**10. WATER QUALITY ANALYSIS**

**Phase 4: Development Part 2**

**ABSTRACT:**

The data used for the research is water quality data, produced during the STREAMES (Stream reach Management, an Expert System) project, initially aiming at producing tools for increasing the quality of European rivers.In this report one will find examples of data imputation, regression, classification, clusterization andfeature selection tasks using machine learning algorithms, such as: random forest, support vector ma-chines, neural networks, k-nearest neighbours, and k-means clustering.

**FEATURES DESCRIPTION:**

1. **ph**: pH of 1. water (0 to 14).
2. **Hardness**: Capacity of water to precipitate soap in mg/L.
3. **Solids**: Total dissolved solids in ppm.
4. **Chloramines**: Amount of Chloramines in ppm.
5. **Sulfate**: Amount of Sulfates dissolved in mg/L.
6. **Conductivity**: Electrical conductivity of water in μS/cm.
7. **Organic\_carbon**: Amount of organic carbon in ppm.
8. **Trihalomethanes**: Amount of Trihalomethanes in μg/L.
9. **Turbidity**: Measure of light emiting property of water in NTU.
10. **Potability**: Indicates if water is safe for human consumption. Potable - 1 and Not potable – 0

**SAMPLE PROGRAM**:

**LIBRARIES**:

import numpy as np

import pandas as pd

from warnings import filterwarnings

from collections import Counter

# Visualizations Libraries

import matplotlib.pyplot as plt

import seaborn as sns

import plotly

import plotly.offline as pyo

import plotly.express as px

import plotly.graph\_objs as go

pyo.init\_notebook\_mode()

import plotly.figure\_factory as ff

import missingno as msno

# Data Pre-processing Libraries

from sklearn.preprocessing import StandardScaler,MinMaxScaler

from sklearn.model\_selection import train\_test\_split

# Modelling Libraries

from sklearn.linear\_model import LogisticRegression,RidgeClassifier,SGDClassifier,PassiveAggressiveClassifier

from sklearn.linear\_model import Perceptron

from sklearn.svm import SVC,LinearSVC,NuSVC

from sklearn.neighbors import KNeighborsClassifier,NearestCentroid

from sklearn.tree import DecisionTreeClassifier

from sklearn.ensemble import RandomForestClassifier,AdaBoostClassifier,GradientBoostingClassifier

from sklearn.naive\_bayes import GaussianNB,BernoulliNB

from sklearn.ensemble import VotingClassifier

# Evaluation & CV Libraries

from sklearn.metrics import precision\_score,accuracy\_score

from sklearn.model\_selection import RandomizedSearchCV,GridSearchCV,RepeatedStratifiedKFold







# Importing The Dataset

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

df.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 3276 entries, 0 to 3275

Data columns (total 10 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

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

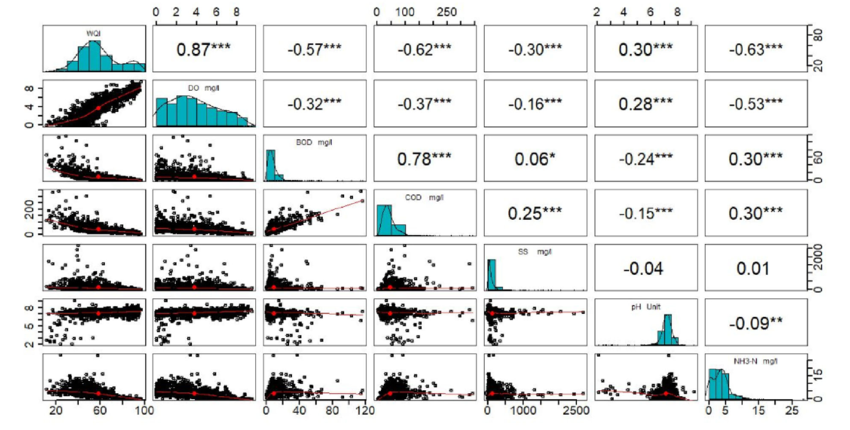
8 Turbidity 3276 non-null float64

9 Potability 3276 non-null int64

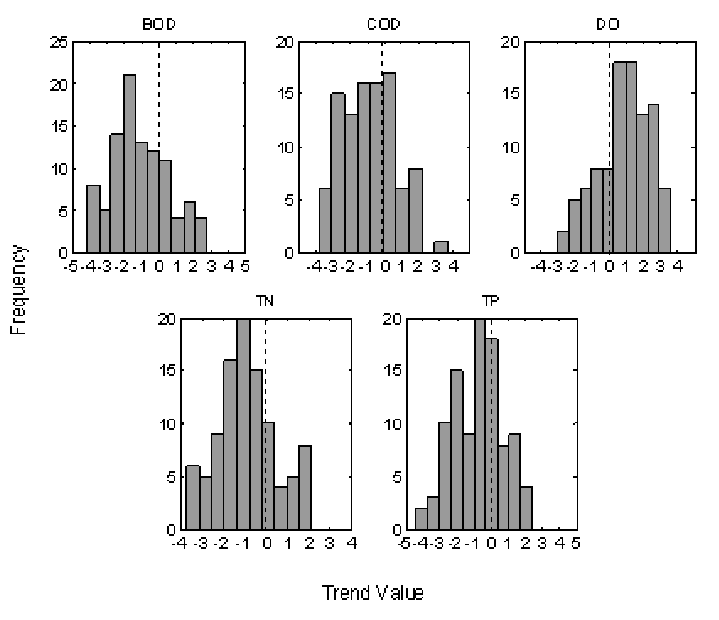
dtypes: float64(9), int64(1)

memory usage: 256.1 KB

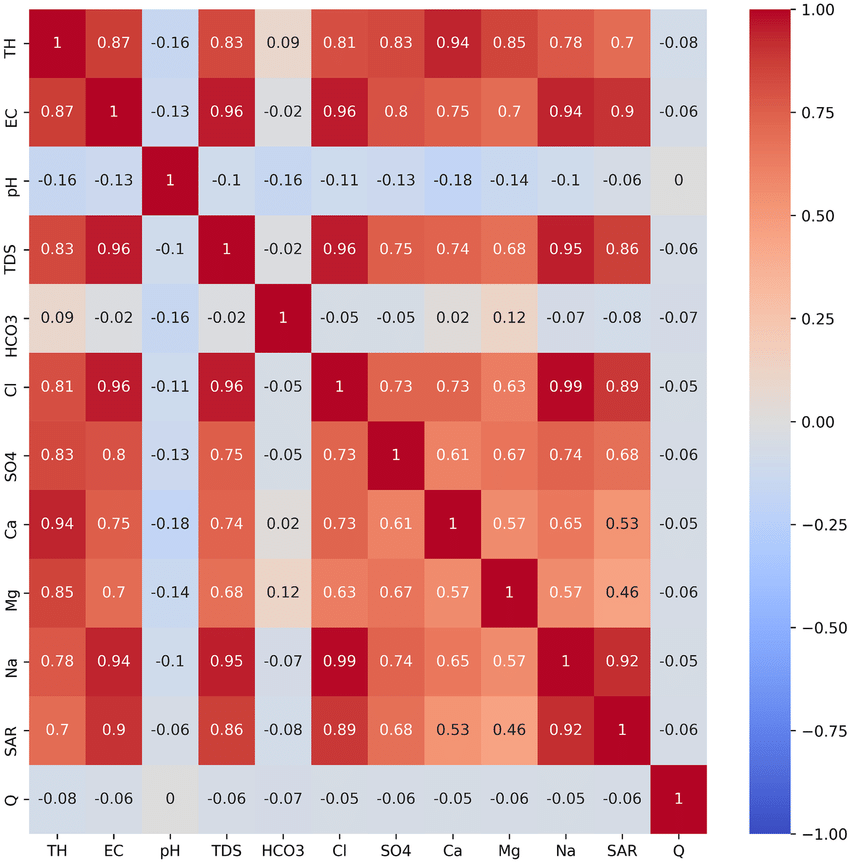
**SACTTER PLOT:**



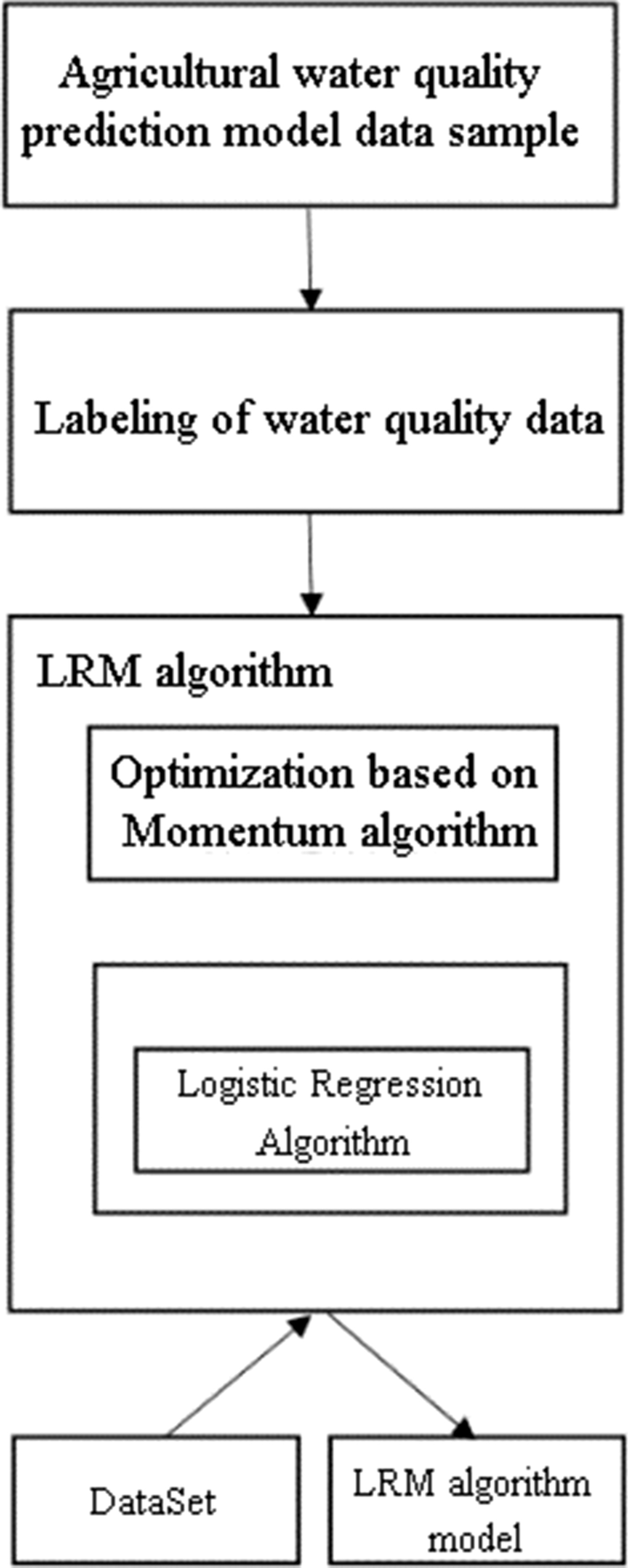
**HISTOGRAM OF WATER QUALITY ANALYSIS:**

****

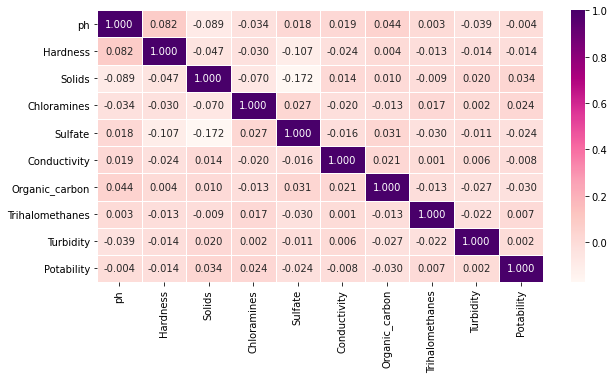
**CORRELATION METRICES:**

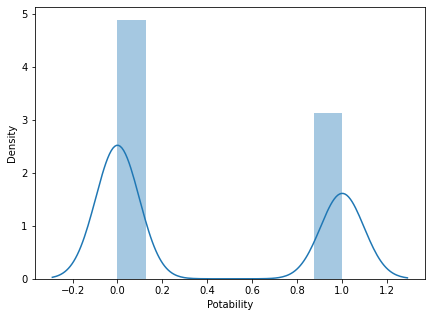
****

**FLOW CHART OF WATER QUALITY PREDICTION MODEL:**

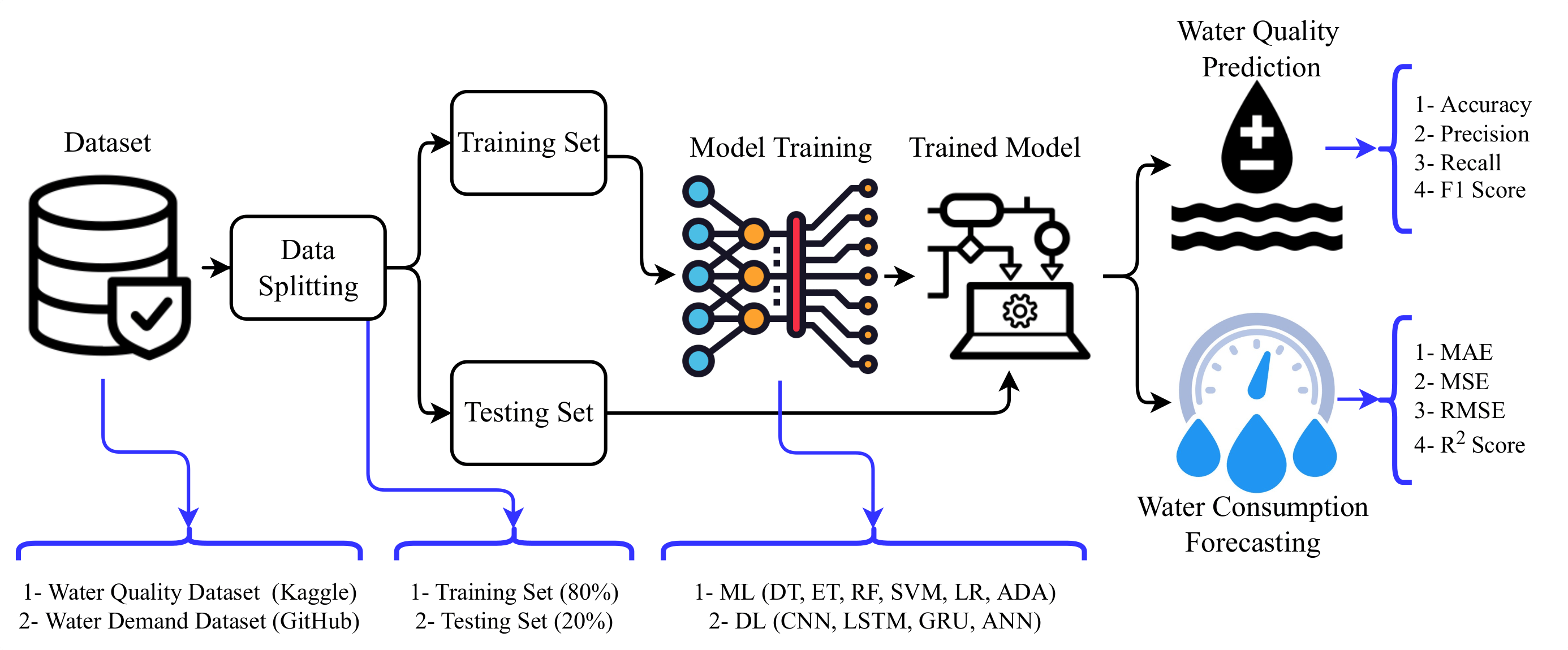
****

**RANDOM FOREST:**

****

****

**DEVELOPEMENT OF WATER QUALITY ANALYSIS**

****

## Conclusion

Both logistic regression and K-NN perform worse in accuracy and sensitifity. Both models perform worse on predicting true possitive value, just having 2.08% sensitifity on logistic regression and 21.42% on K-NN. Both models almost have same specifity performance, 98.8% for linear regression and 95.02% for K-NN. However, if our focus is on predicting that the water is not potabile, so we can choose the linear regression model with higher accuracy and specifity than K-NN model.