SPRINT 1 –DATA PREPROCESSING

DATA PREPROCESSING:

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

- Getting the dataset
- Importing libraries
- Importing datasets
- Analyzing the data
- Finding Missing Data
- Encoding Categorical Data
- Splitting dataset into training and test set
- Feature scaling

IMPORTING LIBRARIES:

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns
```

IMPORTING DATASETS:

```
df = pd.read_csv("water_dataX.csv")
```

ANALYSING THE DATA:

```
df.head();

df.describe();
df.shape

df.info();
```

FINDING MISSING DATA:

```
df.isnull().any();

df.isnull().sum();

for feature in df.columns:
    if df[feature].isnull().sum()>0:
        print(f"{feature} : {round(df[feature].isnull().mean(),4)*100}%")

-------Fill missing values with median

for feature in df.columns:
    df[feature].fillna(df[feature].median() , inplace = True)

------- find dublicate rows in dataset

duplicate = df[df.duplicated()]

duplicate
```

### Finding missing value1

```
d=pd.read_csv("water_dataX.csv")

pd.isnull(d["Solids"])
```

# ##Finding missing value2

```
d=pd.read_csv("water_dataX.csv")

pd.isnull(d["Turbidity"])
```

### Finding missing value3

```
d=pd.read_csv("water_dataX.csv")

pd.isnull(d["ph"])
```

```
-----------removing outliers

Q1 = df.quantile(0.25)

Q3 = df.quantile(0.75)

IQR = Q3 - Q1

print(IQR)
```

SPLITTING DEPENDENT AND INDEPENDENT COLUMN

```
X = df.iloc[: , : -1]

y = df.iloc[ : , -1]
```

SPLITTING DATASET INTO TESTING AND TRAINING:

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.3, random_state= 5)
```