WATER QUALITY ANALYSIS

INTRODUCTION

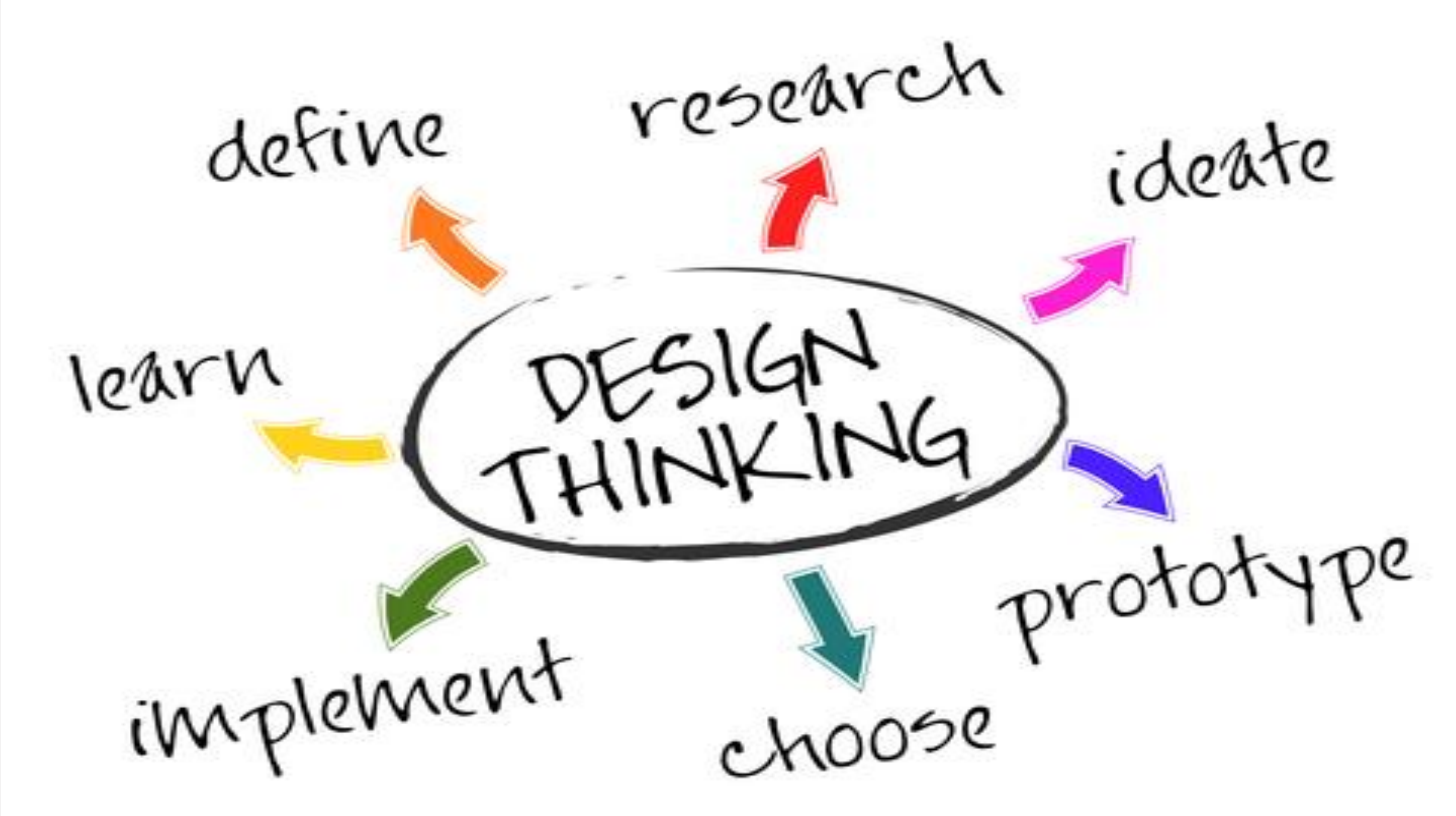
- Water quality analysis is also called hydrochemical analysis. That is to use chemical and physical methods to determine the content of various chemical components in water. Water quality analysis can be divided into three types: simple analysis, complete analysis and special analysis.

PROJECT OBJECTIVE

- River water quality analysis is ultimately performed to ensure safety—specifically, that certain chemical, physical, and biological parameters are within safe limits. Polluted water has many negative effects like threatening fish and shellfish, concentrating pollutants in the food chain, and endangering drinking water.

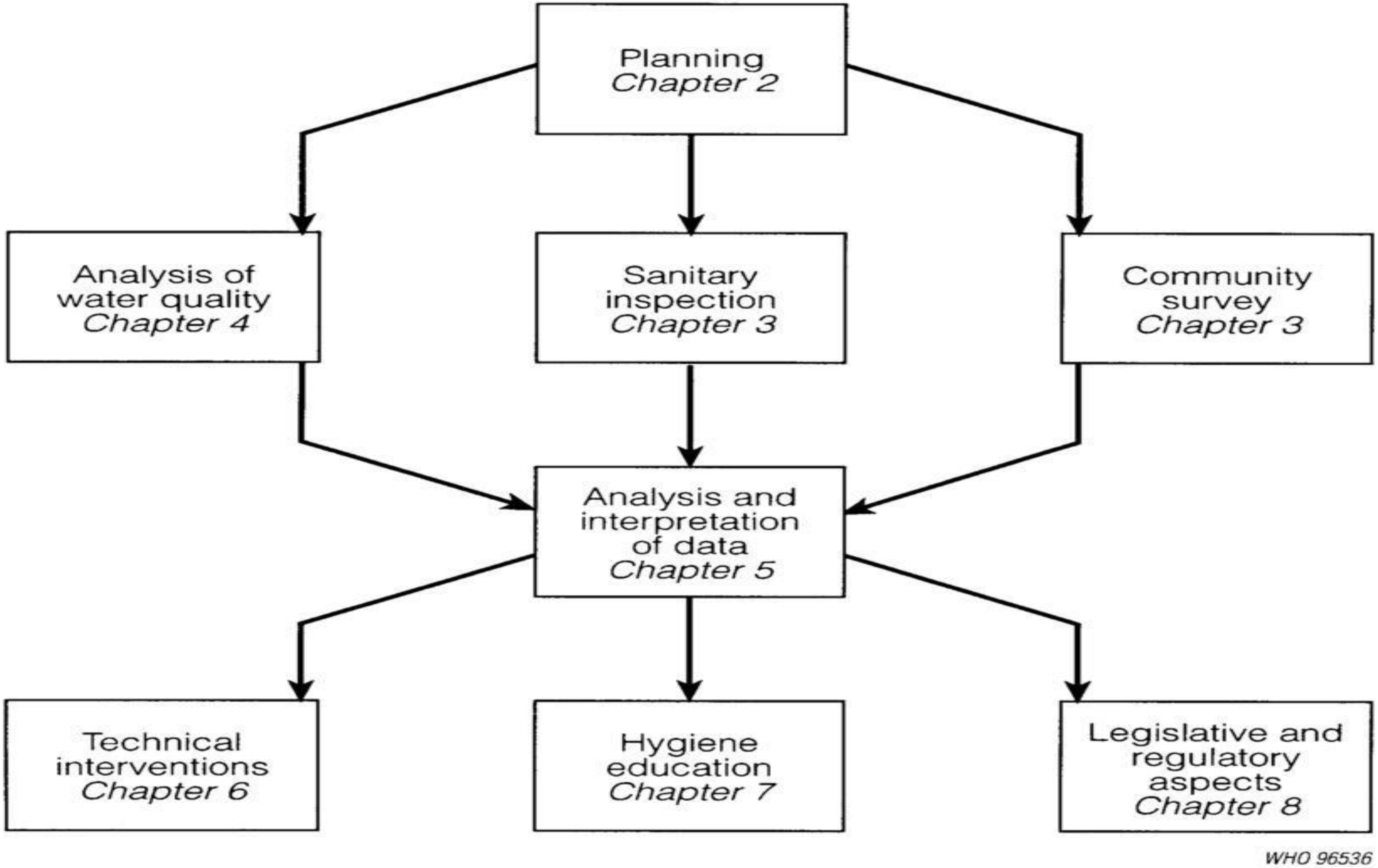
DESIGN THINKING PROCESS

- **Empathize: research your users' needs.**
- **Define: state your users' needs and problems.**
- **Ideate: challenge assumptions and create ideas.**
- **Prototype: start to create solutions.**
- **Test: try your solutions out.**


define
research
ideate
learn
DESIGN THINKING
prototype
implement
choose

DEVELOPMENT PHASES

- **Common steps in involved in water quality analysis are data preprocessing, data splitting model training and testing, and results evaluation. These are the common steps involved in development in almost all ML methods.**

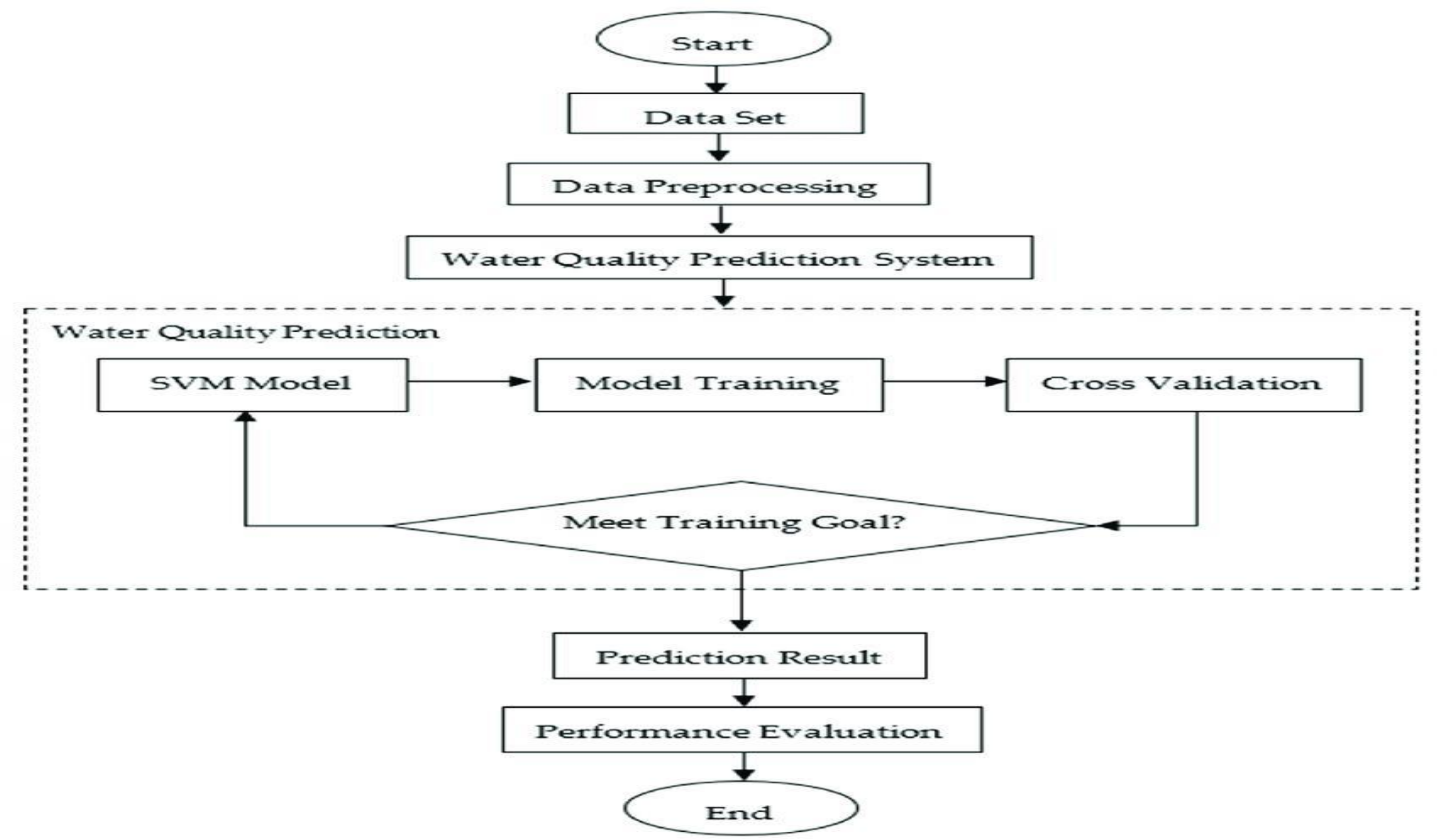

Planning
Chapter 2
Analysis of water quality
Chapter 4
Sanitary inspection
Chapter 3
Community survey
Chapter 3
Analysis and interpretation of data
Chapter 5
Technical interventions
Chapter 6
Hygiene education
Chapter 7
Legislative and regulatory aspects
Chapter 8
WHO 96536

DESCRIBES ANALYSIS OBJECTIVE

- Water quality analysis is also called hydrochemical analysis. That is to **use chemical and physical methods to determine the content of various chemical components in water**. Water quality analysis can be divided into three types: simple analysis, complete analysis and special analysis.

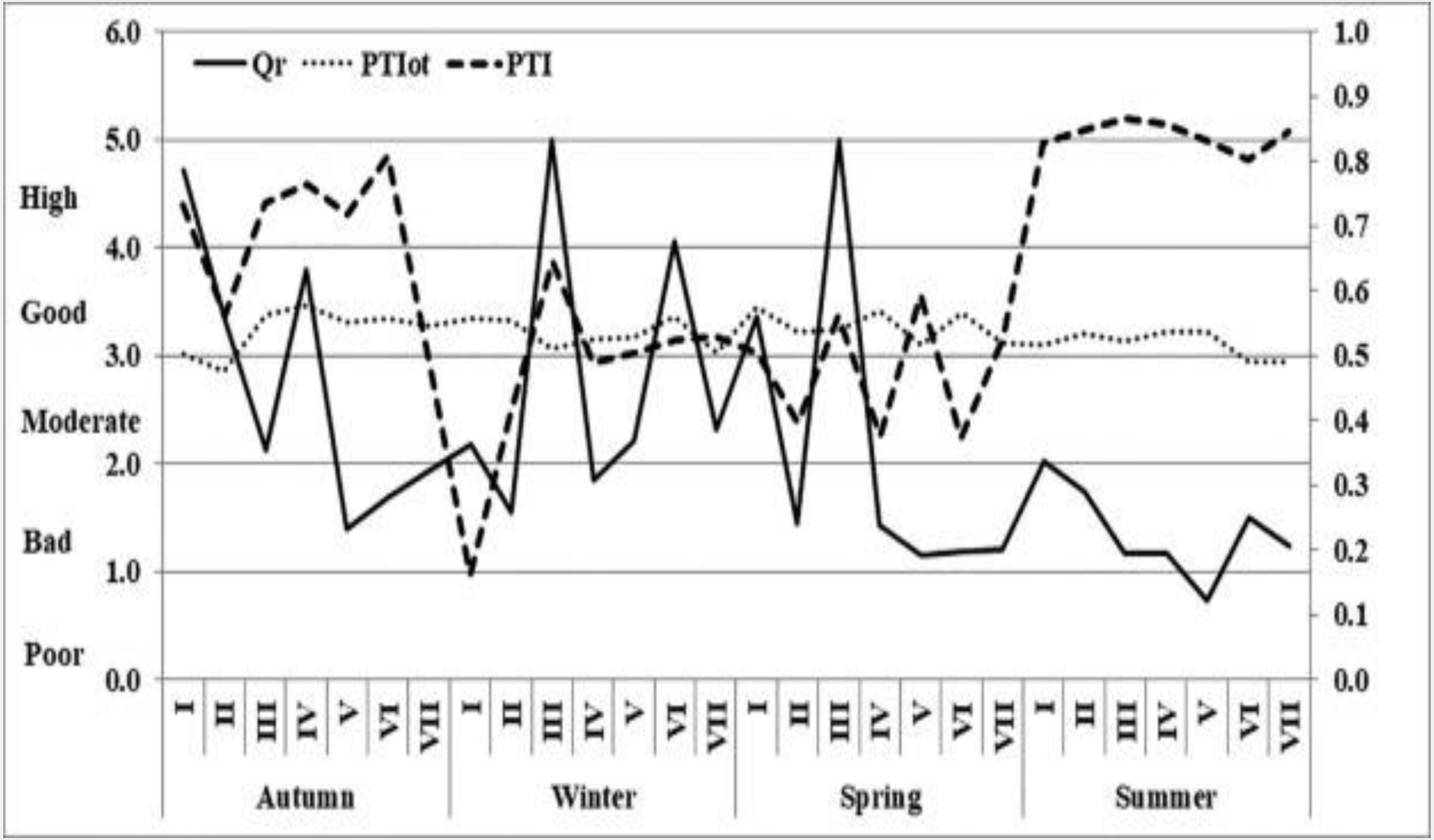
DATA PREPROCESSING

- **In data preprocessing, the raw water dataset is cleansed, decoded (transformed), and normalized for use in machine learning algorithms for model training and testing purposes. DP helps in feeding quality data into the ML models and improves the efficiency of the model training process overall.**

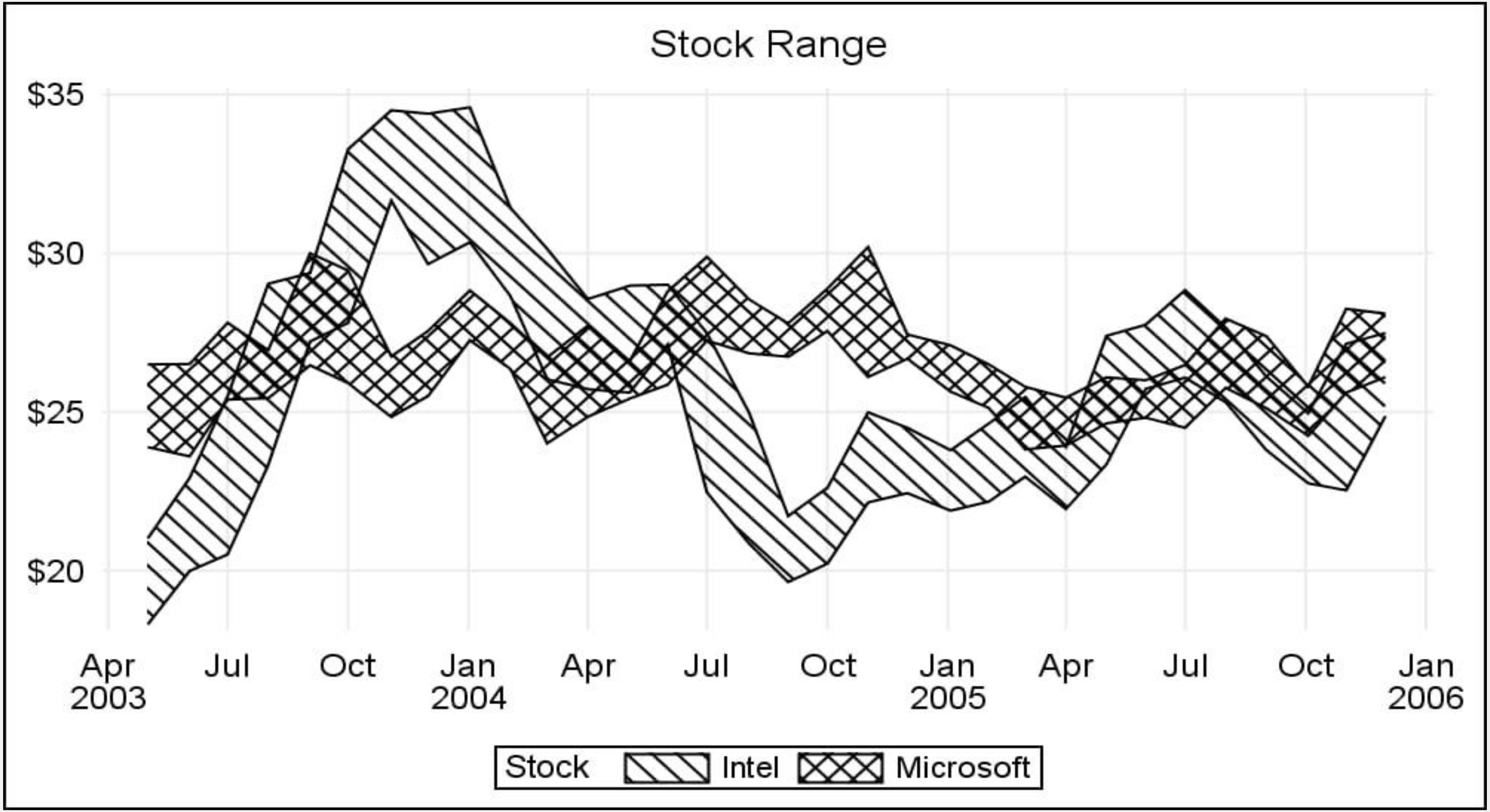

Start
Data Set
Data Preprocessing
Water Quality Prediction System
Water Quality Prediction
SVM Model
Model Training
Cross Validation
Meet Training Goal?
Prediction Result
Performance Evaluation
End

DATA VISUALITATION

- Domestic wastewater and sewage monitoring are essential for protecting public health and ensuring clean water in the environment. Through the Clean Water Act, the Environmental Protection Agency (EPA) and individual municipalities are responsible for directly governing wastewater testing strategies and procedures. The EPA both issues and approves testing methods for a wide variety of contaminants and analytes found in wastewater including trace metals, nonmetals, salts, organic compounds, bacteria, viruses, and particles such as asbestos or silica. Individual municipalities dictate what tests are necessary, how often these tests are conducted, and how data are organized

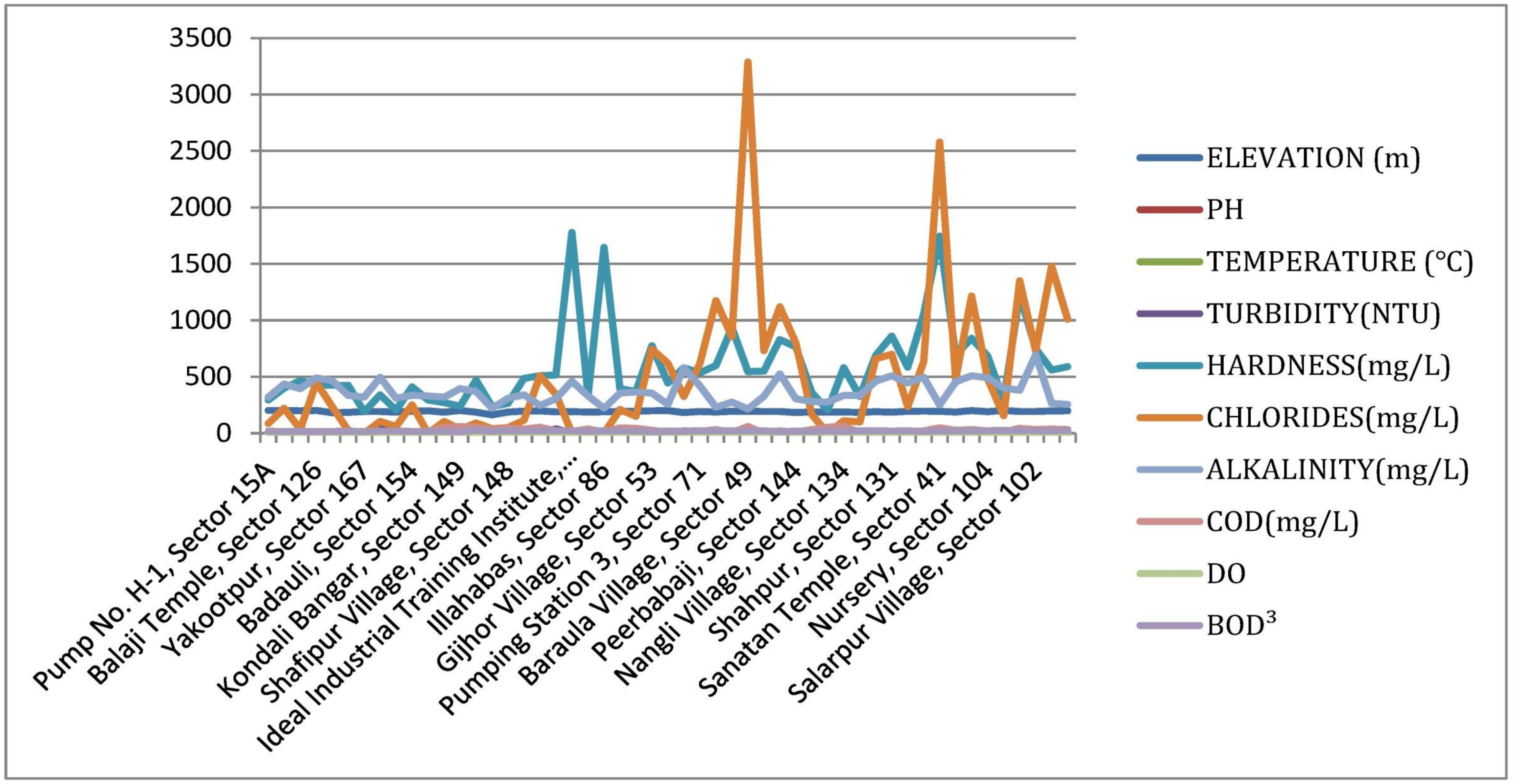


Qr
PTIot
PTI
6.0
5.0
4.0
3.0
2.0
1.0
0.0
High
Good
Moderate
Bad
Poor
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
I II III IV V VI VIII I II III IV V VI VIII I II III IV V VI VIII I II III IV V VI VII
Autumn
Winter
Spring
Summer


Stock Range
$35
$30
$25
$20
Apr 2003
Jul
Oct
Jan 2004
Apr
Jul
Oct
Jan 2005
Apr
Jul
Oct
Jan 2006
Stock
Intel
Microsoft

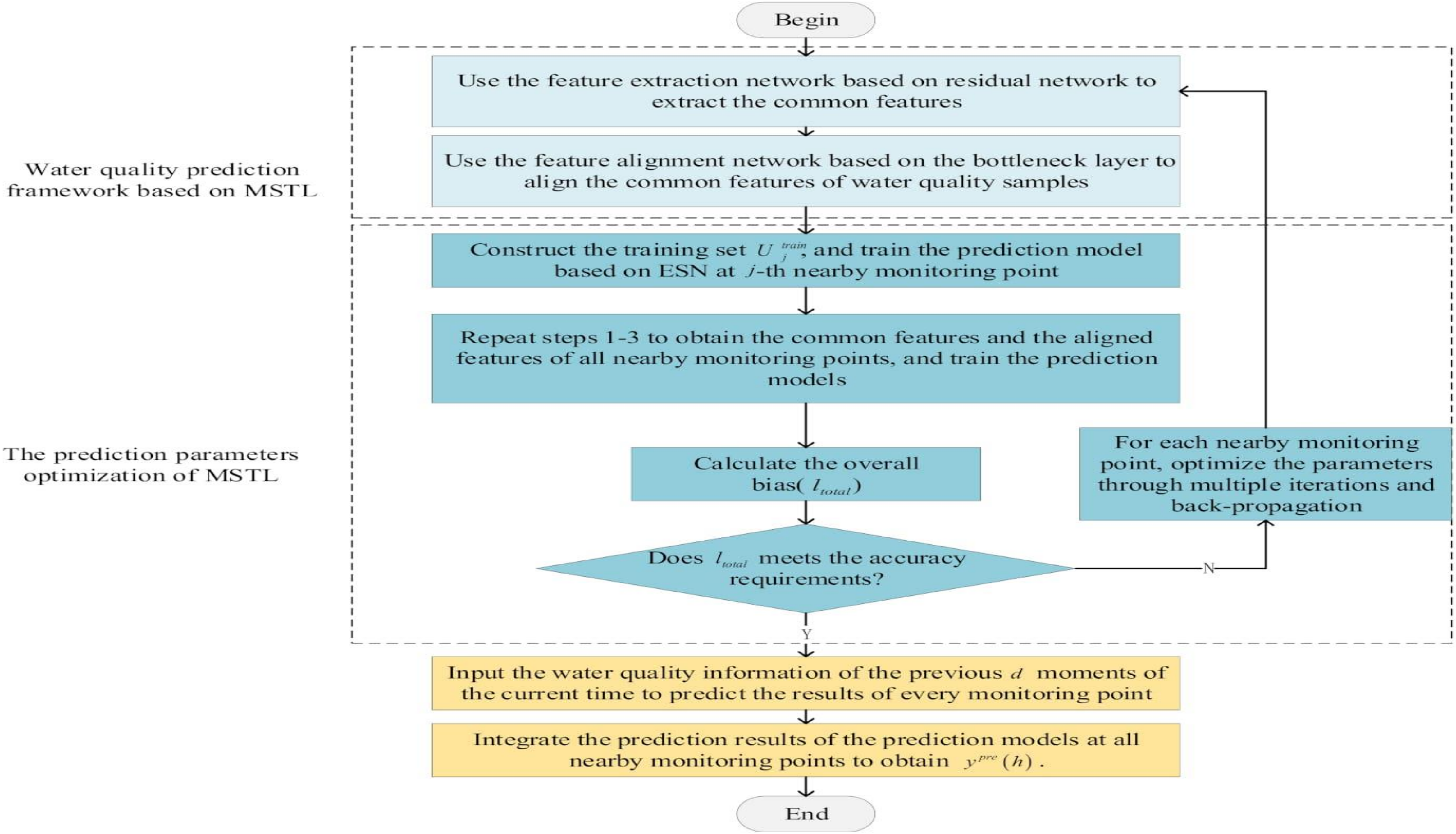
Data visualization **involves presenting data in graphical or pictorial form which makes the information easy to understand**. It helps to explain facts and determine courses of action. It will benefit any field of study that requires innovative ways of presenting large, complex information.

		Spring and Summer	Fall and Winter (Non-storm conditions)	Fall Algal Bloom	Winter Storm Event	Spring Algal Bloom	Red Tide Simulator Event
Water Quality Parameter	Units	April–August 2008	September 2008–March 2009	November 2008	Feb. 16, 2009	Late March–Early April 2009	Mid-April 2009
pH (mean)	pH Units	8.0	8.0	8.0	7.9	7.9	7.9
Temperature (range)	°C	12.0–18.1	9.5–15.8	13.4–15.6	12.8–14.1	11.9–14.2	11.1–13.4
Turbidity (range)	NTU	1.5–4.2	2.0–3.5	1.1–2.0	8–40	1.8–2.8	8–15
Particles (> 2 μm) (mean)	No. per mL	10,530	9,860	12,340	14,110	9,690	12,790
TOC (range)	mg/L	1.0–1.2	1.1–6.0	3.2	2.5	3.4–13.0	7.2
DOC (range)	mg/L	0.9–1.1	1.3–3.8	2.9	2.0	3.1–12.0	4.3
Chlorophyll (typical)	μg/L	2.3–21.2 (at Santa Cruz Wharf)	1.0	2.7	0.7	9.2	30
Algal Cell Count (typical)	Cells per liter	Not counted	15,000	28,000	< 10,000	50,000–160,000	500,000–600,


3500
3000
2500
2000
1500
1000
500
0
Pump No. H-1, Sector 15A
Balaji Temple, Sector 126
Yakootpur, Sector 167
Badauli, Sector 154
Kondali Bangar, Sector 149
Shafipur Village, Sector 148
Ideal Industrial Training Institute,...
Illahabas, Sector 86
Gijhor Village, Sector 53
Pumping Station 3, Sector 71
Baraula Village, Sector 49
Peerbabaji, Sector 144
Nangli Village, Sector 134
Shahpur, Sector 131
Sanatan Temple, Sector 41
Nursery, Sector 104
Salarpur Village, Sector 102
ELEVATION (m)
PH
TEMPERATURE (°C)
TURBIDITY(NTU)
HARDNESS(mg/L)
CHLORIDES(mg/L)
ALKALINITY(mg/L)
COD(mg/L)
DO
BOD³

PREDICTIVE MODEL FOR PORTABILITY

HydrologyWater flow rates at Lick Run Wetland followed a seasonal pattern of wet winter and spring and relatively dry summer and autumn (Fig. 2). Highest flow rates occurred in January through May 1993 with rates ranging from 222 to 248 l/min. During the growing season, inflow ranged from 67 to 104 l/min. Overall, inflow averaged (±std. error) 114±19 l/min. In contrast, inflow prior to construction was estimated from stream sampling, without the advantage of a control structure at the inflow, to be 68±9 HydrologyWater flow rates at Lick Run Wetland followed a seasonal pattern of wet winter and spring and relatively dry summer and autumn (Fig. 2). Highest flow rates occurred in January through May 1993 with rates ranging from 222 to 248 l/min. During the growing season, inflow ranged from 67 to 104 l/min. Overall, inflow averaged (±std. error) 114±19 l/min. In contrast, inflow prior to construction was estimated from stream sampling, without the advantage of a control structure at the inflow, to be 68±9


Begin
Water quality prediction framework based on MSTL
Use the feature extraction network based on residual network to extract the common features
Use the feature alignment network based on the bottleneck layer to align the common features of water quality samples
The prediction parameters optimization of MSTL
Construct the training set U_j^{train}, and train the prediction model based on ESN at j-th nearby monitoring point
Repeat steps 1-3 to obtain the common features and the aligned features of all nearby monitoring points, and train the prediction models
Calculate the overall bias(l_{total})
Does l_{total} meets the accuracy requirements?
N
For each nearby monitoring point, optimize the parameters through multiple iterations and back-propagation
Y
Input the water quality information of the previous d moments of the current time to predict the results of every monitoring point
Integrate the prediction results of the prediction models at all nearby monitoring points to obtain $y^{pre}(h)$.
End

explain how the insights from the analysis can help assess water quality and determine portability

The range of analytical techniques encompasses various **sample preparation protocols, chemical methods – namely titrations, separations, electrochemical and spectroscopic measurements – and biological methods (i.e., biosensors)**. Dissolved oxygen – a vital component in determining the health of aquatic systems.

KEY INSIGHTS

- **After carefully analysing the dataset and what our objectives were, our team decided to proceed using a simple flow process divided into sub-sections and teams.**

DATASET SOURCE

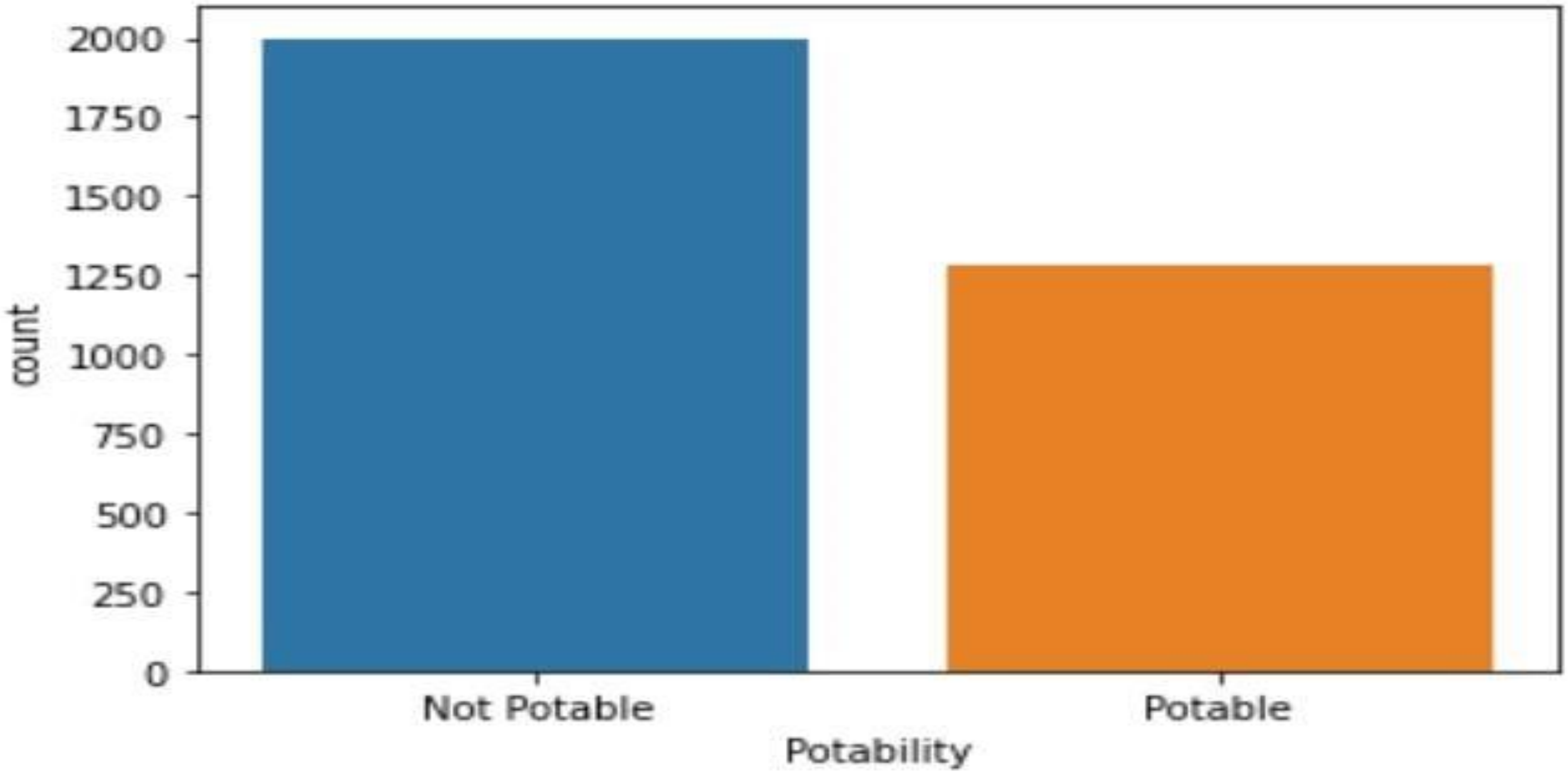
- Your dataset includes information about water sources and their qualities, such as turbidity, hardness, pH, and other parameters. This data was obtained from a crowd-sourced platform called Kaggle.

DATA WRANGLING

- **The dataset has been loaded into a data frame in the notebook using the pd.read_csv() function for further analysis and modeling. To verify the successful reading of the data file and understand its structure, the head() function is used to display a few lines of the dataset. This helps in gaining an overview of the data, including its shape, data types, and the presence of any null values.**
- **During the analysis, it was discovered that the dataset contains three features (variables/columns) with null values.**

SOURCE CODE

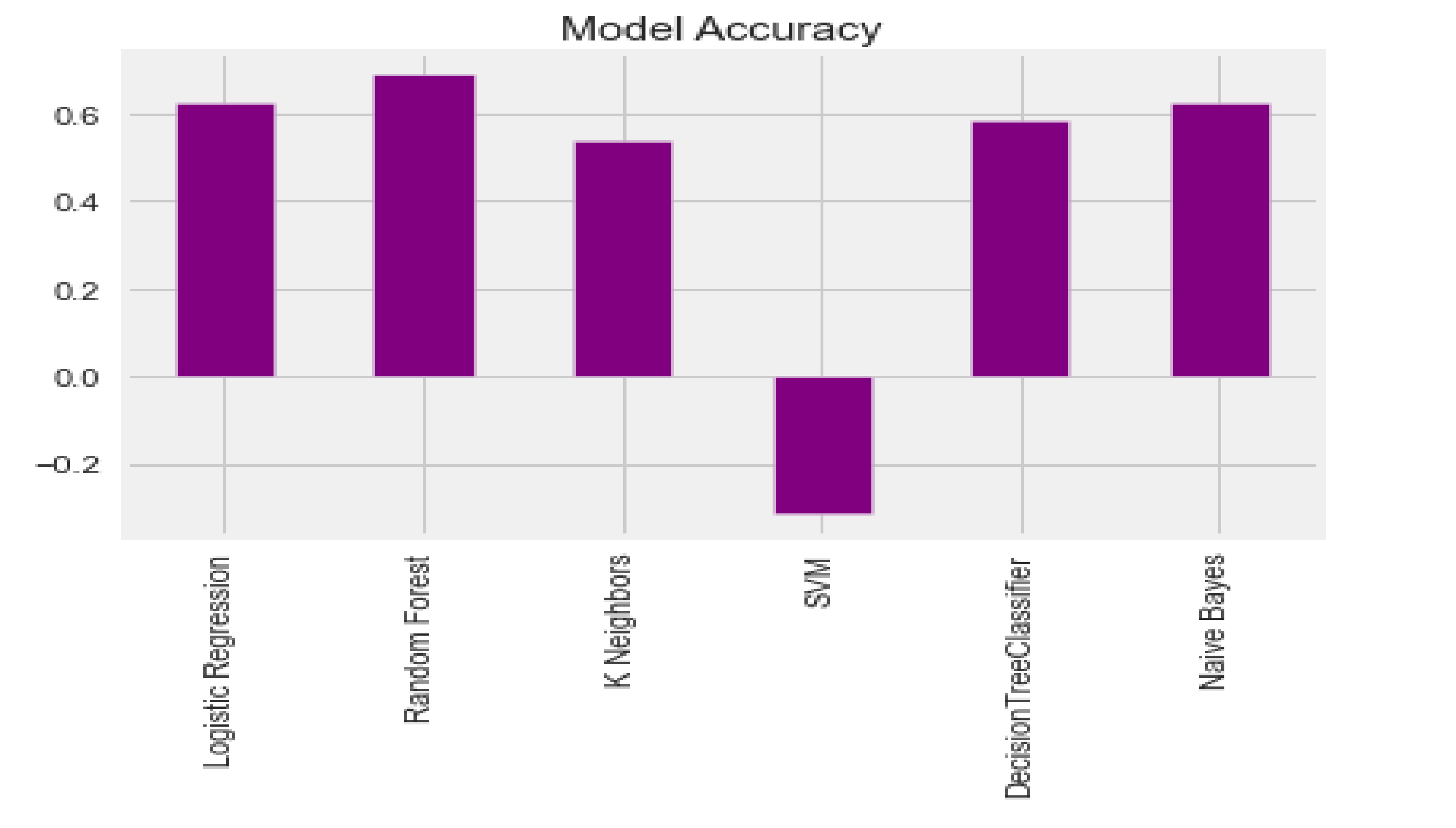
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
import plotly.express as px
import warnings
warnings.filterwarnings('ignore')
print(df.shape)
df.describe()
print(df.columns)
 df.info
print(df.nunique())
ax=sns.countplot(x =
"Potability",data= df, saturation=0.8)
plt.xticks(ticks=[0, 1], labels = ["Not
Potable", "Potable"])
plt.show()
```


2000
1750
1500
1250
1000
750
500
250
0
count
Not Potable
Potable
Potability

```python
models = {'Logistic Regression': LogisticRegression(),
          'Random Forest': RandomForestClassifier(),
          'K Neighbors': KNeighborsClassifier(),
          'SVM': SVR(),
          'DecisionTreeClassifier': DecisionTreeClassifier(),
          'Naive Bayes': GaussianNB()}
```

```python
def scaled_fit_score(models, X_train, X_test, Y_train, Y_test):
    '''
    function to fit and score machine learning models after applying scaling to fix outliers
    parameters
    ---------------------------------------------
    models: dictionary of all scikit learn machine leraning models to fit and evaluate
    X_train: training set of the predictors to fit into model
    X_test: test set of the predictors to fit into model
    Y_train: training set of the depedent varaiable
    Y_test: test set of dependent variable
    '''
#     define a random seed to make same set of prediction appear each time program is run(for reproducability)
    np.random.seed(0)
#     define a dictionary for the model scores
    scaled_model_scores = {}

#     iterate through the models dictionary items
    for name, model in models.items():
        scaled_model = Pipeline([('model', model)])
#         fit model the training set into each model in the dictionary
        scaled_model.fit(X_train, Y_train)
#     get the model score and attach it to each of the model name from model dictionary
        scaled_model_scores[name] = scaled_model.score(X_test, Y_test)
    return scaled_model_scores
```


Model Accuracy
0.6
0.4
0.2
0.0
−0.2
Logistic Regression
Random Forest
K Neighbors
SVM
DecisionTreeClassifier
Naive Bayes

CONCLUSION

- To enhance the accuracy of the classifier, it was concluded that building a neural network would be the best choice, potentially achieving up to 90% accuracy. However, it was deemed that the current classifiers provided satisfactory analysis of water potability.
- Furthermore, it is recommended to deploy the model on the web using a Python framework with good compatibility, preferably Flask. This deployment would enable the model to assist in real-life scenarios by determining the potability of water.
-