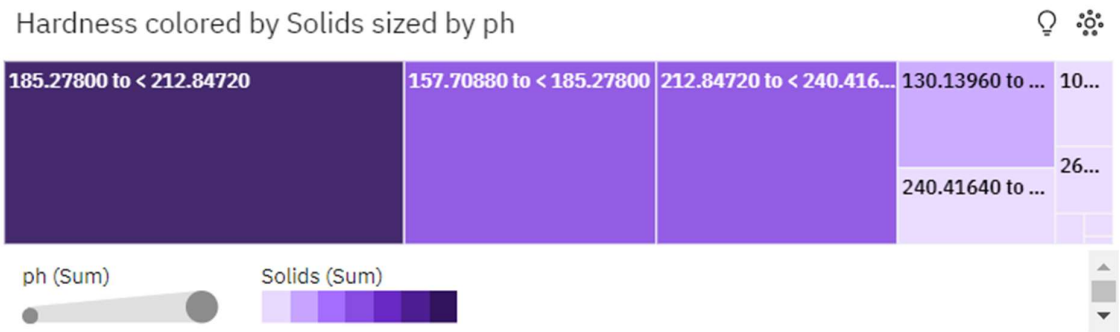
INSIGHTS

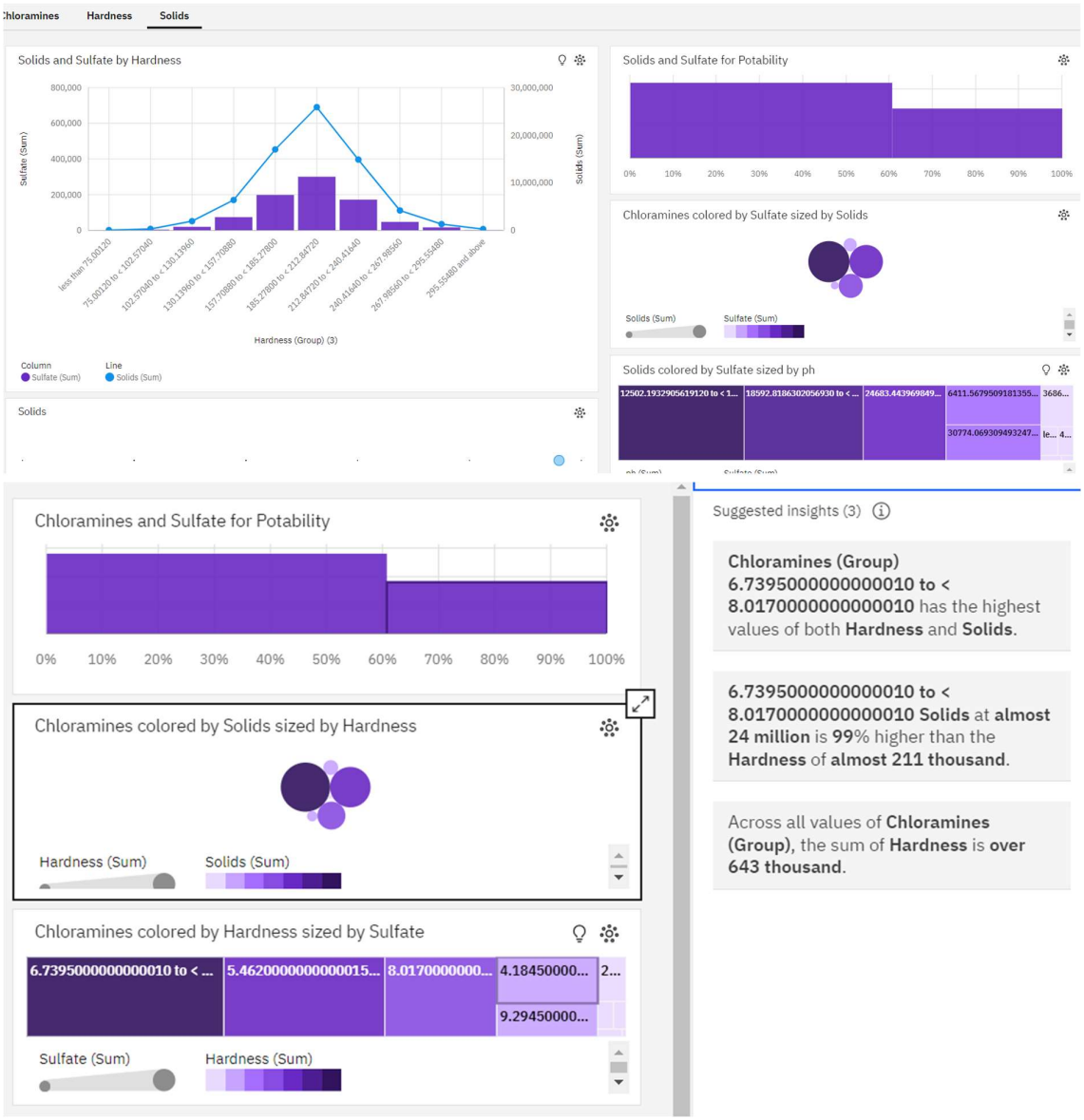
Thus implementation EDA and handing missing values performed and visualize the various parameters from dataset to made lot of insights Through charts and graphs and performed statistical methods to ensured.Standards of datasets.next will implement predictive modelling

IBM Cognos Analytics

visualisation and Insights



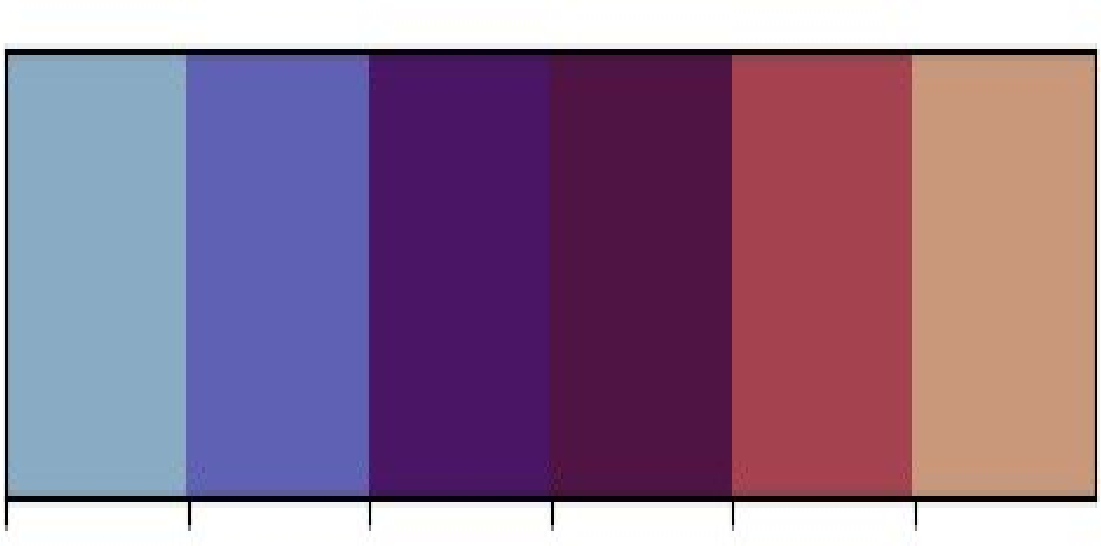




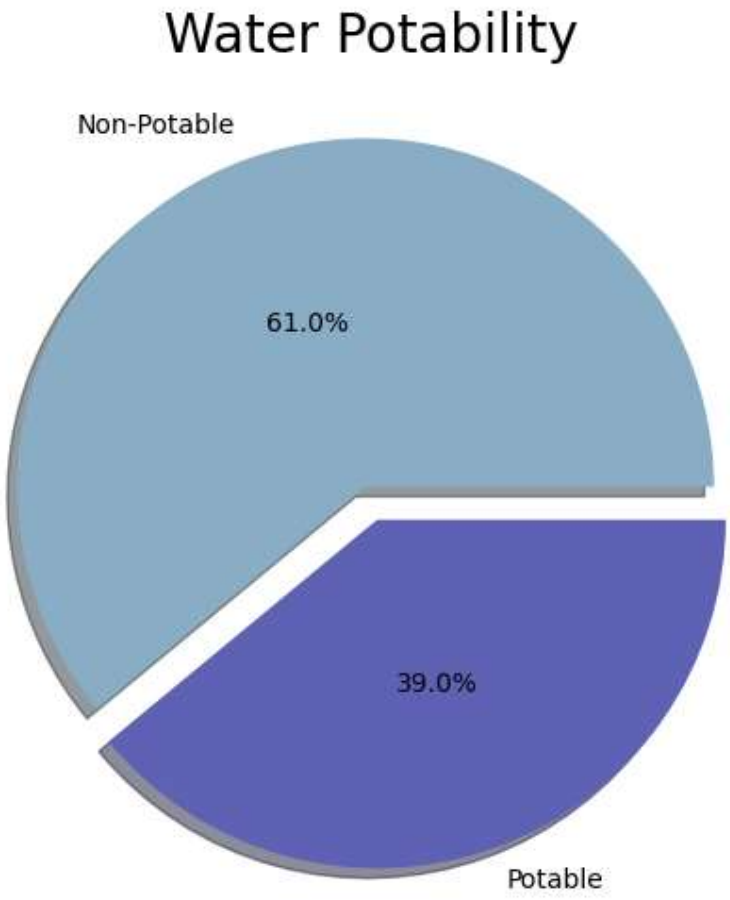
Predictive Modeling: Decide on the machine learning algorithms and features to use for Predicting water potablility

Feature Engineering
Data Visualization

pH Level Chart



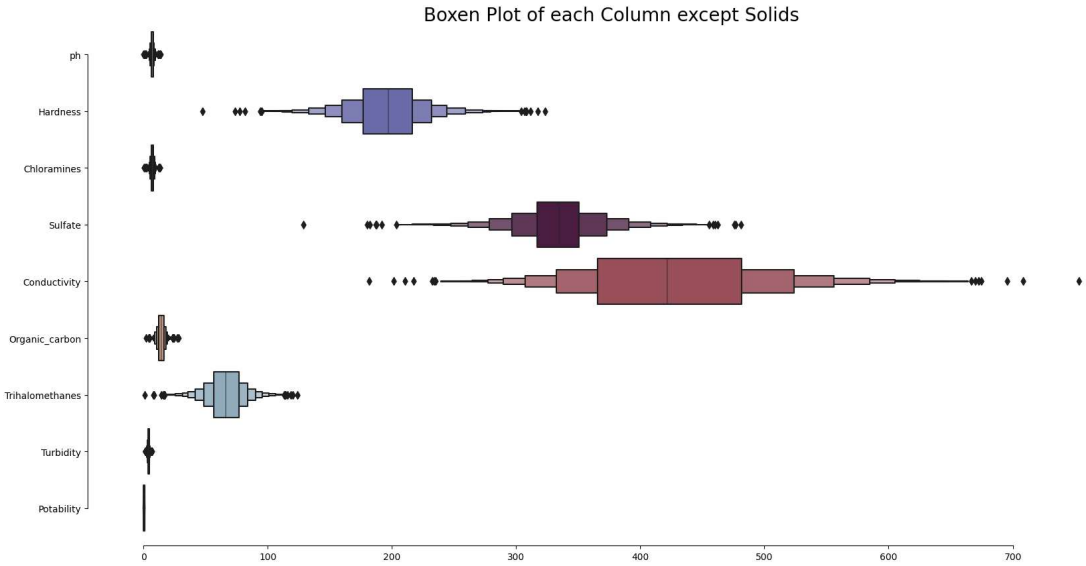
The different colors in this chart indicate the pH level of the given data and the correlation of another parameter
The pH readings are used to roughly determine the type of water, with values above 7 being alkaline and values below 7 being acidic.



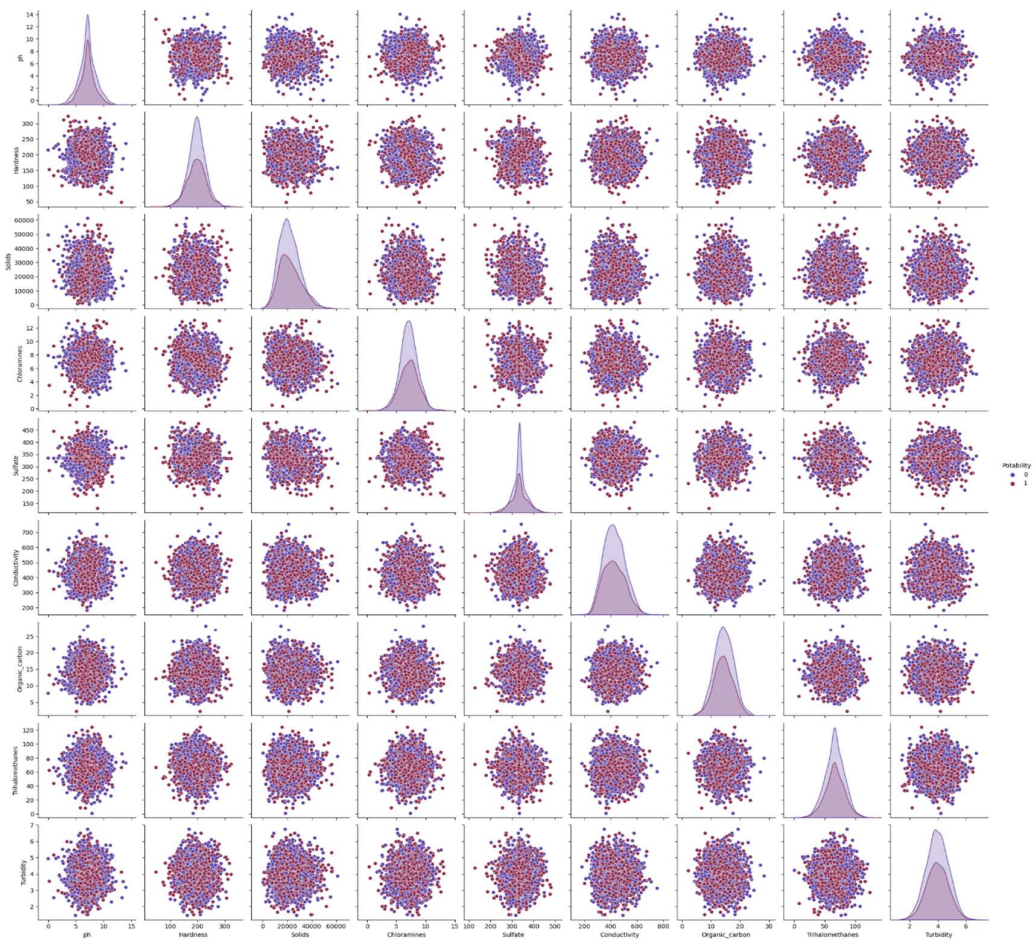
This chart provides insights into the total potability of the water samples. It shows that 61% of the water is non-potable, while 39% is potable

Box plot for each column

There is a box plot for each column in the dataset, which helps visualize the distribution, skewness, and potential outliers of the data



Pairplot



The pair plot shows the relationships between different variables in the dataset. It is a grid of scatterplots, where each variable is plotted against every other variable

Models

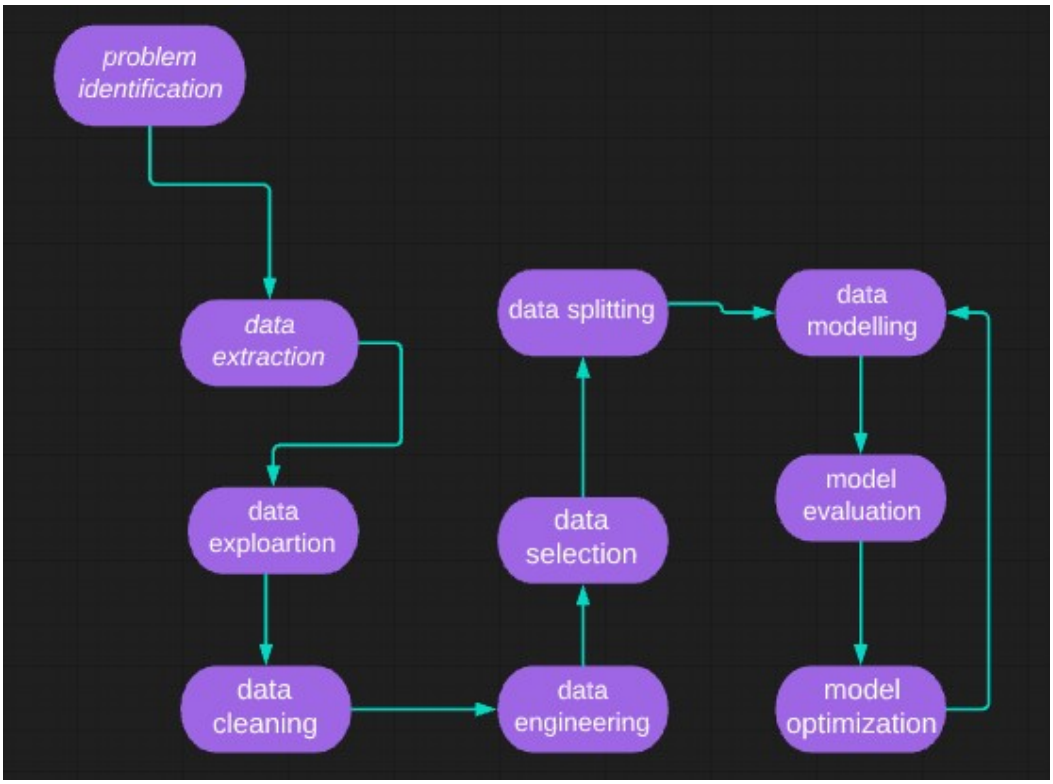
The analysis includes the use of various predictive models to determine water potability based on the water quality parameters. The model scores indicate the accuracy of each model in predicting water potability. The models used in this analysis include Random Forest, Decision Tree, KNN, Support Vector Machines, Logistic Regression, and Naive Bayes`1`. The Random Forest and Decision Tree models have a score of 100%, while the KNN model has a score of 78.27%. The Support Vector Machines model has a score of 73.50%, the Logistic Regression model has a score of 62.64%, and the Naive Bayes model has a score of 61.54%. These scores can be used to

	Model	Score
3	Random Forest	100.00
5	Decision Tree	100.00
1	KNN	78.27
0	Support Vector Machines	73.50
2	Logistic Regression	62.64
4	Naive Bayes	61.54

FInallly

Thus the water quality analaysis project successfully implemented

PROCESS FLOW



DATASET

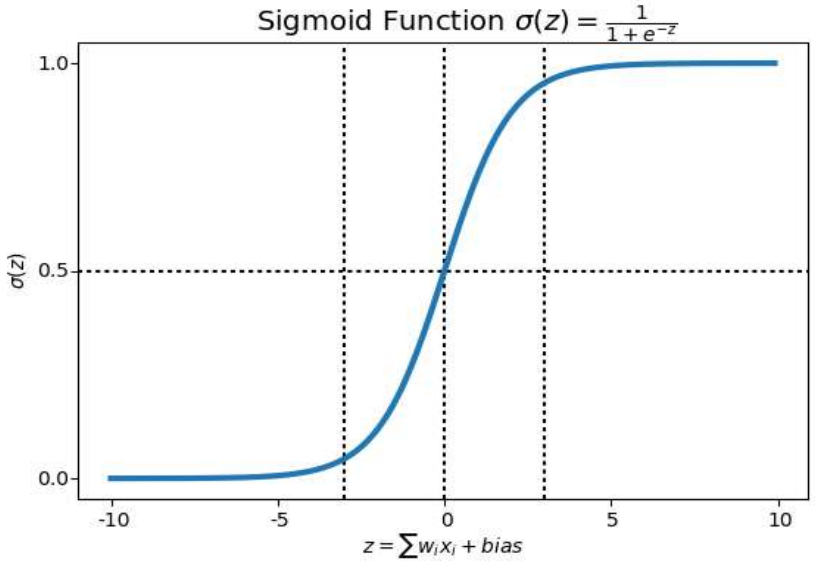
The dataset contains water quality metrics for 3276 different water bodies.

- ph: pH of water (0 to 14).
- Hardness: Capacity of water to precipitate soap in mg/L.
- Solids: Total dissolved solids in ppm.
- Chloramines: Amount of Chloramines in ppm.
- Sulfate: Amount of Sulfates dissolved in mg/L.
- Conductivity: Electrical conductivity of water in µS/cm.
- Organic_carbon: Amount of organic carbon in ppm.
- Trihalomethanes: Amount of Trihalomethanes in µg/L.
- Turbidity: Measure of light emitting property of water in NTU.
- Potability: Indicates if water is safe for human consumption. Potable - 1 and Not potable - 0

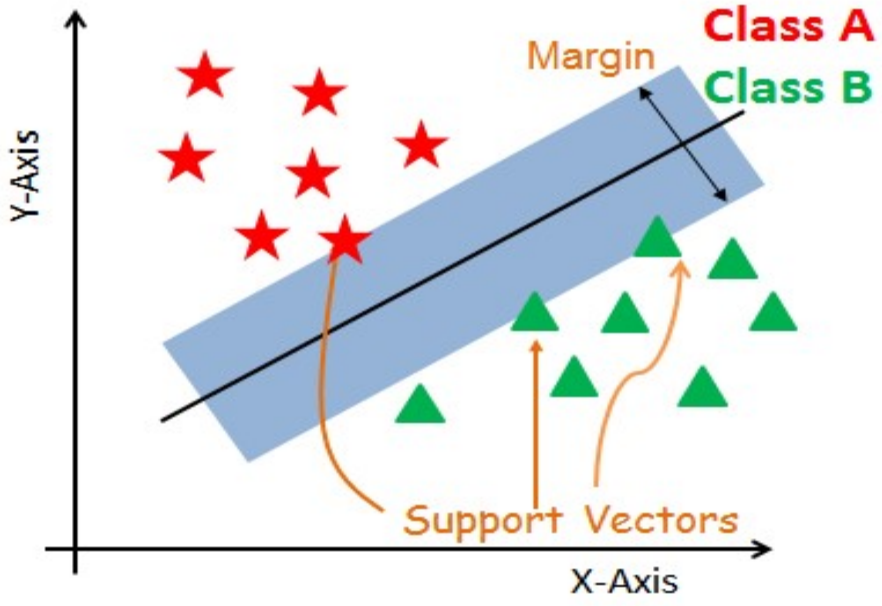
Models used for training

- Logistic Regression - Logistic Regression is named for the function used at the core of the method, the logistic function.
  The logistic function, also called the sigmoid function, was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It's an S-shaped curve that can take any real-valued number and map it into a value

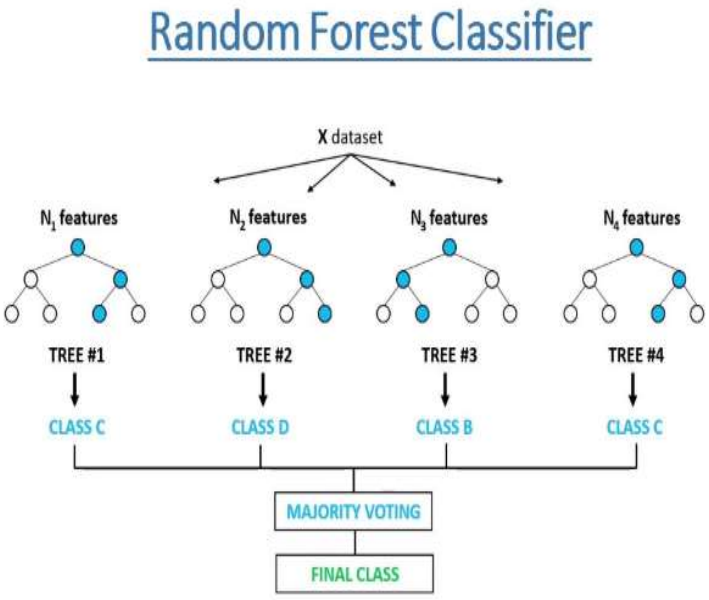
between 0 and 1, but never exactly at those limits.



- Support Vector Classifier - The objective of a Linear SVC (Support Vector Classifier) is to fit the data you provide, returning a "best fit" hyperplane that divides, or categorizes your data.
-



- Random Forest Classifier - A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.



Requirements

The following python libraries were used to perform the various actions on the dataset from loading to preprocessing to visualizing and predicting the results

```
import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

import plotly.express          as ex

import plotly.graph_objs       as go

import plotly.offline          as pyo

import scipy.stats             as stats

import pymc3                   as pm

import theano.tensor           as tt
```

NumPy is the fundamental package needed for scientific computing with Python.

Pandas - Python library used to analyze data.

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

Seaborn - An open-source Python library built on top of matplotlib. It is used for data visualization and exploratory data analysis.

Plotly - provides online graphing, analytics, and statistics tools for individuals and collaboration, as well as scientific graphing libraries for Python, R, MATLAB, Perl, Julia, Arduino, and REST.

Scikit-learn - tool for predictive data analysis built on numpy, scipy and matplotlib.

CONCLUSION

The predictive models Random Forest and Decision Tree got 100% accuracy
To determined the water potability