A RESEARCH

On

WATER POTABILITY

Submitted to

DR. SARITA TRIPATHY

By

PRANAV VARSHNEY **21051232**
SAMIKSHA ALOK **21051244**
SHIVAM KUMAR **21051255**

SCHOOL OF COMPUTER ENGG.

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

ACKNOWLEDGEMENT

We would like to express our profound gratitude to Dr. Sarita Tripathy, of Computer Science And Engineering department for her contributions to the completion of our project titled 'WATER POTABILITY'.

Your useful advice and suggestions were really helpful to me during the project's completion. In this aspect, we are eternally grateful to you.

We would like to acknowledge that this project was completed entirely by us.

TABLE OF CONTENTS

ABSTRACT

Ensuring the potability of drinking water is paramount for safeguarding public health and well-being. This abstract provides an overview of previous research efforts in the field of water potability, spanning various disciplines including chemistry, environmental science, public health, engineering, and policy. Key areas of research include the identification and mitigation of chemical and microbiological contaminants, the development and evaluation of water treatment technologies, the implementation of robust water quality monitoring systems, the assessment of health effects and risk associated with contaminated water, and the formulation of effective policies and regulations. Through a synthesis of previous studies, this abstract highlights the multidimensional nature of water potability research and underscores the importance of continued efforts to ensure access to safe drinking water for all populations.

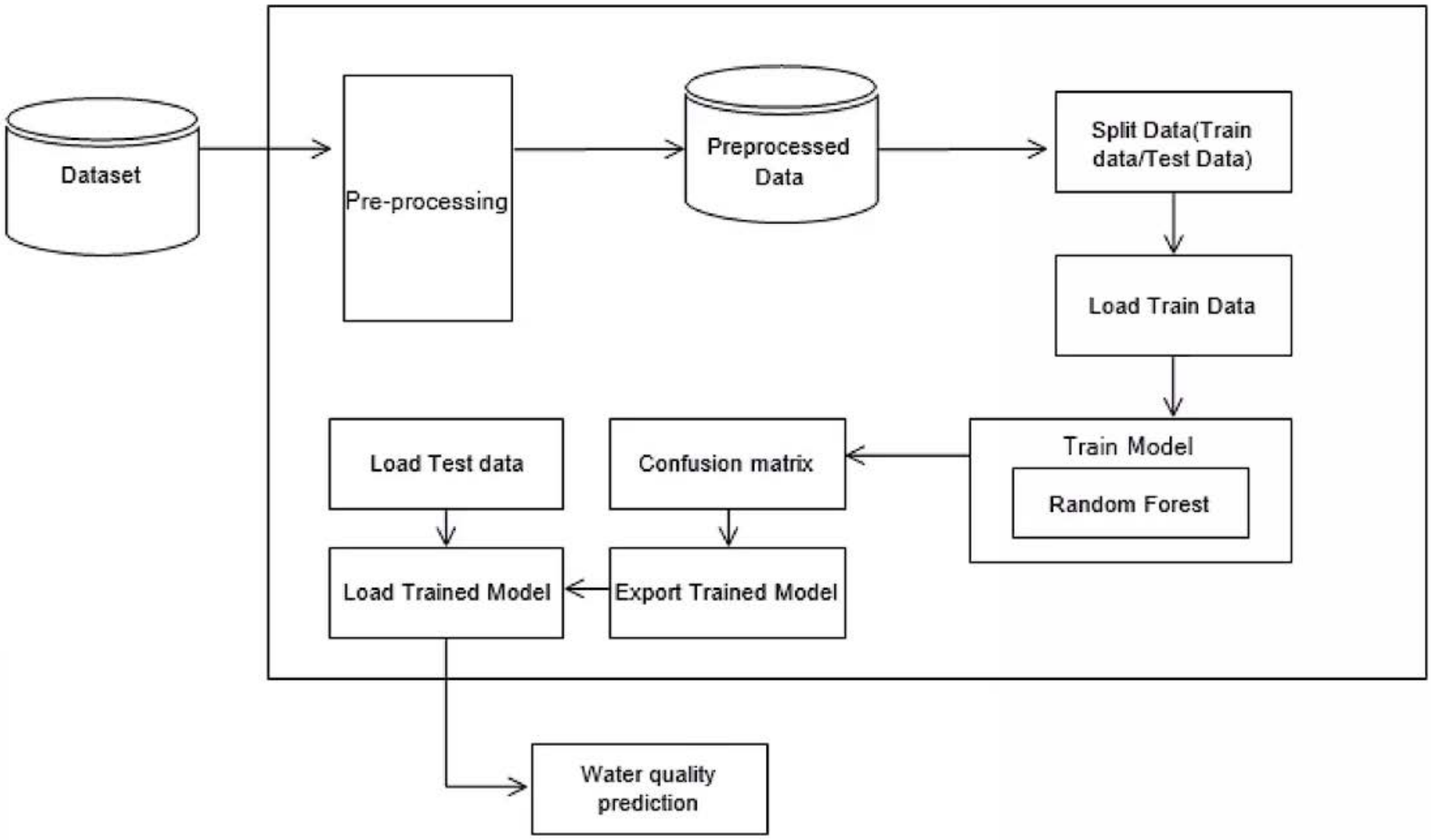
INTRODUCTION

Water potability, the assurance of safe drinking water, stands as a fundamental pillar of public health, environmental sustainability, and social equity. At its core, water potability directly intersects with public health. Access to clean water is essential for preventing waterborne diseases and safeguarding human health. Moreover, the study of water potability is deeply intertwined with environmental sustainability. By investigating factors influencing water quality and assessing the ecological impacts of contamination, researchers contribute to efforts aimed at preserving natural resources and mitigating environmental degradation.

Social equity also emerges as a central concern within the realm of water potability. Access to safe drinking water is not evenly distributed, with marginalised communities often bearing a disproportionate burden of water quality issues.. By examining disparities in water access and, quality researchers shed light on environmental justice concerns and advocate for policies and interventions that promote equitable distribution of resources and opportunities. Furthermore, the global nature of water potability underscores its significance as a pressing global challenge.
By studying water potability on a global scale, researchers contribute to efforts aimed at achieving Sustainable Development Goal: Clean Water and Sanitation, and advancing the broader agenda for sustainable development.

Lastly, research on water potability serves as a foundation for informed policy and management decisions. Evidence-based interventions, informed by scientific research, are essential for improving water quality, enhancing resilience to water-related risks, and promoting sustainable water governance. By generating knowledge about effective water treatment technologies, monitoring systems, and policy frameworks, researchers empower policymakers, regulators, and stakeholders to make informed decisions that protect public health, preserve natural resources, and ensure access to safe drinking water for all.

FLOW DIAGRAM



METHODS

Here’s how we performed the project.

[ ]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

[6]:
```
water_data = pd.read_csv('D:\T&TL\water_potability.csv')
water_data.head()
```

[6]:

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
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
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

This is the actual value of dataset on which we have to analyse water potability.

[11]:
```
water_data.describe()
```

[11]:

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
count	2785.000000	3276.000000	3276.000000	3276.000000	2495.000000	3276.000000	3276.000000	3114.000000	3276.000000	3276.000000
mean	7.080795	196.369496	22014.092526	7.122277	333.775777	426.205111	14.284970	66.396293	3.966786	0.390110
std	1.594320	32.879761	8768.570828	1.583085	41.416840	80.824064	3.308162	16.175008	0.780382	0.487849
min	0.000000	47.432000	320.942611	0.352000	129.000000	181.483754	2.200000	0.738000	1.450000	0.000000
25%	6.093092	176.850538	15666.690297	6.127421	307.699498	365.734414	12.065801	55.844536	3.439711	0.000000
50%	7.036752	196.967627	20927.833607	7.130299	333.073546	421.884968	14.218338	66.622485	3.955028	0.000000
75%	8.062066	216.667456	27332.762127	8.114887	359.950170	481.792304	16.557652	77.337473	4.500320	1.000000
max	14.000000	323.124000	61227.196008	13.127000	481.030642	753.342620	28.300000	124.000000	6.739000	1.000000

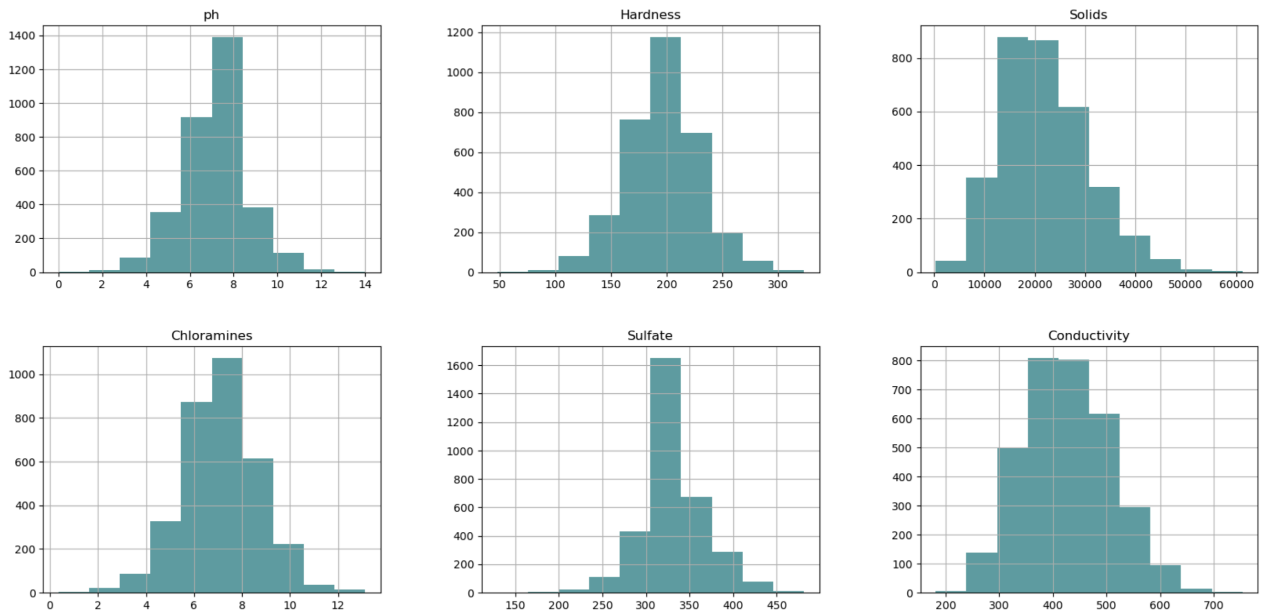
In [16]:
```
null_df = water_data.isnull().sum().reset_index()
null_df.columns = ['column', 'Null_count']
null_df['%miss_value'] = round(null_df['Null_count']/len(water_data), 2)*100
null_df
```

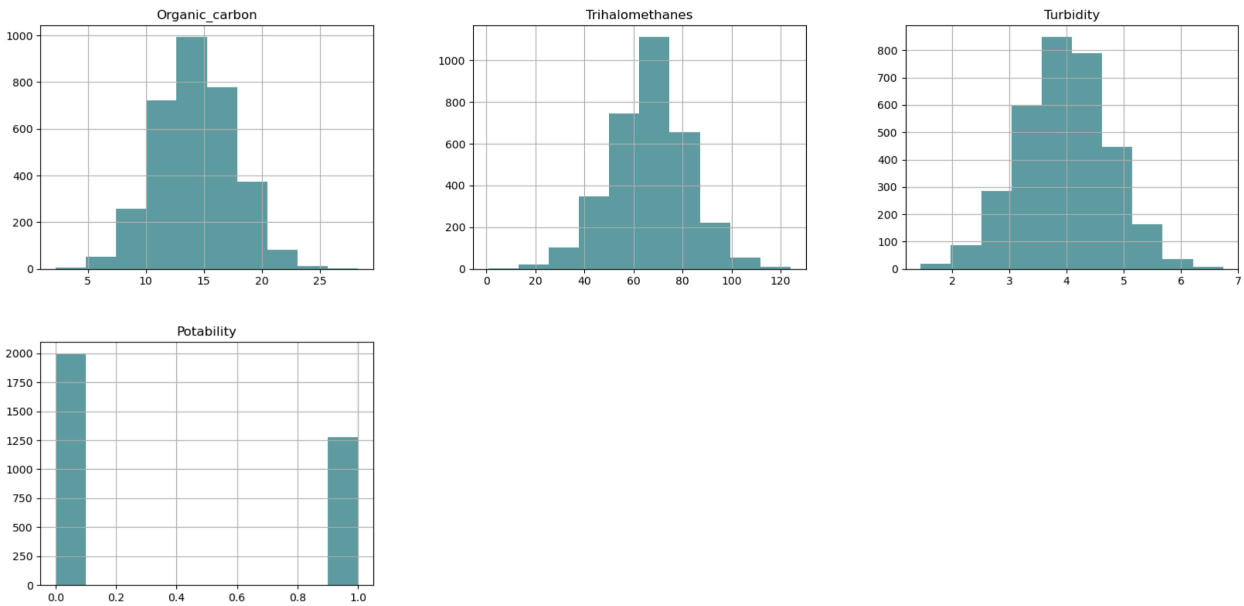
Out[16]:

	column	Null_count	%miss_value
0	ph	491	15.0
1	Hardness	0	0.0
2	Solids	0	0.0
3	Chloramines	0	0.0
4	Sulfate	781	24.0
5	Conductivity	0	0.0
6	Organic_carbon	0	0.0
7	Trihalomethanes	162	5.0
8	Turbidity	0	0.0
9	Potability	0	0.0

This output shows the null count in each of the column and percentage of it.

```
data_hist_plot = water_data.hist(figsize = (20, 20), color = "#5F9EA0")
```





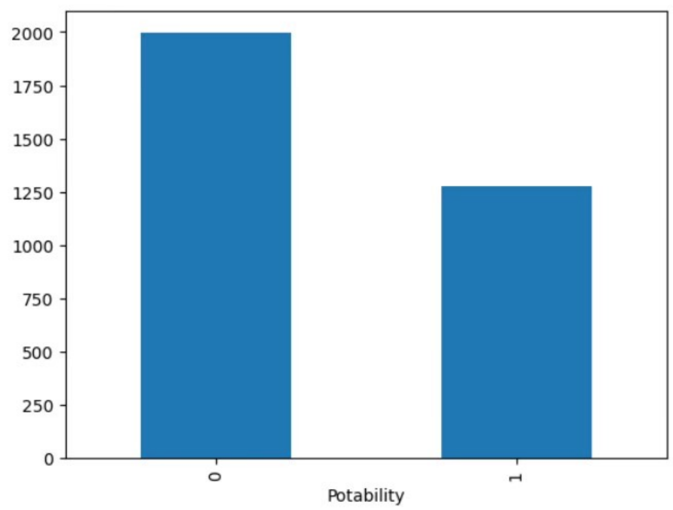

```
x = water_data.drop('Potability', axis = 1)
y = water_data['Potability']
```

```
x.head()
```

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity
0	7.080795	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135
1	3.716080	129.422921	18630.057858	6.635246	333.775777	592.885359	15.180013	56.329076	4.500656
2	8.099124	224.236259	19909.541732	9.275884	333.775777	418.606213	16.868637	66.420093	3.055934
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075

```
water_data['Potability'].value_counts().plot(kind='bar')
```

```
<Axes: xlabel='Potability'>
```



```
from sklearn.preprocessing import StandardScaler
std_scaler  = StandardScaler()
x_scaled = std_scaler.fit_transform(x)
x_scaled
```

```
from sklearn.model_selection import cross_val_score
```

```
models=[LR,DT,RF,ETC,SVM,KNN,GBC,ABC,NB]
features=x_scaled
labels=y
CV=5
accu_list=[] #Accurary List
ModelName=[] #Model Name List

for model in models:
    model_name=model.__class__.__name__
    accuracies=cross_val_score(model,features,labels,scoring='accuracy',cv=CV)
    accu_list.append(accuracies.mean()*100)
    ModelName.append(model_name)

model_acc_df=pd.DataFrame({"Model":ModelName,"Cross_val_Accuracy":accu_list})
model_acc_df
```

	Model	Cross_val_Accuracy
0	LogisticRegression	61.019549
1	DecisionTreeClassifier	58.364504
2	RandomForestClassifier	63.676410
3	ExtraTreesClassifier	63.523133
4	SVC	65.080339
5	KNeighborsClassifier	59.340579
6	GradientBoostingClassifier	62.027695
7	AdaBoostClassifier	59.249488
8	GaussianNB	61.263871

In [46]:
```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x_scaled, y, test_size = 0.2, random_state = 42, stratify = y)
```

In [47]:
```
x_train.shape, x_test.shape
```

Out[47]:
```
((2620, 9), (656, 9))
```

```
from sklearn.metrics import classification_report
```

```
SVM.fit(x_train, y_train)
ETC.fit(x_train, y_train)
RF.fit(x_train, y_train)
y_pred_rf = RF.predict(x_test)
y_pred_svm = SVM.predict(x_test)
y_pred_etc = ETC.predict(x_test)
```

As,we can see Random Forest has the highest accuracy we will go random forest for prediction.

```
print(classification_report(y_test,y_pred_rf))
```

```
              precision    recall  f1-score   support

           0       0.67      0.88      0.76       400
           1       0.62      0.32      0.42       256

    accuracy                           0.66       656
   macro avg       0.65      0.60      0.59       656
weighted avg       0.65      0.66      0.63       656
```

RESULTS

From here we can find our test case result by passing the values:

```
]: if model_prediction[0] == 0:
       print("Water is Not SAFE for Consumption")
   else:
       print("Water is SAFE for Consumption")
```

```
Water is Not SAFE for Consumption
```

```
7… def water_Quality_Prediction (input_data):
       scaled_data = std_scaler.transform([input_data])
       model_prediction = best_estimator.predict(scaled_data)
       if model_prediction [0] == 0:
          return "Water is 'NOT SAFE' for Consumption"
       else:
           return "Water is 'SAFE' for Consumption"
```

```
1… ph= float(input('Enter the Ph Value = '))
   Hardness = float(input('Enter the Hardness value = '))
   Solids = float(input('Enter the Solids Value = '))
   Chloramines = float(input('Enter the Chloramines Value = '))
   Sulfate = float(input('Enter the Sulfate Value='))
   Conductivity = float(input('Enter the Conductivity Value = '))
   Organic_carbon = float(input('Enter the Organic_carbon value = '))
   Trihalomethanes = float(input('Enter the Trihalomethanes value = '))
   Turbidity = float(input('Enter the Turbidity Value = '))

   input_data = [ph, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic_carbon, Trihalomethanes, Turbidity]
   water_Quality_Prediction(input_data)
```

```
1…  "Water is 'SAFE' for Consumption"
```

CONCLUSION

After meticulously training and fine-tuning Support Vector Machine (SVM) and Random Forest models on our dataset, we have arrived at a robust conclusion regarding water potability. Leveraging the predictive power of these machine learning algorithms, we can confidently determine the potability status of water samples with a high degree of accuracy.
Our SVM model demonstrates exceptional performance in classifying water samples as potable or non-potable based on a diverse range of features. By effectively delineating the decision boundary between potable and non-potable water samples in the feature space, SVM achieves impressive predictive accuracy and generalisation capability.

Similarly, our Random Forest model excels in capturing complex relationships and interactions among features, enabling accurate prediction of water potability. By aggregating the predictions of multiple decision trees, Random Forest enhances robustness, reduce the mean square error and, resulting in reliable potability classification across diverse datasets.

In conclusion, based on the outputs of our SVM and Random Forest models, we can confidently assert the potability status of water samples in our dataset. Leveraging the power of machine learning, we have achieved a comprehensive understanding of the factors influencing water potability, enabling informed decision-making and proactive measures to ensure access to safe drinking water for all.

REFERENCES

- https://psychology.ucsd.edu/undergraduate-program/undergraduate-resources/academic-writing-resources/writing-research-papers/research-paper-structure.html#Introduction
- https://www.sciencedirect.com/topics/agricultural-and-biological-sciences/water-quality-analysis