EFFICIENT WATER QUALITY ANALYSIS AND PREDICTION USING MACHINE LEARNING

NALAIYA THIRAN PROJECT BASED LEARNING

ON

PROFESSIONAL READINESS FOR INNOVATION, EMPLOYABILITY AND ENTREPRENEURSHIP

A PROJECT REPORT

S.G.KEERTHANA	410119106024
S.NARMADHA	410119106039
N.PRIYADHARSHINI	410119106051
M.THARANI	410119106068

BACHELOR OF ENGINEERING

IN

ELECTRONIC AND COMMUNICATION ENGINEERING

ADHI COLLEGE OF ENGINEEERING AD TECHNOLOGY

(An ISO Certified Institution Approved by the AICTE,

New Delhi& Affiliated by Anna university)

KANCHIPURAM-631 502

NOVEMBER 2022

ADHI COLLEGE OF ENGINEEERING AD TECHNOLOGY
(An ISO Certified Institution Approved by the AICTE,
New Delhi& Affiliated by Anna university)
KANCHIPURAM-631 502

INTERNAL MENTOR

Mrs. S. ANNIE ANGELINE PREETHI
Assistant Professor
Department Of Electronic And Comunication Engineering
Adhi College Of Engineering And Technology
Kanchipuram-631 502

INTERNAL EVALUATOR

Mr.K.DINESH BABU
Head Of The Department
Department Of Electronic And Comunication Engineering
Adhi College Of Engineering And Technology
Kanchipuram-631 502

INDUSTRY MENTOR

MISS.LALITHA GAYATHRI
IBM

ABSTRACT

Water makes up about 70% of the earth's surface and is one of the most important sources vital to sustaining life. Rapid urbanization and industrialization have led to a deterioration of water quality at an alarming rate, resulting in harrowing diseases. Water quality has been conventionally estimated through expensive and time-consuming lab and statistical analyses, which render the contemporary notion of real-time monitoring moot. The alarming consequences of poor water quality necessitate an alternative method, which is quicker and inexpensive. With this motivation, this research explores a series of supervised machine learning algorithms to estimate the water quality index (WQI), which is a singular index to describe the general quality of water, and the water quality class (WQC), which is a distinctive class defined on the basis of the WQI. The proposed methodology employs four input parameters, namely, temperature, turbidity, pH and total dissolved solids. Of all the employed algorithms, gradient boosting, with a learning rate of 0.1 and polynomial regression, with a degree of 2, predict the WQI most efficiently, having a mean absolute error (MAE) of 1.9642 and 2.7273, respectively. Whereas multi-layer perceptron (MLP), with a configuration of (3,7), classifies the WQC most efficiently, with an accuracy of 0.8507. The proposed methodology achieves reasonable accuracy using a minimal number of parameters to validate the possibility of its use in real time water quality detection systems.

TABLE OF CONTENTS

CHAPTER-1

INTORDUCTION

The water quality index (WQI) has been used to identify threats to water quality and to support better water resource management. This study combines a machine learning algorithm, WQI, and remote sensing spectral indices (difference index, DI; ratio index, RI; and normalized difference index, NDI) through fractional derivatives methods and in turn establishes a model for estimating and assessing the WQI. The results show that the calculated WQI values range between 56.61 and 2,886.51. We also explore the relationship between reflectance data and the WQI. The number of bands with correlation coefficients passing a significance test at 0.01 first increases and then decreases with a peak appearing after 1.6 orders.

This paper intends to address this issue by suggesting a model based upon Machine Learning techniques in order to predict the future water quality trends of a particular area with the help of current water quality data.

CHAPTER-2
OBJECTIVE

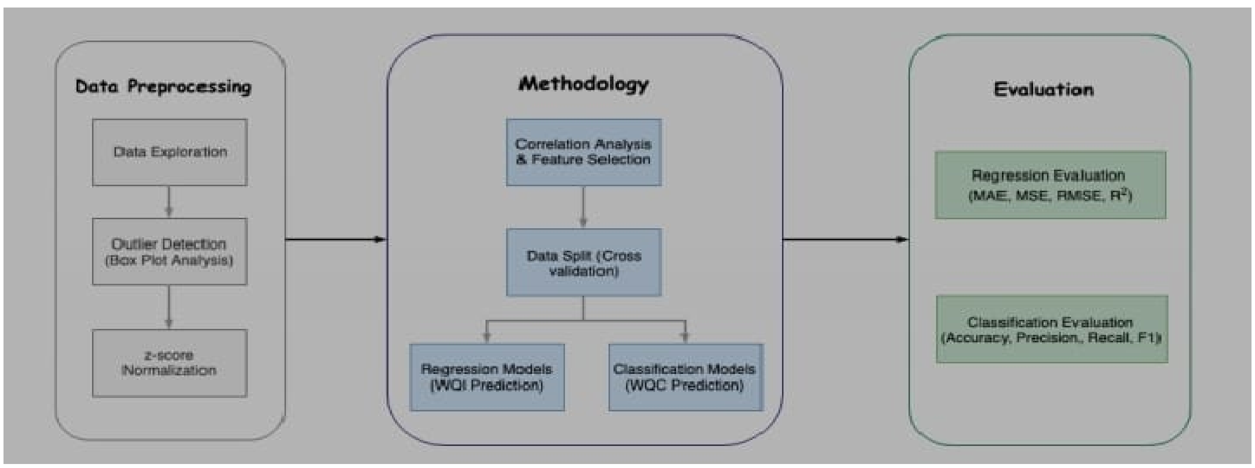
The goal of this research is to develop efficient models to predict values of water quality parameters based upon their present values.

The basic idea of this research is to devise a comprehensive methodology that analyzes and predicts water quality of particular regions with the help of certain water quality parameters. These parameters include physical, biological or chemical factors which influence water quality. There are certain quality standards set up by international organizations like World Health Organization (WHO) and Environmental Protection Agency (EPA), which serve as a benchmark for determining the quality of water. In its document “Parameters of Water Quality”, EPA mentions a total of 101 parameters.

After much experimentation, the results reflect that gradient boosting and polynomial regression predict the WQI best with a mean absolute error (MAE) of 1.9642 and 2.7273, respectively, whereas multi-layer perceptron (MLP) classifies the WQC best, with an accuracy of 0.8507.

CHAPTER-3
DATA PREPROCESSING

The data used for this research was obtained from PCRWR and it was cleaned by performing a box plot analysis, discussed in this section. After the data were cleaned, they were normalized using q-value normalization to convert them to the range of 0–100 to calculate the WQI using six available parameters. Once the WQI was calculated, all original values were normalized using z-score, so they were on the same scale. The complete procedure is detailed next.




```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import seaborn as sns
        from collections import defaultdict
        from sklearn.model_selection import train_test_split
        from sklearn.preprocessing import StandardScaler
        from sklearn.metrics import accuracy_score, f1_score, r2_score
        from sklearn.ensemble import RandomForestRegressor, AdaBoostClassifier,
        from sklearn.linear_model import LinearRegression
```

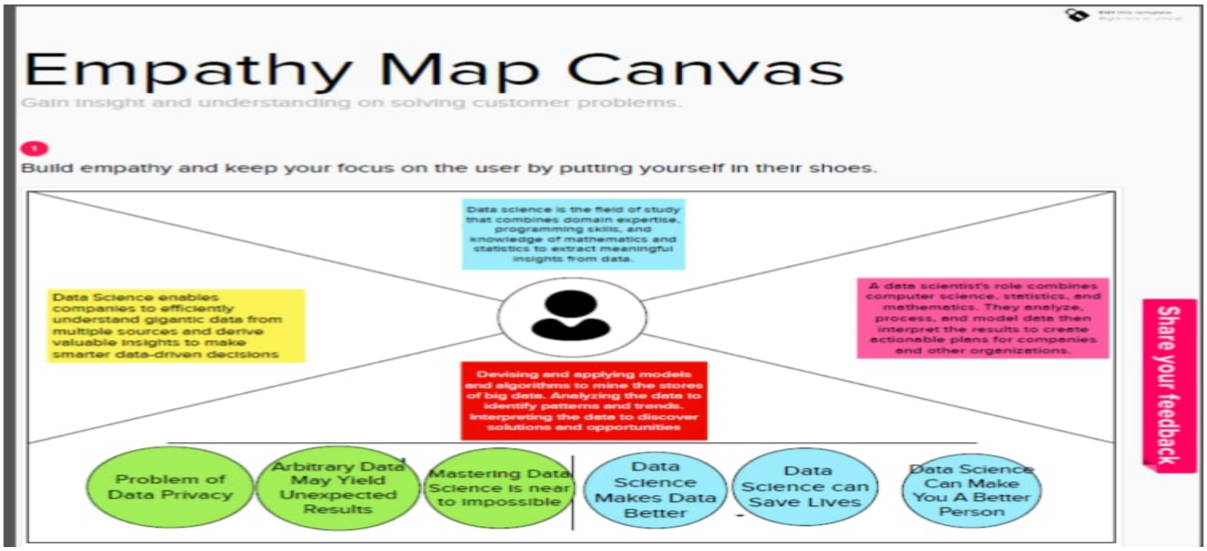
CHAPTER-4
IDEATION PHASE

4.1 LITRATURE SURVEY:

PROJECT NAME	AUTHORS NAME	OUTCOME
Predictive Analysis of Water Quality Parameters using Deep Learning	Archana Solanki , Himanshu Agrawal, Kanchan Khare	Merit of the unsupervised learning algorithms are evaluated on the basis of metrics such as mean absolute error and mean square error to examine the error rate of prediction.
Predicting and Analyzing Water Quality using Machine Learning	Yafra Khan , Chai Soo See	The deteriorating quality of natural water resources like lakes, streams and estuaries, is one of the direst and most worrisome issues faced by humanity.The goal of this study is to develop a water quality prediction model with the help of water quality factors using Artificial Neural Network (ANN) and time-series analysis.For the

		purpose of evaluating the performance of model, the performance evaluation measures used are Mean-Squared Error (MSE), Root Mean-Squared Error (RMSE) and Regression Analysis.
Efficient Water Quality Prediction Using Supervised Machine Learning	Umair Ahmed, Rafia Mumtaz, Hirra Anwar, Asad A.shah, Rabia Irfan, Jose Garcia-nieto	Rapid urbanization and industrialization have led to a deterioration of water quality an alarming rate, resulting in harrowing diseases.In this motivation, this research explores a series of supervised machine learning algorithms to estimate the water quality index , which is a singular index to describe the general quality of water, and the water quality class (WQC), which is a distinctive class defined on the basis of the WQI.

4.2 EMPATHY MAP:



4.3 BRAINSTROM:

CHAPTER-5

PROJECT DESIGN PHASE -I

5.1 PROPOSED SOLUTION :

S.NO	PARAMETER	DESCRIPTION
2	Idea/Solution Describtion	The Data to develop a water quality prediction model with the help of water quality factors using Artificial Neural Network (ANN) and time-series analysis.
3	Novelty/ Uniqueness	The data includes the measurements of 4 parameters which affect and influence water quality. For the purpose of evaluating the performance of model, the performance evaluation measures used are Mean-Squared Error (MSE), Root Mean-Squared Error (RMSE) and Regression Analysis.
4	Social Imapct/Customer	Surface waters and aquifers can be

	Satisfaction	contaminated by various chemicals, microbes. Disinfection of drinking water has dramatically reduced the prevalence of waterborne diseases by the evaluating the data
5	Business Model (Revenue Model)	Machine learning can provide solutions for water pollution controll, water quality improvement and watershed ecosystem security management.
6	Scalability of the Solution	The solution can be used almost various source of water quality factors , watersheds and so on.Thus it is scalable for all types of prediction.

5.2 PROBLEM SOLUTION FIT:

Focus on J&P, tap into BE, understand RC

1.CUSTOMER SEGMENT(S) CS

The aim of the world's water use is for agriculture, industry, and electricity. The most common water uses include: Drinking and Household Needs. And also analysis the water quality to drinking purpose.

6. CUSTOMER CONSTRAINTS CC

If the water is not at standard quality it is an serious thread to all the people. Because water essential one for all to sustain.

5. AVAILABLE SOLUTIONS AS

The main solution is to analysis the water quality for the purpose of drinking,household,agriculture due to the healthy life of living things

The available solution is finding water quality index (WQI) and water quality class(WQC).

2. JOBS-TO-BE-DONE / PROBLEMS J&P

It is very difficult to find the pure drinking Identify the associated casual factor. water.Because it need more proof to be an qualified water.The rising water pollution ,resulting in lab testing to imperative reliability and accuracy and directly include the drinking water. The main problem is impurities present in the water.

9. PROBLEM ROOT CAUSE RC

Identify appropriate solution. Collect sufficient amount of data, Root Cause Analysis (RCA) is a comprehensive term encompassing a collection of problem solving methods used to identify the real cause of a non-conformance or quality problem. Root Cause Analysis is the process of defining understanding and solving a problem.

7. BEHAVIOUR BE

Water quality analyst analyse the quality and develop policies and plans for control the factor which produce impurities.They conduct chemical,physical and biological test to define water quality standard.

J&P, tap into BE, understand RC

3. TRIGGERS TR

This triggers to discover the pattern in user data and then make prediction based on intricate pattern for analyzing the quality of water. It also helps to improve the efficiency of water and more protected to drink water.

4. EMOTIONS: BEFORE / AFTER EM

Before there is no technology ,customer faced many problems ,they have solutions but it does not sacrifice the customer to analyse the water quality so it cause problem in health issue like disease such as diarrhoea, dysentery, hepatitis, typhoid, polio and cholera.. But now a days it is decreased .The problems are also cleared and sacrifice the water due to the methods of finding pure water by using Water monitoring system .

10. YOUR SOLUTION SL

Using Advanced Artificial Intelligence seven significant parameters and developed models were evaluated based on some statistical parameters based on naïve bayes algorithm, K Nearest Neighbour(KNN), Support Vector Machine(SVM) and Linear regression algorithm,

8.CHANNELS of BEHAVIOUR

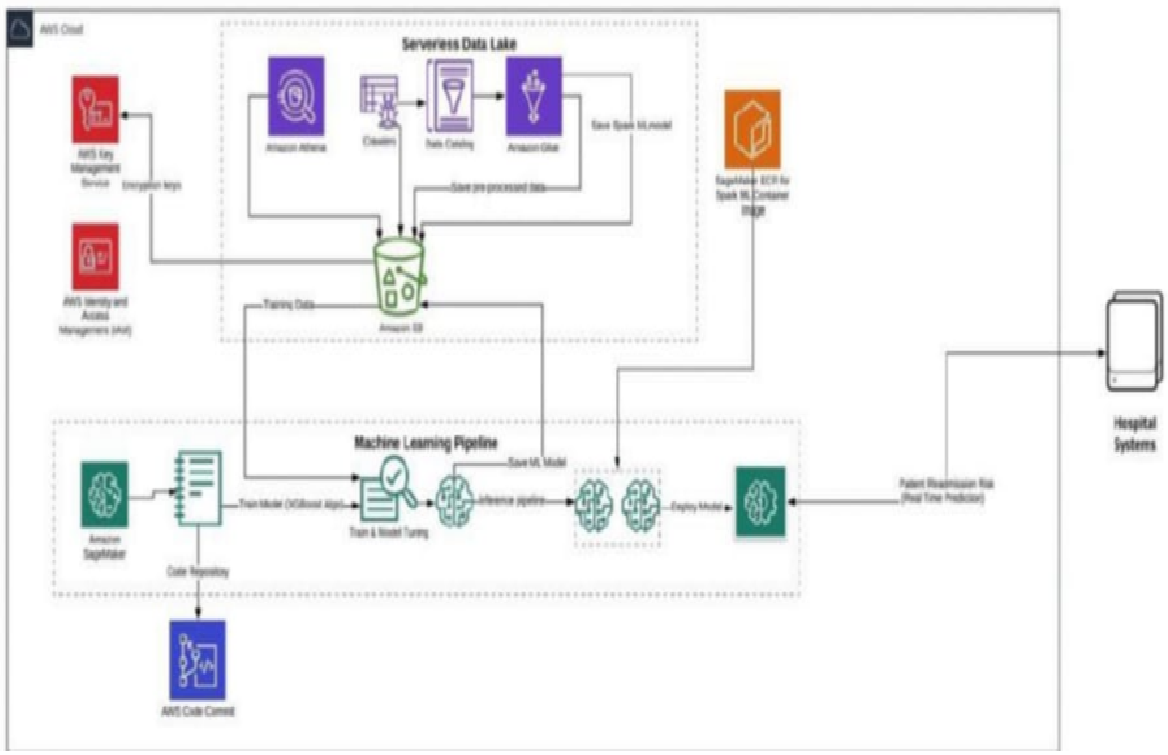
ONLINE

Helps to notify the data preprocessing information.

OFFLINE

Helps to notify the data preprocessing information.

5.3 SOLUTION ARCHITECTURE:



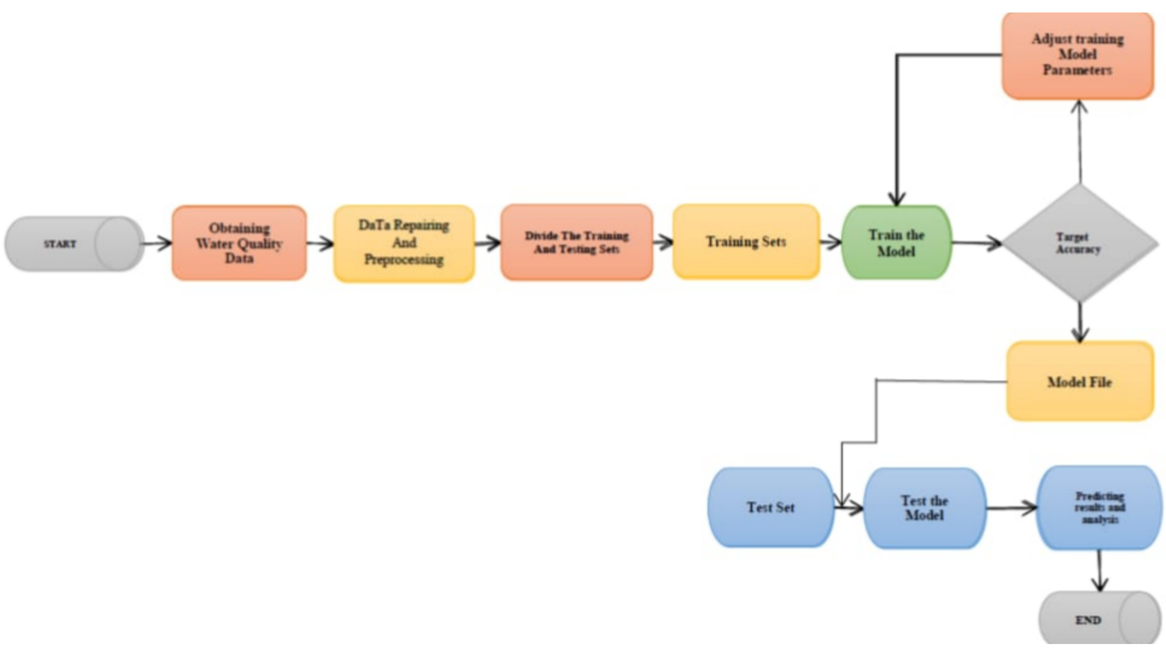
CHAPTER-6
PROJECT DESGN PHASE -II

6.1 CUSTOMER JOURNEY MAP:

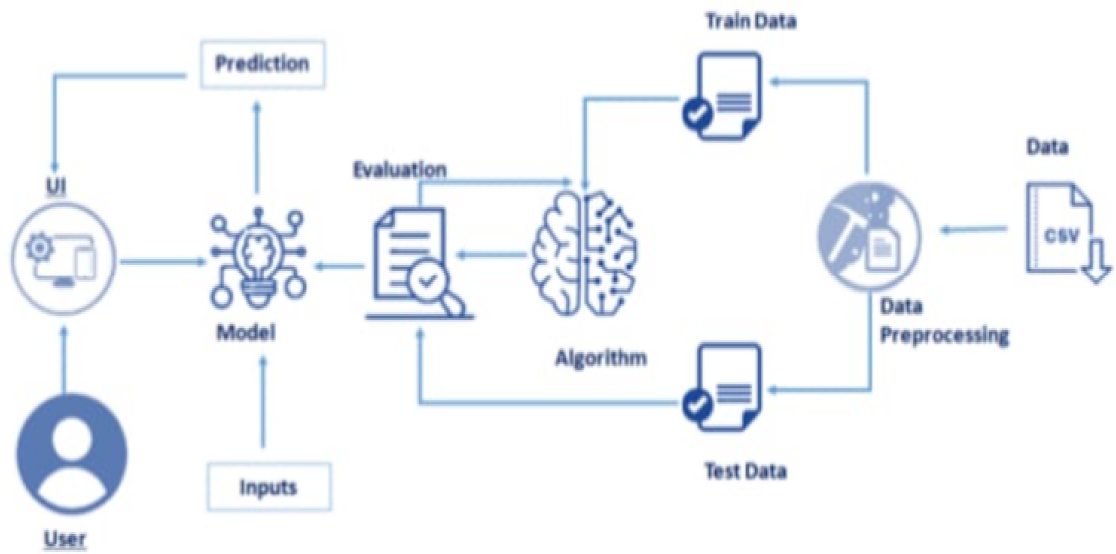
User Journey Map

PHASES	REQURIEMENTS NEED	SAMPLE COLLECTION	DATA ANALYSIS	INFORMATION UTILIZATION
STEPS	• selection of parameters • selection of methods • precision and accuracy	clean the sample containers and choose the filter pore size.minimize microbiale activities.select sample prevention method	measurements of six parameters and analyse the data collected.The unnecessary data will be rejected.Being analyse the data and interpret result.	Finally the data collected is tested and predic the good condition of the water.it will be detected by using the advanced artificial intelligence algorthims.
FEELINGS	☺	☹	😐	□
	• less unused features • less development rework • Some defects may occur	Highly specificity for target compounds.detection limits below regulatory trigger criteria.The resonable throughput for sample collection is more quantity id difficult.	Difficult to manage over time and with large data set.Require operation to submit it data,sometimes its configuration is required.	Usually feasible under exchange grands to a final result but it is challenging to accomplish the specific result to produce.
PAIN POINTS	• undocumented process • conflict requirement • need of more resources	Lack of technology and human resources occur sometimes.Storage and transportation issue happens.Technical hurdles is one of the pain point.	Collecting of water quality data can be expensive.Maintaining and repairing equippment costs can be rack up quickly overtime.sometime in correct may be an problem.	It still has a high require component.Good quality needed for all .To measure the required parameter of water.
OPPURTUNITIES	• lower cost of developement • Higher level of needs • More beneficial Members	Sampling reducestime and cost of research studies.The quality of water is always a better with sample collection.It provides much quicker result.	Appropriate data submission gives and excellent output.Then it is easy to verify the parameters and can predict the water quality.	The utilization of data in decision making allows us to make decisions based on evidence and also speedup the things by making it easier to share the perception.it also has the advantage of making it easier to verify the result in future.

6.2 DATA FLOW DIAGRAMS:



6.3 TECHNOLOGY ARCHITECTURE:



CHAPTER-7
PROJECT PLANNING PHASE

7.1 PREPARE MILESTONE AND ACTIVITY LIST:

S.NO	MILESTONE	DESCRIPTION	DURATION	WORKING STATUS
1	Prerequisites	Prerequisties are all the needs at the requirement level needed for the execution of the different phases	1 WEEK	Completed
2	Ideation	Ideation process is where you generate ideas and solutions through sessions such as sketching,prototyping,Brainstroming,Worst Possible,ideas,and Wealth of Other techniquies.	1 WEEK	Completed
3	Project design phase	Project design is an early phase of a project where the project's key features, structure, criteria for success, and major deliverables are planned out. The aim is to develop one or more designs that can be used to achieved the desire goals.	1 WEEK	Completed
4	Project Plannin gPhase	In the Planning Phase, the Project Manager works with the team to create the technical design, task list,resources,communicat	1 WEEK	Completed

		ionplan,budgetand initialschedule for project.		
5	Data Collection and Data preprocessing	A Data collection is a process of gathering and measuring information on variables to ensure accuracy and facilitate analysis. It help to solve the critical workloads.	1 WEEK	Completed
6	Model Building	Model Building is used for project visualization to provide information about the proposed state. It helps to identify the quality of objectives and it formulate the conceptual model.	4 WEEKS	Completed
7	Develop Application	A web application is application software that runs in a web browser, unlike software programs that run locally and natively on the operating system of the device.	4 WEEKS	Completed
8	Project develop ment phase	Project development is the process of planning and allocating resourcesto fully develop a project or product from concept to go-live.	4 WEEKS	Completed

7.2 SPRINT DELIVERY:

PRODUCT BACKLOG, SPRINT SCHEDULE AND ESTIMATION:

SPRINT	FUNCTIONAL REQUIREMENTS	USER STORY NUMBER	USER STORY/ TASK	STORY POINTS	PRIORITY	TEAM MEMBERS
Sprint 1	Regestration	USN 1	As a user, I can register for the application by entering my email, password, and confirming my password	2	HIGH	S.G.KEERTHANA
Sprint 2	User Confrimation	USN 2	As a user, I will receive confirmation email once I have registered for getting the data set.	1	HIGH	S.NARMADHA
Sprint 3	Login	USN 3	As a user, I can login to the application by entering my email and password.	1	HIGH	N.PRIYADHARSHINI
Sprint 4	Home Page	USN 4	As a user, I can find the data set to analyse waterquality.	2	HIGH	M.THARANI

PROJECTOR TRACKER , VELOCITY& BURN DOWN CHART

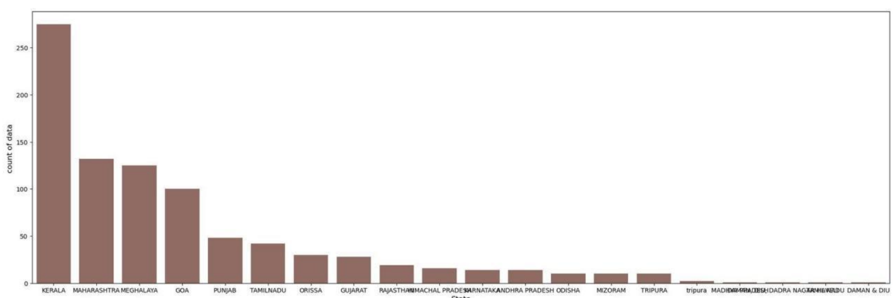
SPRINT	TOTAL STORY POINTS	DURATION	SPRINT START DATE	SPRINT END DATE (Plannned)	STORY POINTS COMPLETED (as on Planned End Date)	SPRINT RELEASE DATE(Acutal)
Sprint 1	20	4 Days	24 OCT 2022	27 OCT 2022	20	29 OCT 2022
Sprint 2	20	5 Days	28 OCT 2022	01 NOV 2022	20	04 NOV 2022
Sprint 3	20	8 Days	02 NOV 2022	09 NOV 2022	20	11 NOV 2022
Sprint 4	20	9 Days	10 NOV 2022	10 NOV 2022	20	19 NOV 2022

CHAPTER-8

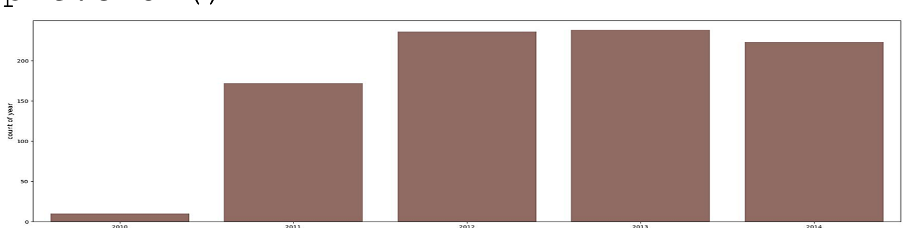
PROJECT DEVELOPEMENT PACKAGE

8.1 PROJECT DEVELOPEMENT-DELIVERY OF SPRINT -1

```
import pandas as pd
 import numpy as np
 import os
 import matplotlib.pyplot as plt
 import seaborn as sns
plt.figure(figsize=(25,8))
sns.barplot(int_level.index,int_level.values,alpha=0.9,color=color[5])
plt.ylabel('count of data ',fontsize=12)
plt.xlabel('State',fontsize=12)
plt.show()
```

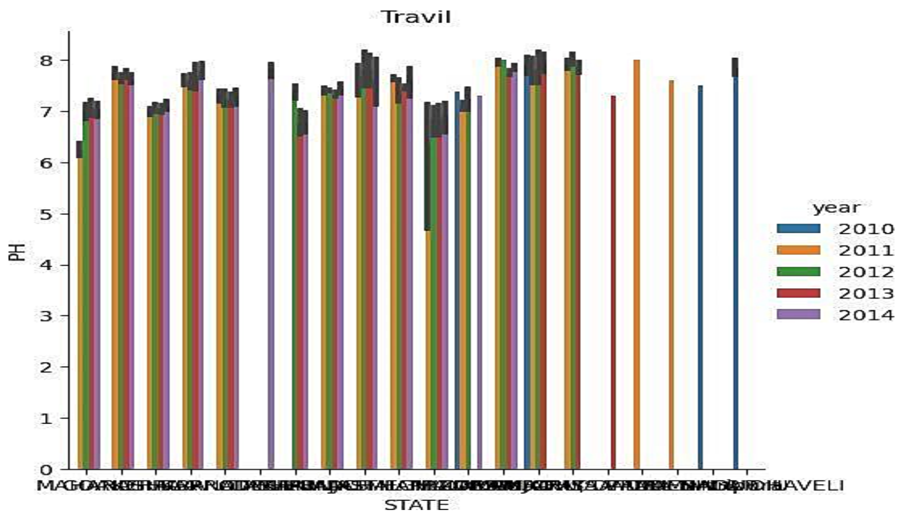


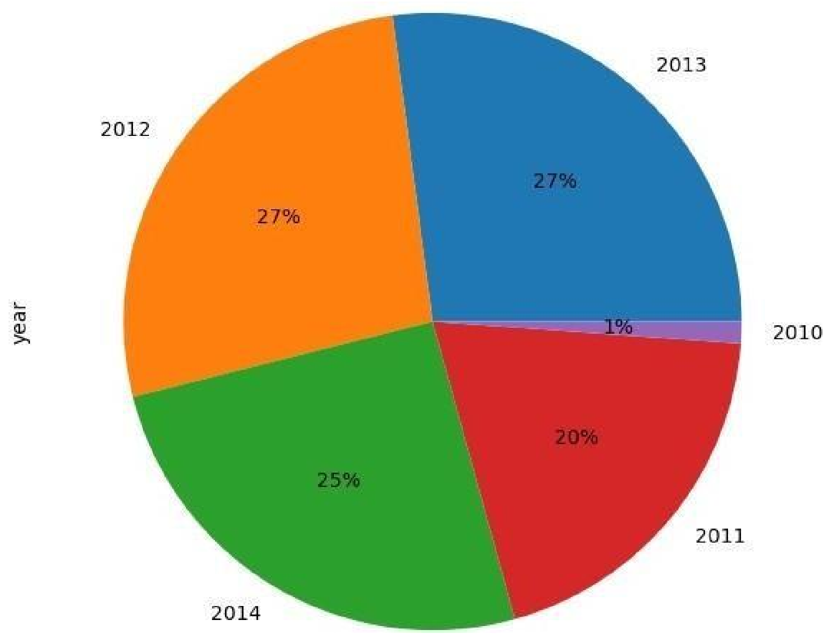

```
plt.figure(figsize=(25,8))
sns.barplot(int_level.index,int_level.values,alpha=0.9,
color=color[5])
plt.ylabel('count of year',fontsize=12)
plt.xlabel('year',fontsize=12)
plt.show()
```




```
plt.figure(figsize=(20,20))
```

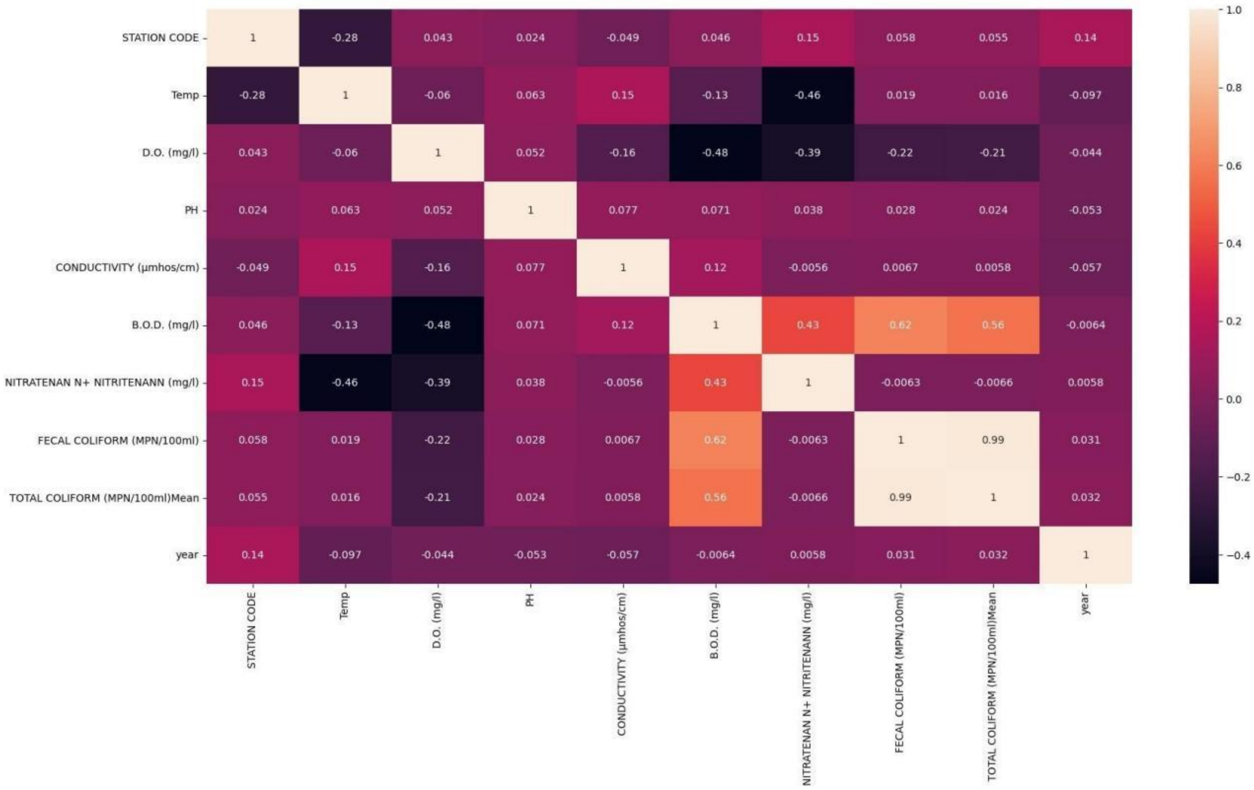
```
g=sns.catplot(data=df,kind="bar",x="STATE",y="PH",hue="
year")
plt.title("Travil")
Text(0.5, 1.0, 'Travil')
```



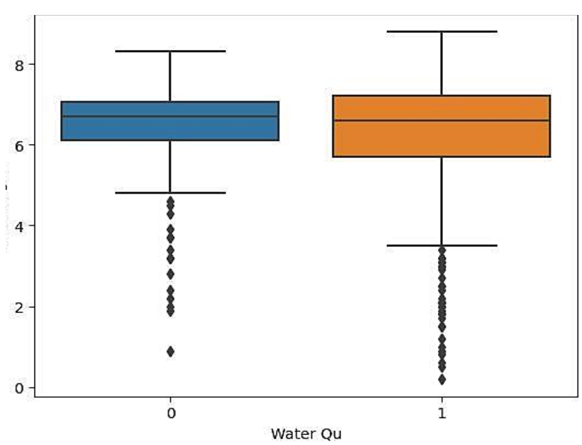


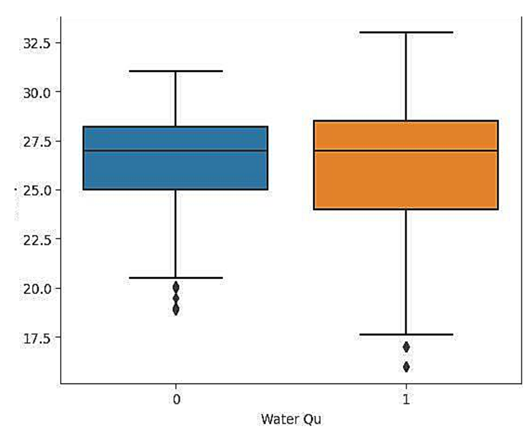
8.2 PROJECT DEVELOPEMENT-DELIVERY OF SPRINT -2

```
plt.figure(figsize=(20,10))
sns.heatmap(df.corr(),annot=True)
plt.show()
```




```
df.drop(df.index[(df[col]>col_Higher)],inplace=True,axis=0)
sns.boxplot(x='Water Qu',y=df[col],data=df)
plt.show()
```





LinearRegression

```
#fit the Linear regression model
regressor= LinearRegression()
regressor.fit(x_train, y_train)
y_pred= regressor.predict(x_test)
#x_pred= regressor.predict(x_train)
ypred_pd=pd.DataFrame({'WQ':y_test.values,'WQ_Pred':y_pred})
ypred_pd['predicted']=ypred_pd['WQ_Pred'].map(lambda x:1 if x>0.5 else0)
ypred_pd['WQ']=ypred_pd['WQ'].map(lambda x:1 if x>0.7 else 0)
ypred_pd.head()
WQ
WQ_Pred predicted
confusion=confusion_matrix(ypred_pd['WQ'],ypred_pd['predicted'])
print(confusion)
print(accuracy_score(ypred_pd['WQ'],ypred_pd['predicted']))
0.9344262295081968
```

Decision Tree

```
# Fit the desiontree regression
clf_gini = DecisionTreeRegressor(random_state = 0)
clf_gini.fit(x_train, y_train)
y_pred = clf_gini.predict(x_test)
ypred_pd=pd.DataFrame({'WQ':y_test.values,'WQ_Pred':y_pred})
ypred_pd['predicted']=ypred_pd['WQ_Pred'].map(lambda x:1 if x>0.7 else0)
ypred_pd['WQ']=ypred_pd['WQ'].map(lambda x:1 if x>0.7 else 0)
ypred_pd.head()
WQ
WQ_Pred predicted
print('Model accuracy score with criterion gini index: {0:0.4f}'.
format(accuracy_score(ypred_pd['WQ'],ypred_pd['predicted'])))
Model accuracy score with criterion gini index: 0.9180
```

8.3 PROJECT DEVELOPEMENT-DELIVERY OF SPRINT -3:

Initial Analysis

In [3]:
```
df.shape
```

Out[3]:
```
(3276, 10)
```

In [4]:
```
df.info()
```

```
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
```

Except Target feature, other features are float and continueous value. we can convert the Portability into Categoring feature

In [5]:
```
df.nunique()
```

Out[5]:
```
ph                 2785
Hardness           3276
Solids             3276
Chloramines        3276
Sulfate            2495
Conductivity       3276
Organic_carbon     3276
Trihalomethanes    3114
Turbidity          3276
Potability            2
dtype: int64
```

Statistical Analysis

In [7]:
```
df.describe().T.style.background_gradient(subset=['mean','std','50%','count'], cmap='PuBu')
```

Out[7]:

	count	mean	std	min	25%	50%	75%	max
ph	2785.000000	7.080795	1.594320	0.000000	6.093092	7.036752	8.062066	14.000000
Hardness	3276.000000	196.369496	32.879761	47.432000	176.850538	196.967627	216.667456	323.124000
Solids	3276.000000	22014.092526	8768.570828	320.942611	15666.690297	20927.833607	27332.762127	61227.196008
Chloramines	3276.000000	7.122277	1.583085	0.352000	6.127421	7.130299	8.114887	13.127000
Sulfate	2495.000000	333.775777	41.416840	129.000000	307.699498	333.073546	359.950170	481.030642
Conductivity	3276.000000	426.205111	80.824064	181.483754	365.734414	421.884968	481.792304	753.342620
Organic_carbon	3276.000000	14.284970	3.308162	2.200000	12.065801	14.218338	16.557652	28.300000
Trihalomethanes	3114.000000	66.396293	16.175008	0.738000	55.844536	66.622485	77.337473	124.000000
Turbidity	3276.000000	3.966786	0.780382	1.450000	3.439711	3.955028	4.500320	6.739000

From the above table, we can see that the count of each feature are not same. so there must me some null values. Feature Solids has the high mean and standard deviation comparted to other feature. so the distribution must be high. However, the above description is for overall population. lets try the same for 2 samples based on Portability feature

In [8]:
```
#Portability is 1 - means good for Human
df[df['Potability']==1].describe().T.style.background_gradient(subset=['mean','std','50%','count'], cmap='PuBu')
```

Out[8]:

	count	mean	std	min	25%	50%	75%	max
ph	1101.000000	7.073783	1.448048	0.227499	6.179312	7.036752	7.933068	13.175402
Hardness	1278.000000	195.800744	35.547041	47.432000	174.330531	196.632907	218.003420	323.124000
Solids	1278.000000	22383.991018	9101.010208	728.750830	15668.985035	21199.386614	27973.236446	56488.672413
Chloramines	1278.000000	7.169338	1.702988	0.352000	6.094134	7.215163	8.199261	13.127000
Sulfate	985.000000	332.566990	47.692818	129.000000	300.763772	331.838167	365.941346	481.030642
Conductivity	1278.000000	425.383800	82.048446	201.619737	360.939023	420.712729	484.155911	695.369528
Organic carbon	1278.000000	14.160893	3.263907	2.200000	12.033897	14.162809	16.356245	23.604298

In [9]:
```python
# Portability is 0 - means not good for Human
df[df['Potability']==0].describe().T.style.background_gradient(subset=['mean','std','50%','count'], cmap='RdBu')
```

Out[9]:

	count	mean	std	min	25%	50%	75%	max
ph	1684.000000	7.085378	1.683499	0.000000	6.037723	7.035456	8.155510	14.000000
Hardness	1998.000000	196.733292	31.057540	98.452931	177.823265	197.123423	216.120687	304.235912
Solids	1998.000000	21777.490788	8543.068788	320.942611	15663.057382	20809.618280	27006.249009	61227.196008
Chloramines	1998.000000	7.092175	1.501045	1.683993	6.155640	7.090334	8.066462	12.653362
Sulfate	1510.000000	334.564290	36.745549	203.444521	311.264006	333.389426	356.853897	460.107069
Conductivity	1998.000000	426.730454	80.047317	181.483754	368.498530	422.229331	480.677198	753.342620
Organic_carbon	1998.000000	14.364335	3.334554	4.371899	12.101057	14.293508	16.649485	28.300000
Trihalomethanes	1891.000000	66.303555	16.079320	0.738000	55.706530	66.542198	77.277704	120.030077
Turbidity	1998.000000	3.965800	0.780282	1.450000	3.444062	3.948076	4.496106	6.739000

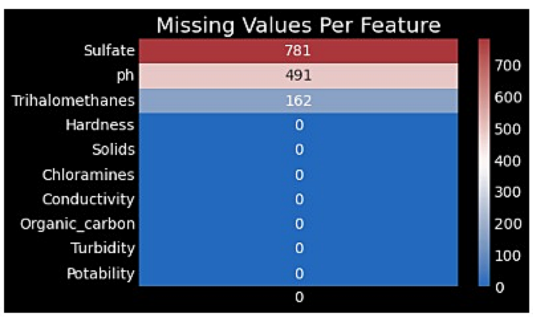
Mean and std of almost all features are similar for both samples. there are few differnces in Solids feature. Further analysis using hypothetical testing could help us to identify the significance.

8.4 PROJECT DEVELOPEMENT-DELIVERY OF SPRINT -4:

Check for missing values

In [10]:
```python
plt.title('Missing Values Per Feature')
nans = df.isna().sum().sort_values(ascending=False).to_frame()
sns.heatmap(nans,annot=True,fmt='d',cmap='vlag')
```

Out[10]:



In [11]:
```python
df[df['Sulfate'].isnull()]
```

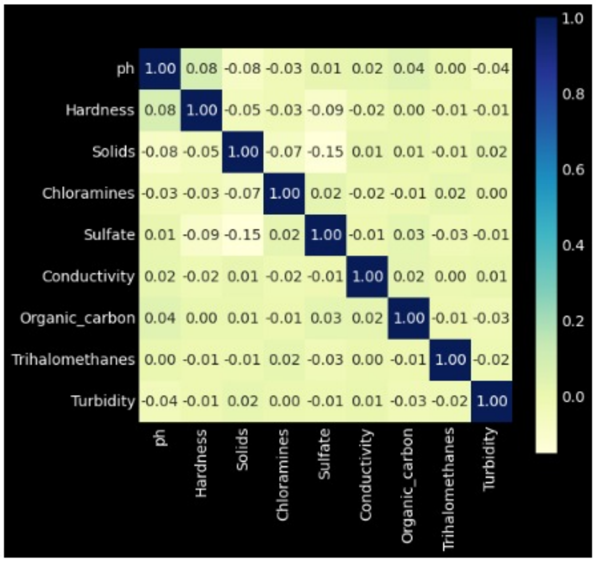
Out[11]:

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
11	7.974522	218.693300	18767.656682	8.110385	NaN	364.098230	14.525746	76.485911	4.011718	0
14	7.496232	205.344982	28388.004887	5.072558	NaN	444.645352	13.228311	70.300213	4.777382	0
16	7.051786	211.049406	30980.600787	10.094796	NaN	315.141267	20.397022	56.651604	4.268429	0
...	...	...	...	...	...	...	...	...	...	...
3266	8.372910	169.087052	14622.745494	7.547984	NaN	464.525552	11.083027	38.435151	4.906358	1
3272	7.808856	193.553212	17329.802160	8.061362	NaN	392.449580	19.903225	NaN	2.798243	1

Exploratory Data Analysis

In [16]:

```
Corrmat = df.corr()
plt.subplots(figsize=(7,7))
sns.heatmap(Corrmat, cmap="YlGnBu", square = True, annot=True, fmt='.2f')
plt.show()
```



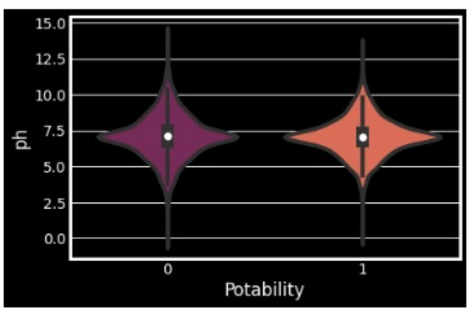
In [17]:

```
fig =  ex.pie (df, names = "Potability", hole = 0.4, template = "plotly_dark")
fig.show ()
```

In [18]:

```
sns.violinplot(x='Potability', y='ph', data=df, palette='rocket')
```

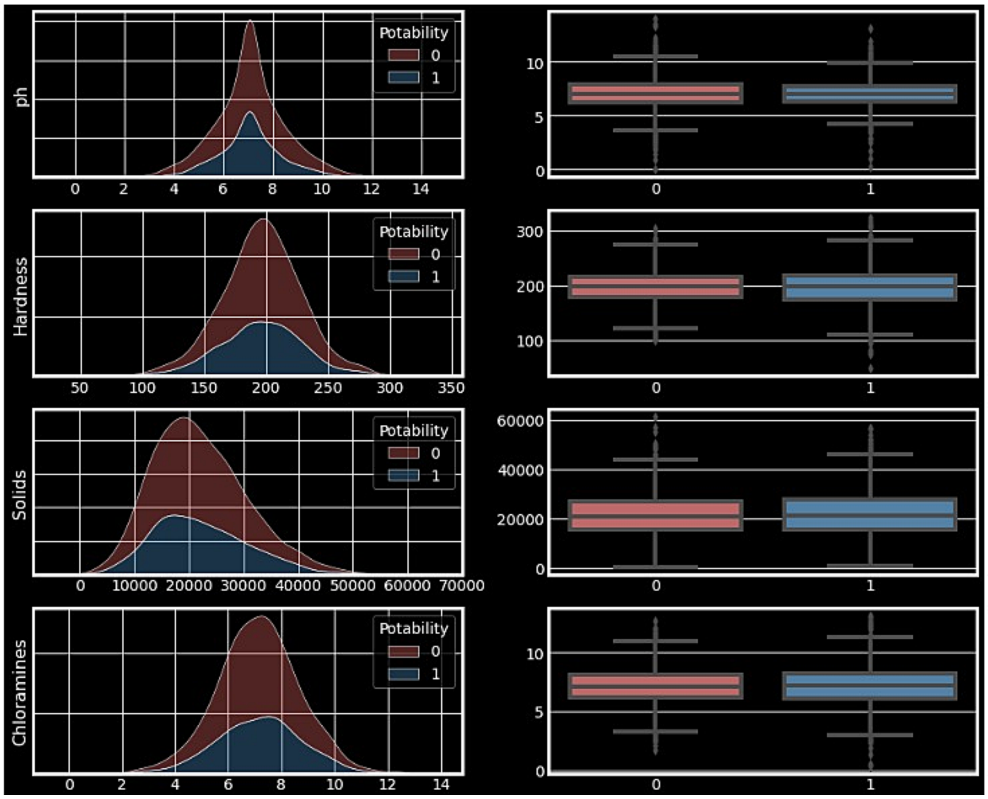
Out[18]:



In [19]:

```
print('Boxplot and density distribution of different features by Potability\n')

fig, ax = plt.subplots(ncols=2, nrows=9, figsize=(14, 28))

features = list(df.columns.drop('Potability'))
i=0
for cols in features:
    sns.kdeplot(df[cols], fill=True, alpha=0.4, hue = df.Potability,
                palette=('indianred', 'steelblue'), multiple='stack', ax=ax[i,0])

    sns.boxplot(data= df, y=cols, x='Potability', ax=ax[i, 1],
               palette=('indianred', 'steelblue'))
    ax[i,0].set_xlabel(' ')
    ax[i,1].set_xlabel(' ')
    ax[i,1].set_ylabel(' ')
    ax[i,1].xaxis.set_tick_params(labelsize=14)
    ax[i,0].tick_params(left=False, labelleft=False)
    ax[i,0].set_ylabel(cols, fontsize=16)
    i=i+1

plt.show()
```

Boxplot and density distribution of different features by Potability



CONCLUSION

This paper analyzes and forecasts the values of water quality parameters, in order to determine the concentration of Chlorophyll, Dissolved Oxygen, Turbidity and Specific Conductance and analyzes the results. The time series data used has been acquired from USGS National Water Information System (NWIS), with data from the year of 2014. The specified monitoring station is a channel situated in the State of New York. The measurements of water quality parameters were scaled between 0 and 1 for better data handling. Artificial Neural Network (ANN) with Nonlinear Autoregressive (NAR) time series has been used with Scaled Conjugate gradient (SCG) as training algorithm. Four ANN models depicting the four selected water quality parameters have been developed and analyzed. The performance measures that are used to depict the result are Regression, Mean Squared Error and Root Mean Squared Error . The results of the conducted tests provide an insight about the prediction efficiency and accuracy of the proposed model with the help of performance measures. The proposed model comprising of ANN-NAR proves to a reliable one with the prediction accuracy indicating much improved values, with the lowest MSE being 3.7x10-4 for turbidity and the best Regression value for Specific Conductance (0.99). The future of water quality modeling seems to be very bright and remarkable with the continuous improvement in technology day by day. Besides further improvements in prediction accuracy, there needs to be a more user-centric approach towards tackling the water quality issues, by involving all the relevant stakeholders, using user-friendly tools and an interactive environment so that the solution actually benefits the target users in tackling water quality issuesIt will hopefully result in curtailment of people consuming poor quality water and consequently de-escalate harrowing diseases like typhoid and diarrhea. In this regard, the application of a prescriptive analysis from the expected values would lead to future facilities to support decision and policy makers.

REFERENCE

[1]HAN Silva, A Rosato, R Altilio, and M Panella, Water quality prediction based on wavelet neural networks and remote sensing, IEEE, in 2018 International Joint Conference on Neural Networks (IJCNN) (Rio de Janeiro, Brazil), 2018, pp. 1-6.

[2]S Chatterjee, S Sarkar, N Dey, S Sen, T Goto, and NC Debnath, Water quality prediction: multi objective genetic algorithm coupled artificial neural network based approach, IEEE, in 2017 IEEE 15th International Conference on Industrial Informatics (INDIN) (Emden, Germany), 2017, pp. 963-968.

[3]WC Leong, A Bahadori, J Zhang, and Z Ahmad, Prediction of water quality index (WQI) using support vector machine (SVM) and least square-support vector machine (LS-SVM), International Journal of River Basin Management, Vol. 19, 2021, pp. 149-156.

[4]TH Aldhyani, M Al-Yaari, H Alkahtani, and M Maashi, Water quality prediction using artificial intelligence algorithms, Applied Bionics and Biomechanics, Vol. 2020, 2020, pp. 6659314.

[5]P Liu, J Wang, AK Sangaiah, Y Xie, and X Yin, Analysis and prediction of water quality using LSTM deep neural networks in IoT environment, Sustainability, Vol. 11, 2019, pp. 2058.