10. WATER QUALITY ANALYSIS

Phase 5:DAC_Phase5

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 and feature selection tasks using machine learning algorithms, such as: random forest, support vector ma-chines, neural networks, k-nearest neighbours, and k-means clustering.

OBJECTIVES:

The objective of water quality monitoring is to obtain quantitative information on the physical, chemical, and biological characteristics of water via statistical sampling (Sanders et al. 1987). The type of information sought depends on the objectives of the monitoring programme.

Design Thinking:

Analysis Objectives:

The objective of water quality monitoring is to obtain quantitative information on the physical, chemical, and biological characteristics of water via statistical sampling

Data Collection:

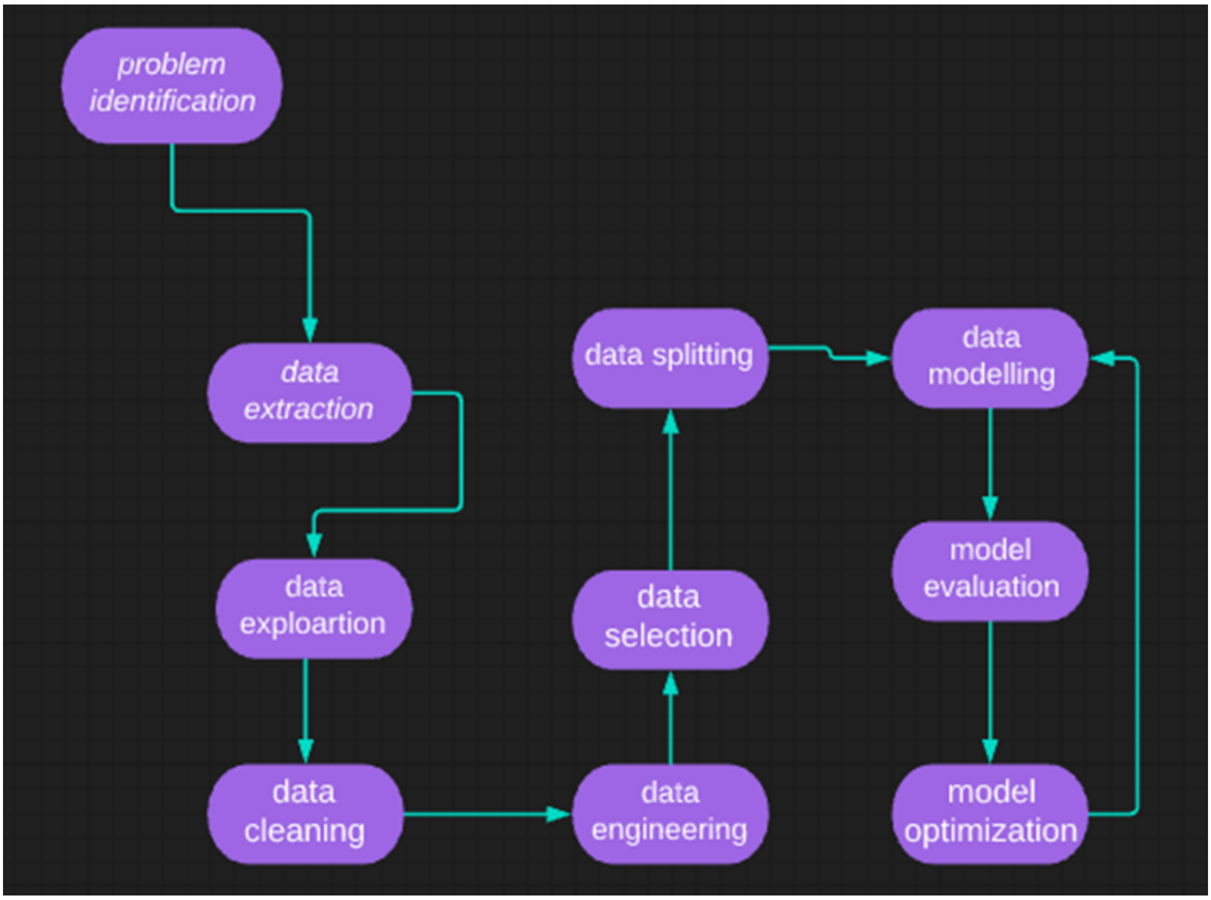
Water quality analysis is also called hydrochemical analysis. That is to use chemical and physical methods to determine the content of various chemical components in water. Water quality analysis can be divided into three types: simple analysis, complete analysis and special analysis.

Predictive Modeling:

We can define predictive models as quantitative mathematical projections that use statistical classifiers to determine the probability of a specific water quality event in the future. Predictive modelling can also be applied to an unknown water quality event, even after it has occurred.

Project Steps :

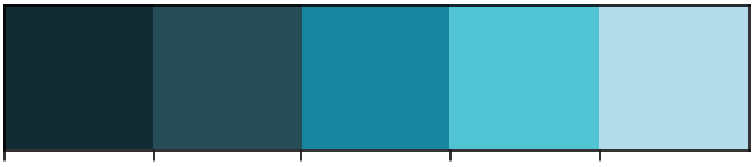
Common steps in involved in water quality analysis are data preprocessing, data splitting model training and testing, and results evaluation. These are the common steps involved in development in almost all ML methods

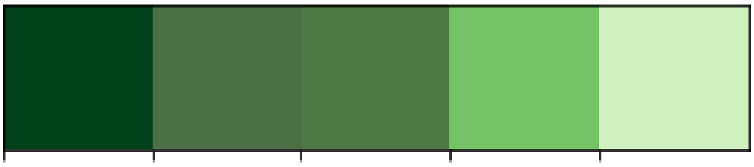


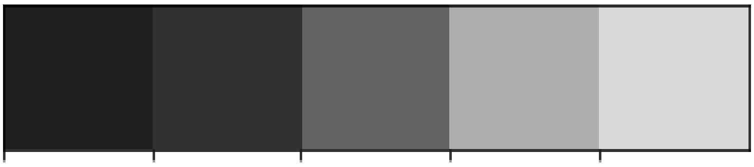
SAMPLE PROGRAM:

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
```



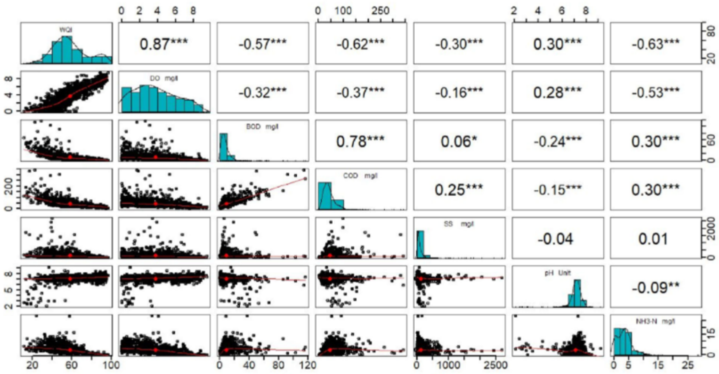




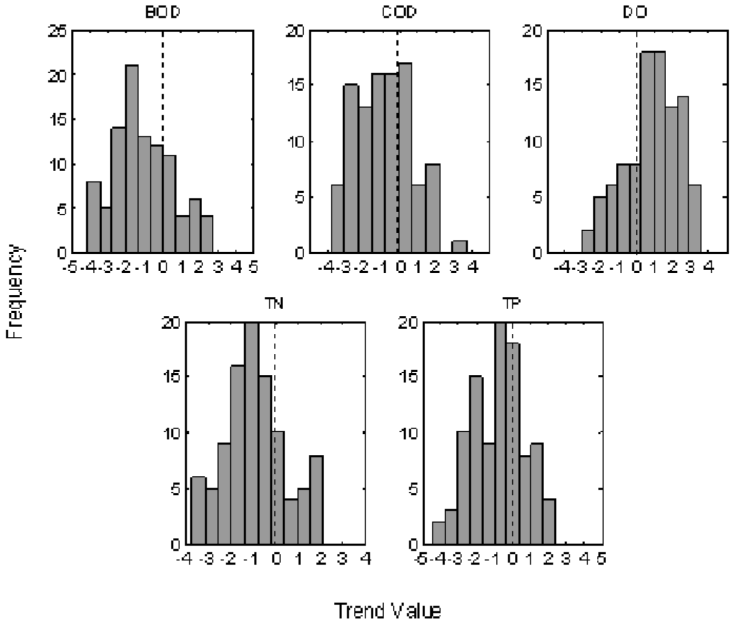
Importing The Dataset

```
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
```

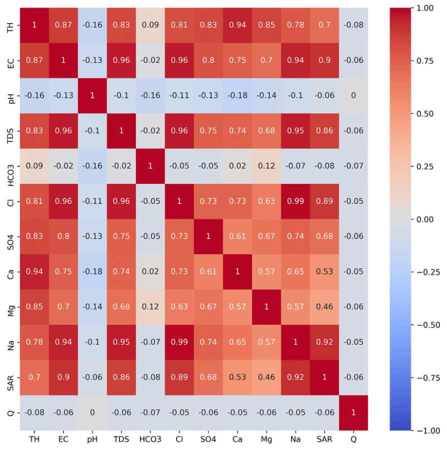
SACTTER PLOT:



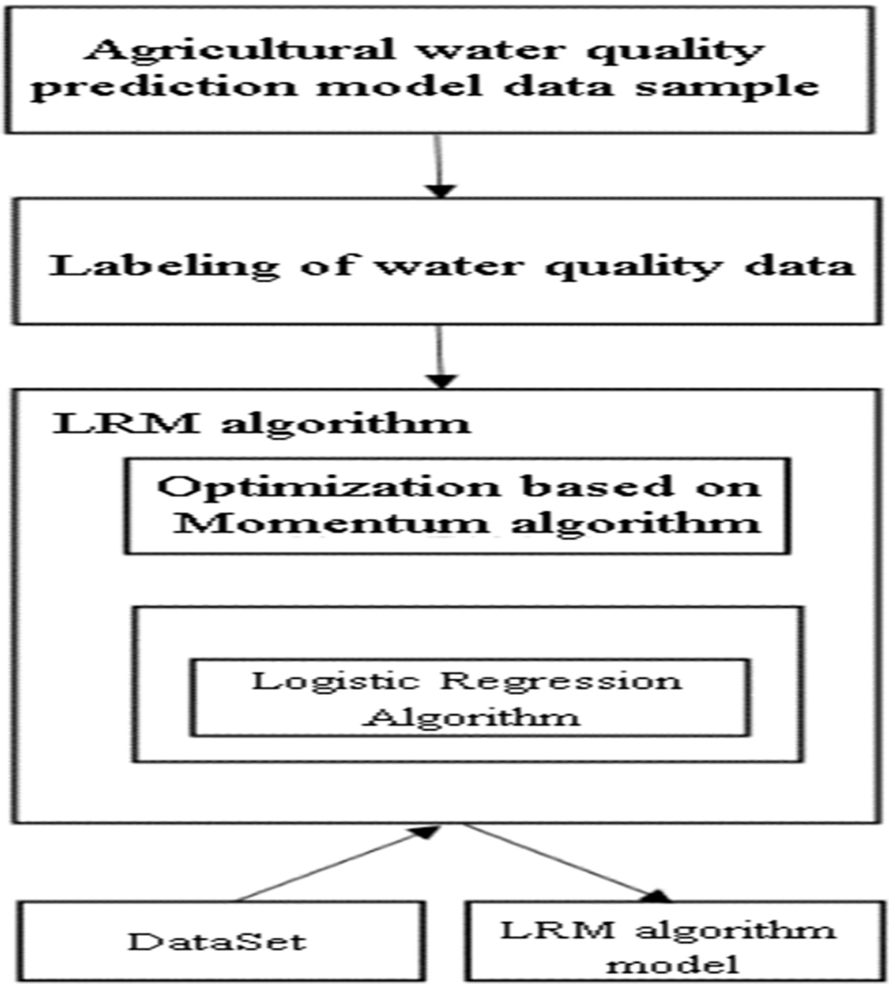
HISTOGRAM OF WATER QUALITY ANALYSIS:



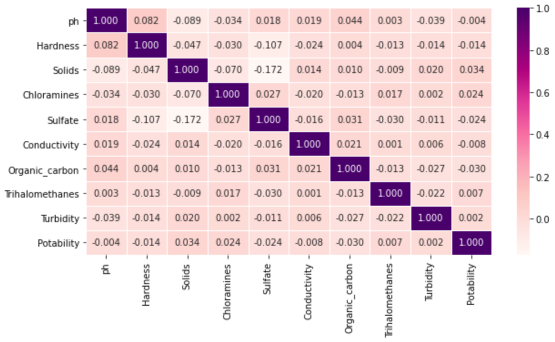
CORRELATION METRICES:

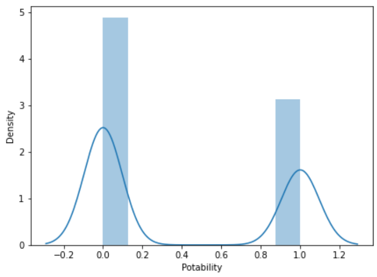


FLOW CHART OF WATER QUALITY PREDICTION MODEL:

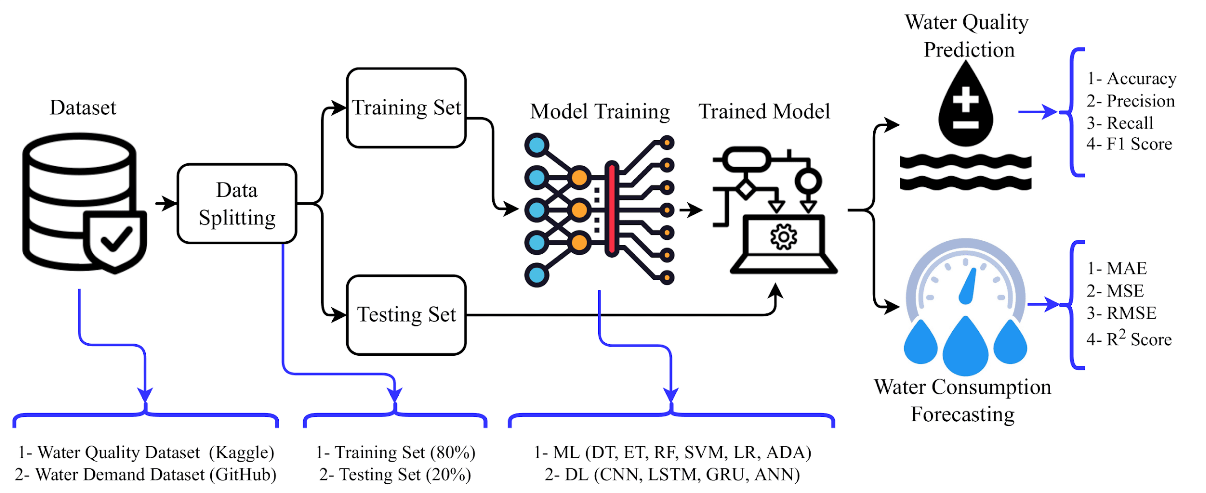

Agricultural water quality prediction model data sample
Labeling of water quality data
LRM algorithm
Optimization based on Momentum algorithm
Logistic Regression Algorithm
DataSet
LRM algorithm model

RANDOM FOREST:

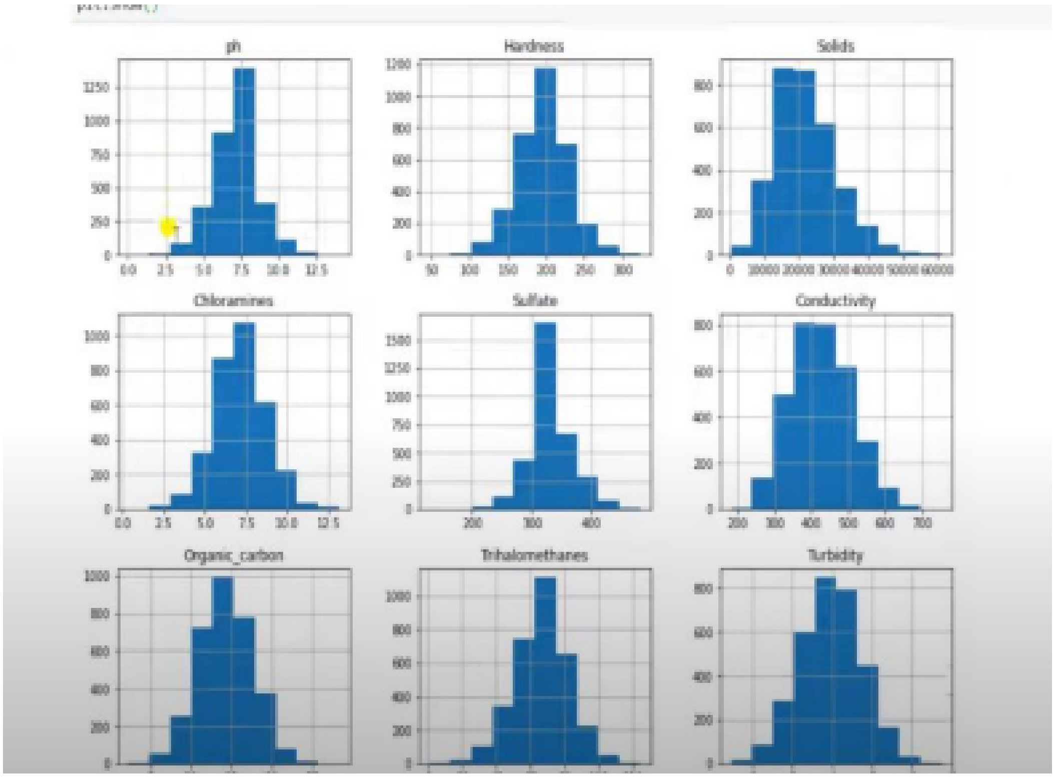


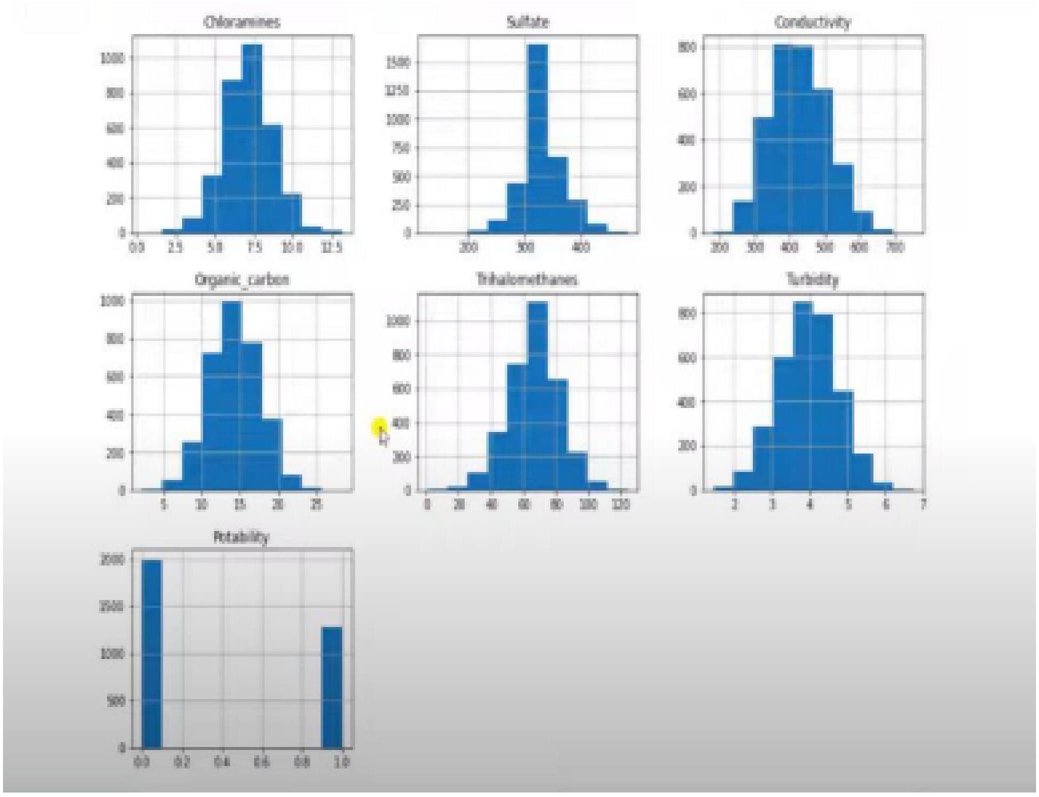


DEVELOPEMENT OF WATER QUALITY ANALYSIS



Expected Output :





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.