# Project Development Phase
## Project Development – Delivery Of Sprint-1

| | |
|---|---|
| Team ID | PNT2022TMID17967 |
| Project Name | Project – Efficient Water Quality Analysis and Prediction using Machine Learning |

Importing the libraries

```
In [2]: import numpy as np
        import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plt
        import warnings
```

Reading the dataset

```
In [4]: data=pd.read_csv(r'D:/PRIEE/water_dataX.csv',encoding='latin1')
        data
```

Out[4]:

| | STATION CODE | LOCATIONS | STATE | Temp | D.O. (mg/l) | PH | CONDUCTIVITY (µmhos/cm) | B.O.D. (mg/l) | NITRATENAN N+ NITRITENANN (mg/l) | C (MF |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1393 | DAMANGANGA AT D/S OF MADHUBAN, DAMAN | DAMAN & DIU | 30.6 | 6.7 | 7.5 | 203 | NAN | 0.1 | |
| 1 | 1399 | ZUARI AT D/S OF PT. WHERE KUMBARJRIA CANAL JOI... | GOA | 29.8 | 5.7 | 7.2 | 189 | 2 | 0.2 | |
| 2 | 1475 | ZUARI AT PANCHAWADI | GOA | 29.5 | 6.3 | 6.9 | 179 | 1.7 | 0.1 | |
| 3 | 3181 | RIVER ZUARI AT BORIM BRIDGE | GOA | 29.7 | 5.8 | 6.9 | 64 | 3.8 | 0.5 | |
| 4 | 3182 | RIVER ZUARI AT MARCAIM JETTY | GOA | 29.5 | 5.8 | 7.3 | 83 | 1.9 | 0.4 | |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |

## Analyze the data

```
In [5]: data.head()
```

Out[5]:

| | STATION CODE | LOCATIONS | STATE | Temp | D.O. (mg/l) | PH | CONDUCTIVITY (µmhos/cm) | B.O.D. (mg/l) | NITRATENAN N+ NITRITENANN (mg/l) | FECAL COLIFORM (MPN/100ml) | TOTAL COLIFORM (MPN/100ml)Mean | year |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1393 | DAMANGANGA AT D/S OF MADHUBAN, DAMAN | DAMAN & DIU | 30.6 | 6.7 | 7.5 | 203 | NAN | 0.1 | 11 | 27 | 2014 |
| 1 | 1399 | ZUARI AT D/S OF PT. WHERE KUMBARJRIA CANAL JOI... | GOA | 29.8 | 5.7 | 7.2 | 189 | 2 | 0.2 | 4953 | 8391 | 2014 |
| 2 | 1475 | ZUARI AT PANCHAWADI | GOA | 29.5 | 6.3 | 6.9 | 179 | 1.7 | 0.1 | 3243 | 5330 | 2014 |
| 3 | 3181 | RIVER ZUARI AT BORIM BRIDGE | GOA | 29.7 | 5.8 | 6.9 | 64 | 3.8 | 0.5 | 5382 | 8443 | 2014 |
| 4 | 3182 | RIVER ZUARI AT MARCAIM JETTY | GOA | 29.5 | 5.8 | 7.3 | 83 | 1.9 | 0.4 | 3428 | 5500 | 2014 |

```
In [6]: data.describe()
```

Out[6]:

| | year |
|---|---|
| count | 1991.000000 |
| mean | 2010.038172 |
| std | 3.057333 |
| min | 2003.000000 |
| 25% | 2008.000000 |
| 50% | 2011.000000 |
| 75% | 2013.000000 |
| max | 2014.000000 |

```
In [7]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1991 entries, 0 to 1990
Data columns (total 12 columns):
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   STATION CODE                       1991 non-null   object
 1   LOCATIONS                          1991 non-null   object
 2   STATE                              1991 non-null   object
 3   Temp                               1991 non-null   object
 4   D.O. (mg/l)                        1991 non-null   object
 5   PH                                 1991 non-null   object
 6   CONDUCTIVITY (µmhos/cm)            1991 non-null   object
 7   B.O.D. (mg/l)                      1991 non-null   object
 8   NITRATENAN N+ NITRITENANN (mg/l)   1991 non-null   object
 9   FECAL COLIFORM (MPN/100ml)         1991 non-null   object
 10  TOTAL COLIFORM (MPN/100ml)Mean     1991 non-null   object
 11  year                               1991 non-null   int64
dtypes: int64(1), object(11)
memory usage: 186.8+ KB
```

```
In [8]: data.shape
```

```
Out[8]: (1991, 12)
```

## Handling missing values

```
In [9]: data.isnull().any()
```

```
Out[9]: STATION CODE                       False
        LOCATIONS                          False
        STATE                              False
        Temp                               False
        D.O. (mg/l)                        False
        PH                                 False
        CONDUCTIVITY (µmhos/cm)            False
        B.O.D. (mg/l)                      False
        NITRATENAN N+ NITRITENANN (mg/l)   False
        FECAL COLIFORM (MPN/100ml)         False
        TOTAL COLIFORM (MPN/100ml)Mean     False
        year                               False
        dtype: bool
```

## Handling missing values 2

```
In [10]: data.dtypes
```

```
Out[10]: STATION CODE                       object
         LOCATIONS                          object
         STATE                              object
         Temp                               object
         D.O. (mg/l)                        object
         PH                                 object
         CONDUCTIVITY (µmhos/cm)            object
         B.O.D. (mg/l)                      object
         NITRATENAN N+ NITRITENANN (mg/l)   object
         FECAL COLIFORM (MPN/100ml)         object
         TOTAL COLIFORM (MPN/100ml)Mean     object
         year                                int64
         dtype: object
```

```
In [11]: data['Temp']=pd.to_numeric(data['Temp'],errors='coerce')
         data['D.O. (mg/l)']=pd.to_numeric(data['D.O. (mg/l)'],errors='coerce')
         data['PH']=pd.to_numeric(data['PH'],errors='coerce')
         data['B.O.D. (mg/l)']=pd.to_numeric(data['B.O.D. (mg/l)'],errors='coerce')
         data['CONDUCTIVITY (µmhos/cm)']=pd.to_numeric(data['CONDUCTIVITY (µmhos/cm)'],errors='coerce')
         data['NITRATENAN N+ NITRITENANN (mg/l)']=pd.to_numeric(data['NITRATENAN N+ NITRITENANN (mg/l)'],errors='coerce')
         data['TOTAL COLIFORM (MPN/100ml)Mean']=pd.to_numeric(data['TOTAL COLIFORM (MPN/100ml)Mean'],errors='coerce')
         data.dtypes
```

```
Out[11]: STATION CODE                       object
         LOCATIONS                          object
         STATE                              object
         Temp                              float64
         D.O. (mg/l)                       float64
         PH                                float64
         CONDUCTIVITY (µmhos/cm)           float64
         B.O.D. (mg/l)                     float64
         NITRATENAN N+ NITRITENANN (mg/l)  float64
         FECAL COLIFORM (MPN/100ml)         object
         TOTAL COLIFORM (MPN/100ml)Mean    float64
         year                                int64
         dtype: object
```

```
[13]: data.isnull().sum()
```

```
t[13]: STATION CODE                       0
       LOCATIONS                          0
       STATE                              0
       Temp                              92
       D.O. (mg/l)                       31
       PH                                 8
       CONDUCTIVITY (µmhos/cm)           25
       B.O.D. (mg/l)                     43
       NITRATENAN N+ NITRITENANN (mg/l)  225
       FECAL COLIFORM (MPN/100ml)         0
       TOTAL COLIFORM (MPN/100ml)Mean    132
       year                               0
       dtype: int64
```

## Handling missing values 3

```
In [14]: data['Temp'].fillna(data['Temp'].mean(),inplace=True)
         data['D.O. (mg/l)'].fillna(data['D.O. (mg/l)'].mean(),inplace=True)
         data['PH'].fillna(data['PH'].mean(),inplace=True)
         data['B.O.D. (mg/l)'].fillna(data['B.O.D. (mg/l)'].mean(),inplace=True)
         data['CONDUCTIVITY (µmhos/cm)'].fillna(data['CONDUCTIVITY (µmhos/cm)'].mean(),inplace=True)
         data['NITRATENAN N+ NITRITENANN (mg/l)'].fillna(data['NITRATENAN N+ NITRITENANN (mg/l)'].mean(),inplace=True)
         data['TOTAL COLIFORM (MPN/100ml)Mean'].fillna(data['TOTAL COLIFORM (MPN/100ml)Mean'].mean(),inplace=True)
```

```
In [15]: data.drop(['FECAL COLIFORM (MPN/100ml)'],axis=1,inplace=True)
```

```
In [17]: data=data.rename(columns={'D.O. (mg/l)':'do'})
         data=data.rename(columns={'CONDUCTIVITY (µmhos/cm)':'co'})
         data=data.rename(columns={'B.O.D. (mg/l)':'bod'})
         data=data.rename(columns={'NITRATENAN N+ NITRITENANN (mg/l)':'na'})
         data=data.rename(columns={'TOTAL COLIFORM (MPN/100ml)Mean':'tc'})
         data=data.rename(columns={'STATION CODE':'station'})
         data=data.rename(columns={'LOCATIONS':'location'})
         data=data.rename(columns={'STATE':'state'})
         data=data.rename(columns={'PH':'ph'})
```