Data Analytics Internship Program 2024

Final Project Presentation

Developing a Sustainable Water Management System [SDG: 06]

Unique ID: IBM3465
Team DaSci ’24
Sri Sri University

CSRBOX®
Doing Good in a **Better** Way

Team Members:

- Lokesh Patra [*Team S’PoC*]
- Indigibilli Harshit
- Subhangee Das
- Mayank Sharad Kapse
- Subham Nayak

Introduction

OverView Of the Project

Urban areas face significant challenges related to water management, including inefficient distribution, water wastage, and inadequate access to clean water. These issues are exacerbated by population growth, climate change, and aging infrastructure.

Objectives

- Optimize Water Distribution
- Reduce Water Wastage
- Ensure Sustainable Access to Clean Water
- Enhance Predictive Capabilities
- Improve Infrastructure Management
- Increase Public Awareness and Engagement
- Leverage Technology for Real-time Monitoring
- Facilitate Data-Driven Decision Making
- Support Climate Resilience
- Promote Innovation in Water Management

Problem Identification

Problem Statement

The project aims to design a smart water management system that leverages data from various sources to optimize water distribution, reduce water waste, and ensure sustainable access to clean water in urban areas. A key objective of this project is to identify the most effective machine learning algorithm for accurately predicting water potability based on various features. By providing insights into the factors that influence water quality, this system will help ensure safe and reliable drinking water for urban populations, thereby preventing water-borne diseases and promoting public health.

Timelines [*SDG Relevance*]

- **1977 Mar del Plata Conference**: Established an Action Plan on "*Community Water Supply*" and declared the right to access drinking water.
- **1993 World Water Day & 2013 World Toilet Day**: Designated by the UN General Assembly to raise awareness.
- **2000 Millennium Development Declaration**: Aimed to halve the proportion of people without access to safe drinking water and basic sanitation by 2015.
- **2003 UN-Water Establishment**: Created to coordinate efforts on water and sanitation issues among UN entities and international organizations.
- **2016 International Decade for Action**: Adopted to support SDG 6 and other water-related targets, with a focus on sustainable development from 2018 to 2028.

6 CLEAN WATER AND SANITATION

ENSURE AVAILABILITY AND SUSTAINABLE MANAGEMENT OF WATER AND SANITATION FOR ALL

SAFE DRINKING WATER, SANITATION AND HYGIENE

STILL OUT OF REACH

FOR BILLIONS

IN 2022

2.2 BILLION PEOPLE LACKED SAFELY MANAGED DRINKING WATER

3.5 BILLION PEOPLE LACKED SAFELY MANAGED SANITATION

2.2 BILLION PEOPLE LACKED BASIC HAND WASHING FACILITIES

TO MEET 2030 TARGETS, PACE OF PROGRESS WILL HAVE TO ACCELERATE

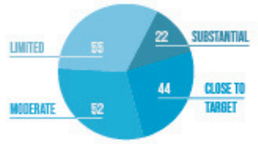
- 6x DRINKING WATER
- 5x FOR SANITATION
- 3x HYGIENE

2.4 BILLION PEOPLE LIVE IN WATER-STRESSED COUNTRIES (2020)

81% OF SPECIES DEPENDENT ON INLAND WETLANDS HAVE DECLINED SINCE 1970

INTEGRATED WATER-RESOURCES-MANAGEMENT IMPLEMENTATION NEEDS ACCELERATION

NUMBER OF COUNTRIES PER PROGRESS LEVEL



THE SUSTAINABLE DEVELOPMENT GOALS REPORT 2023: SPECIAL EDITION- UNSTATS.UN.ORG/SDGS/REPORT/2023/

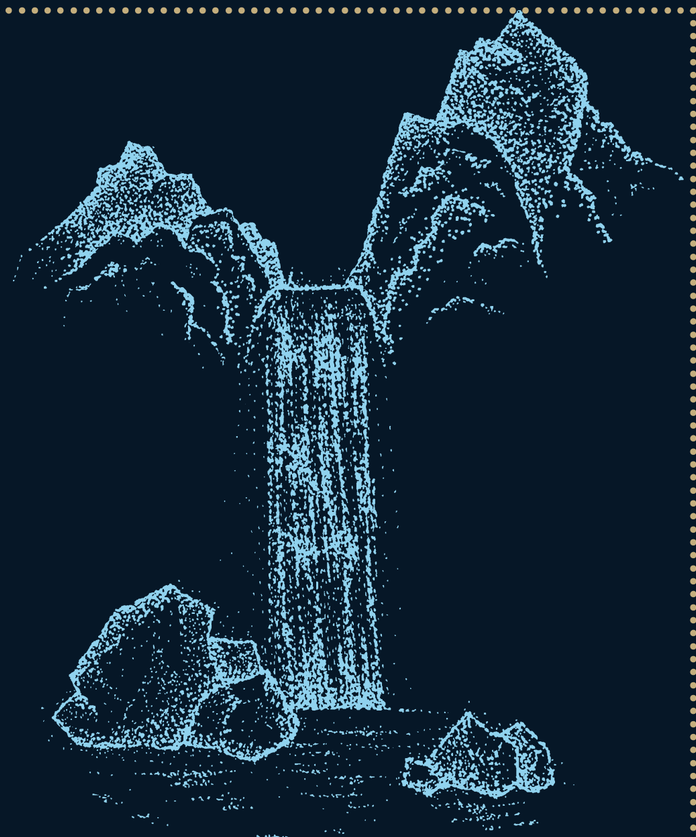
Data Collection

Data Sources /Collection /Techbonies

Water Flow and Quality Sensors
Weather Data
Population Growth and Urbanization Trends
Water Demand Patterns

Machine Learning and Predictive Analytics
Smart Meter Technology
Automated Control Systems

Concept Note ~



Methodologies

Data Analysis /Transfomartion

Collection /Analysis /Optimisation
User Engagement
Infrastructure Enhancement

Concept Note ~

Data Source

***Data is taken from Kaggle. There are several attributes in the data set on the basis of which we are predicting, analyzing, visualizing, and evaluation.*

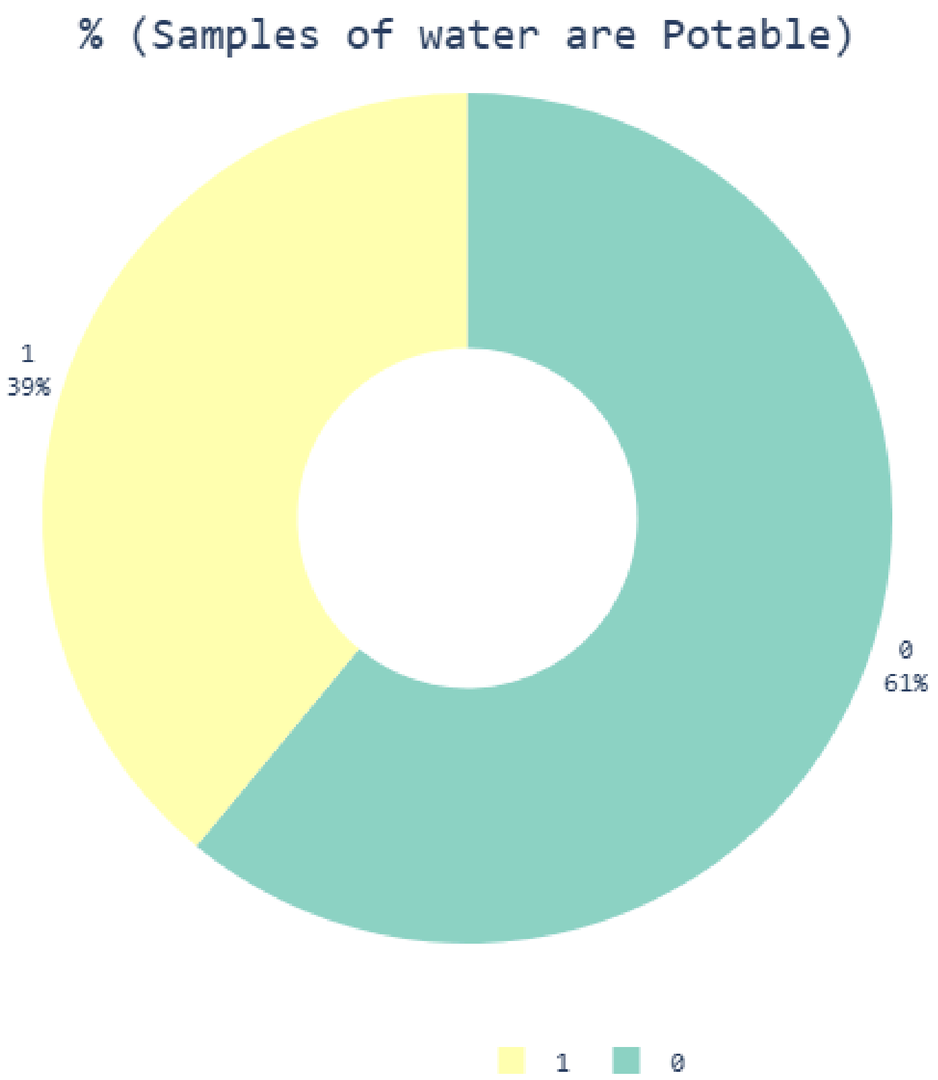
	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	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

After Cleaning

***Used .fillna() method in Pandas to fill missing values (NaN) in the Water DataFrame.*

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	7.036752	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	333.073546	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	333.073546	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

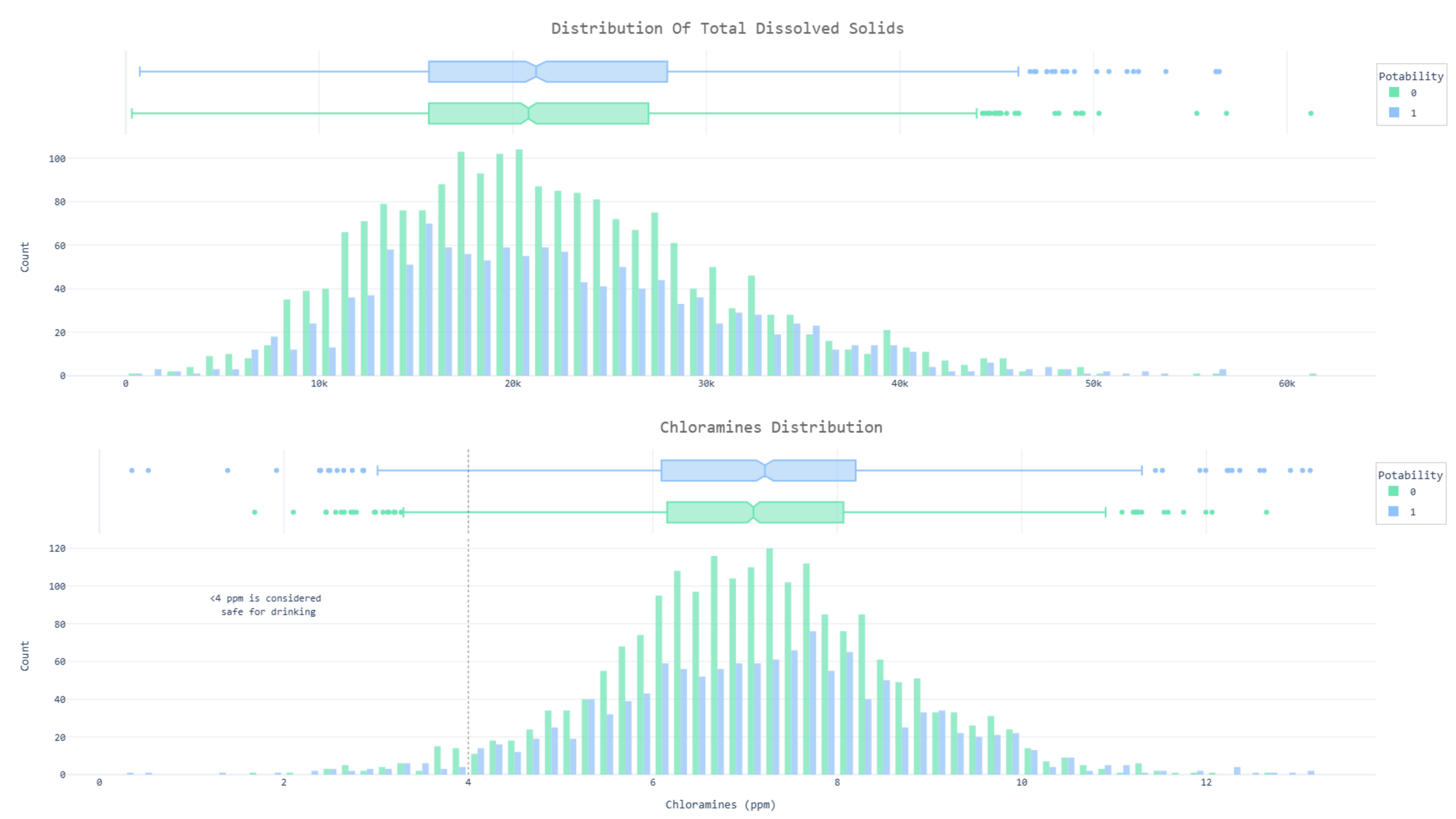
Amount Of Potable WaterBodies 💧

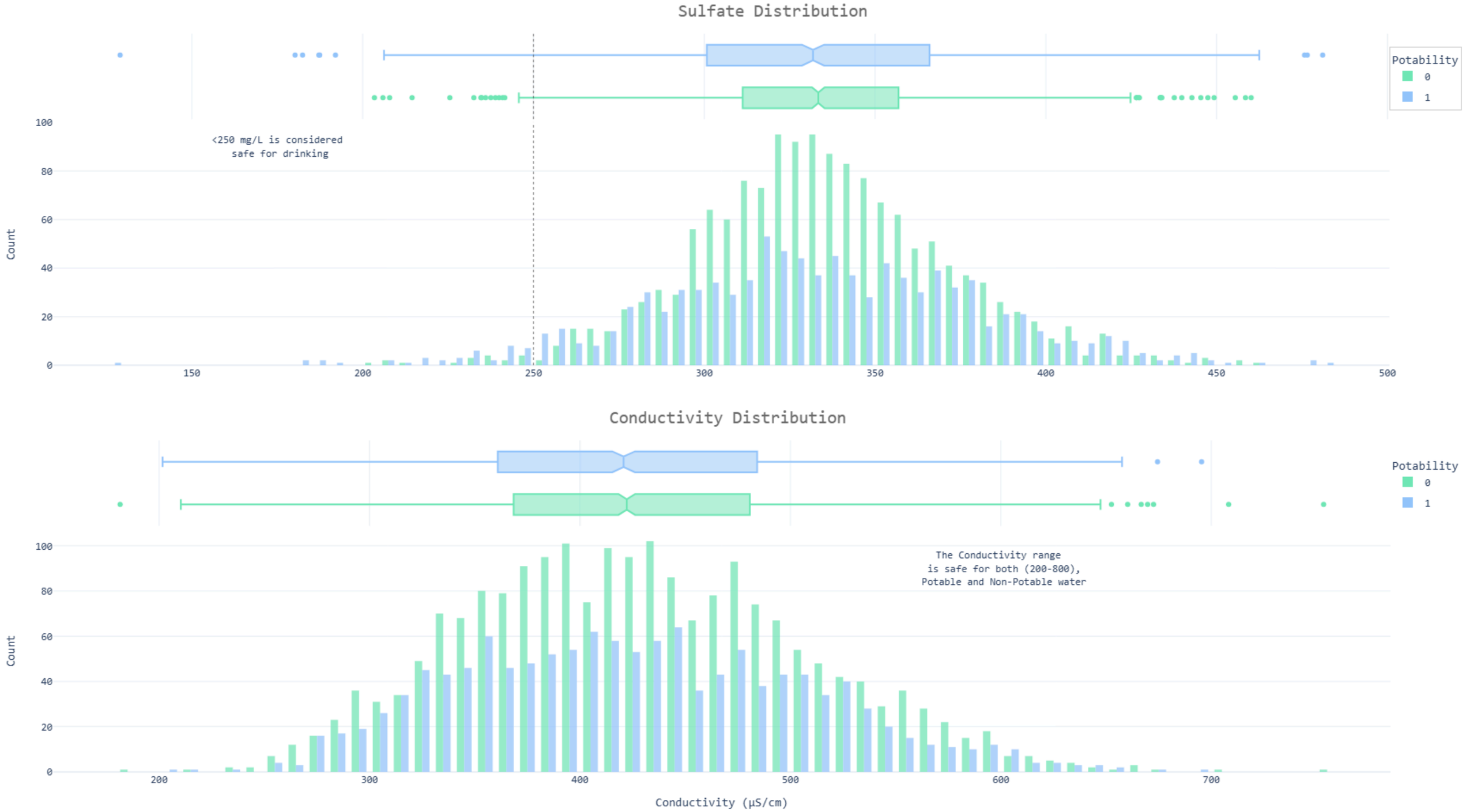


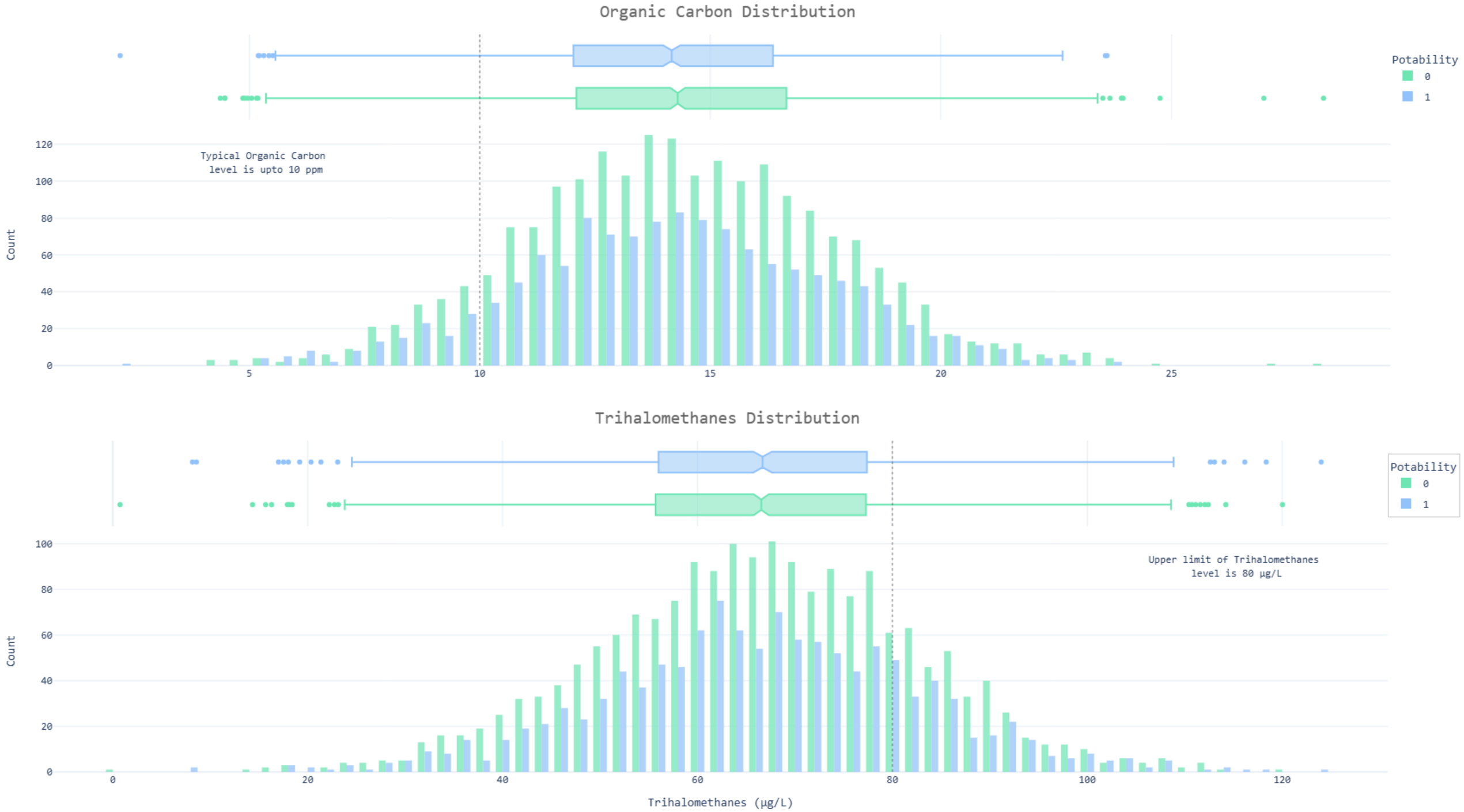
39% Drinkable
61% UnDrinkable

1278 Drinkable
1998 UnDrinkable

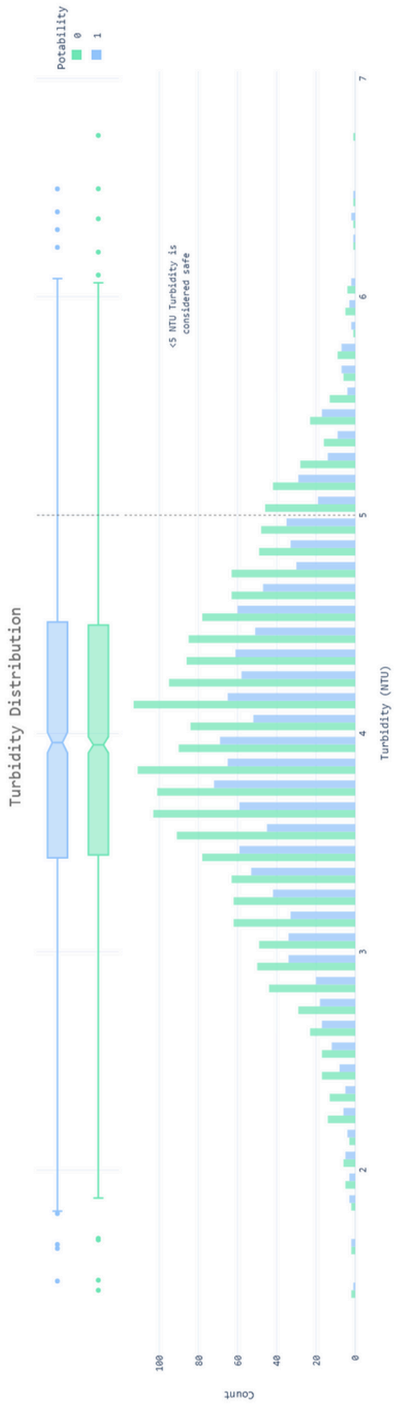

Hardness Distribution
Potability
0
1
<76 mg/L is considered soft
Between 76 and 150 (mg/L) is moderately hard
Between 151 and 300 (mg/L) is considered hard
>300 mg/L is considered very hard
Count
0
20
40
60
80
100
120
140
50
100
150
200
250
300
350
pH Level Distribution
Potability
0
1
<7 is Acidic
>7 is Basic
Count
0
20
40
60
80
0
2
4
6
8
10
12
14
pH Level

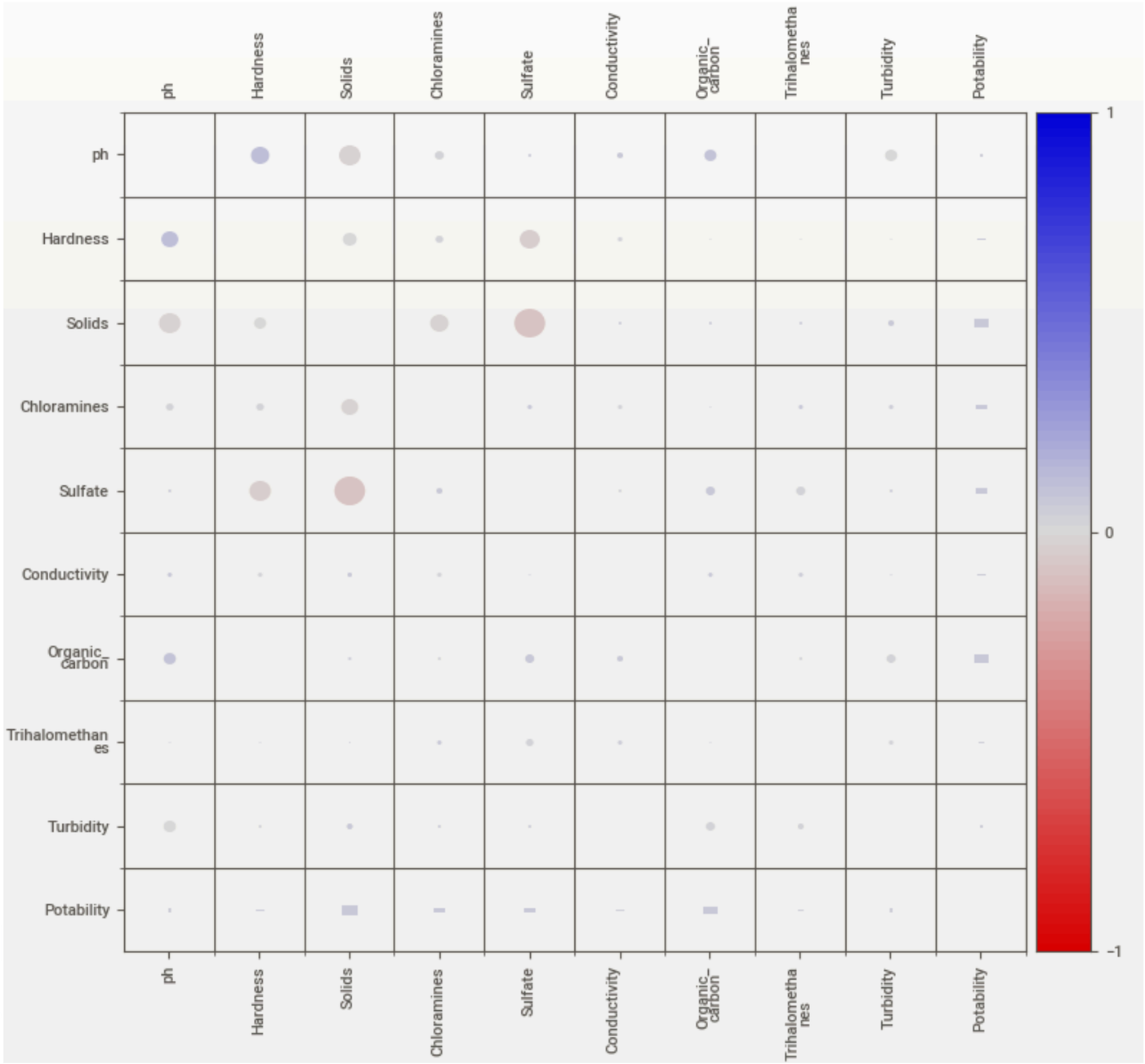

Distribution Of Total Dissolved Solids
Potability
0
1
Count
0
20
40
60
80
100
0
10k
20k
30k
40k
50k
60k
Chloramines Distribution
Potability
0
1
<4 ppm is considered
safe for drinking
Count
0
20
40
60
80
100
120
0
2
4
6
8
10
12
Chloramines (ppm)


Sulfate Distribution
Potability
0
1
<250 mg/L is considered
safe for drinking
Count
100
80
60
40
20
0
150
200
250
300
350
400
450
500
Conductivity Distribution
Potability
0
1
The Conductivity range
is safe for both (200-800),
Potable and Non-Potable water
Count
100
80
60
40
20
0
200
300
400
500
600
700
Conductivity (µS/cm)

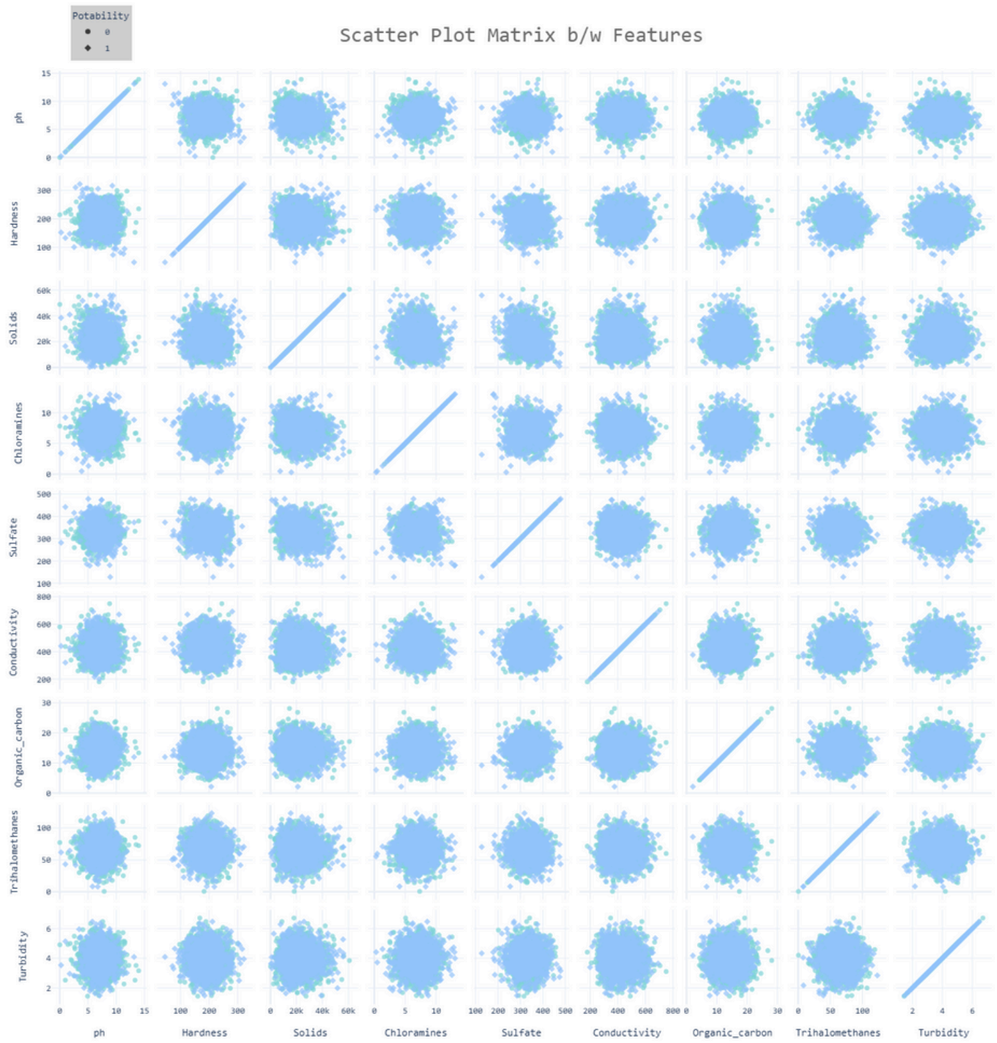

Organic Carbon Distribution
Potability
0
1
Typical Organic Carbon
level is upto 10 ppm
Count
120
100
80
60
40
20
0
5
10
15
20
25
Trihalomethanes Distribution
Potability
0
1
Upper limit of Trihalomethanes
level is 80 µg/L
Count
100
80
60
40
20
0
0
20
40
60
80
100
120
Trihalomethanes (µg/L)

**Association between Features

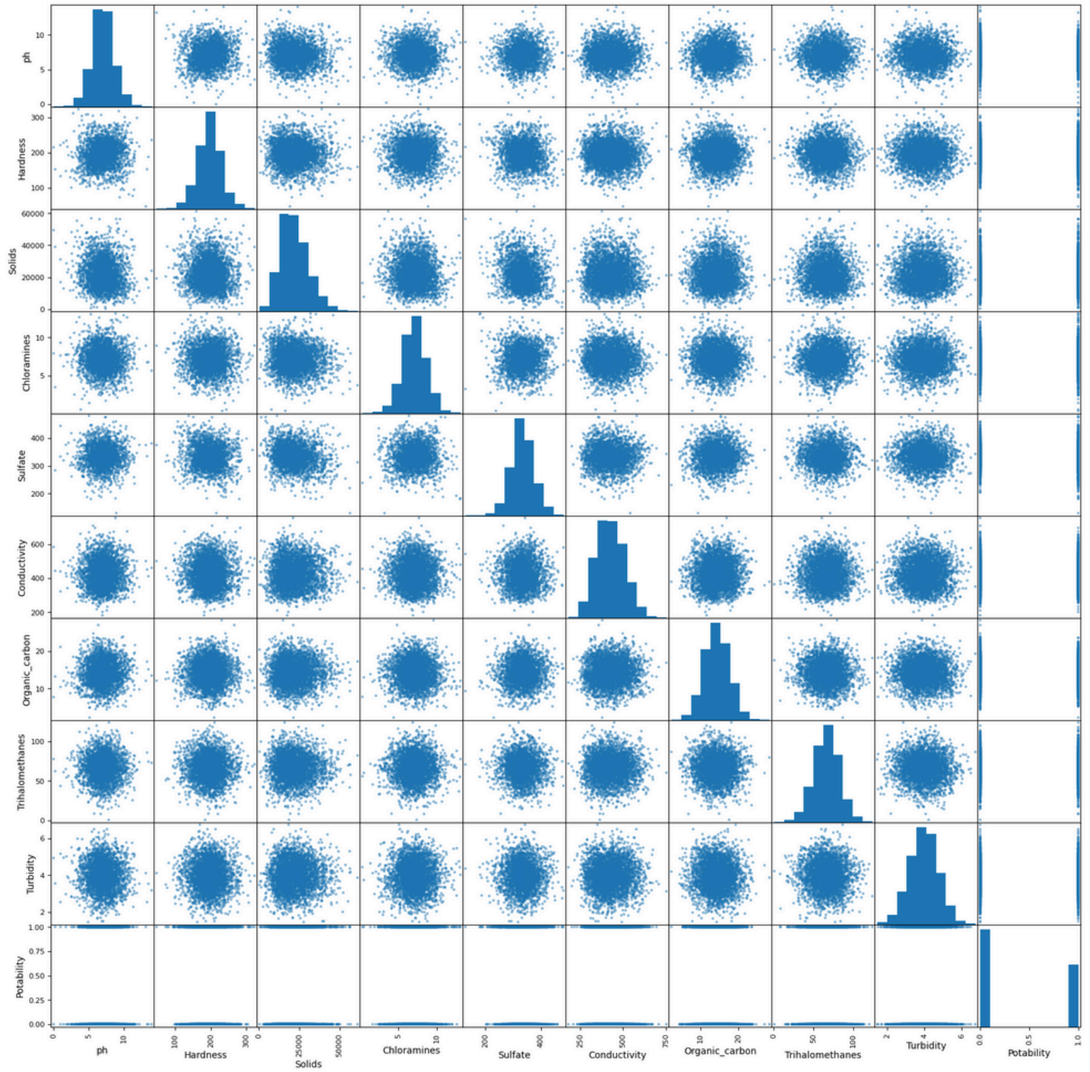




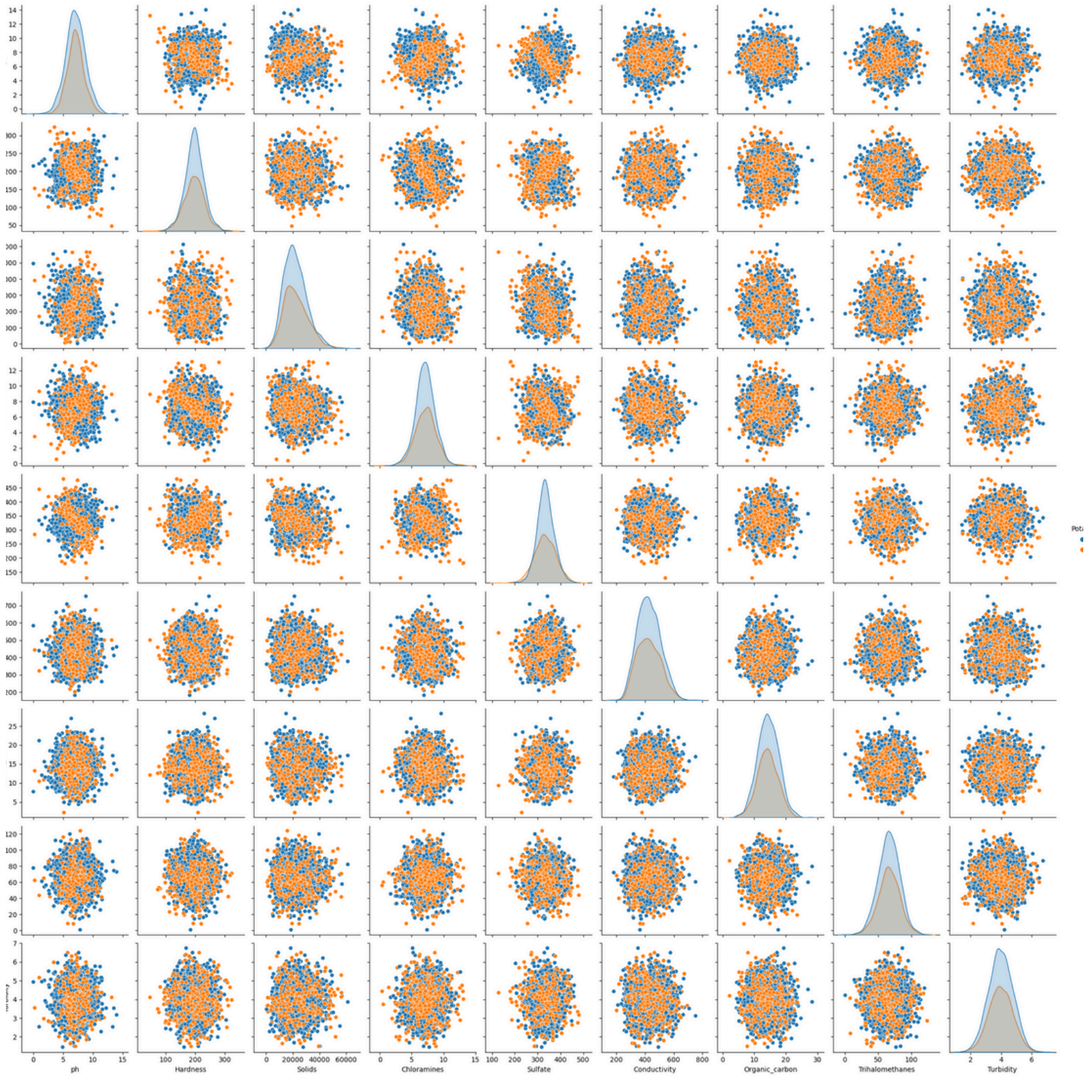
**Scatter Plot



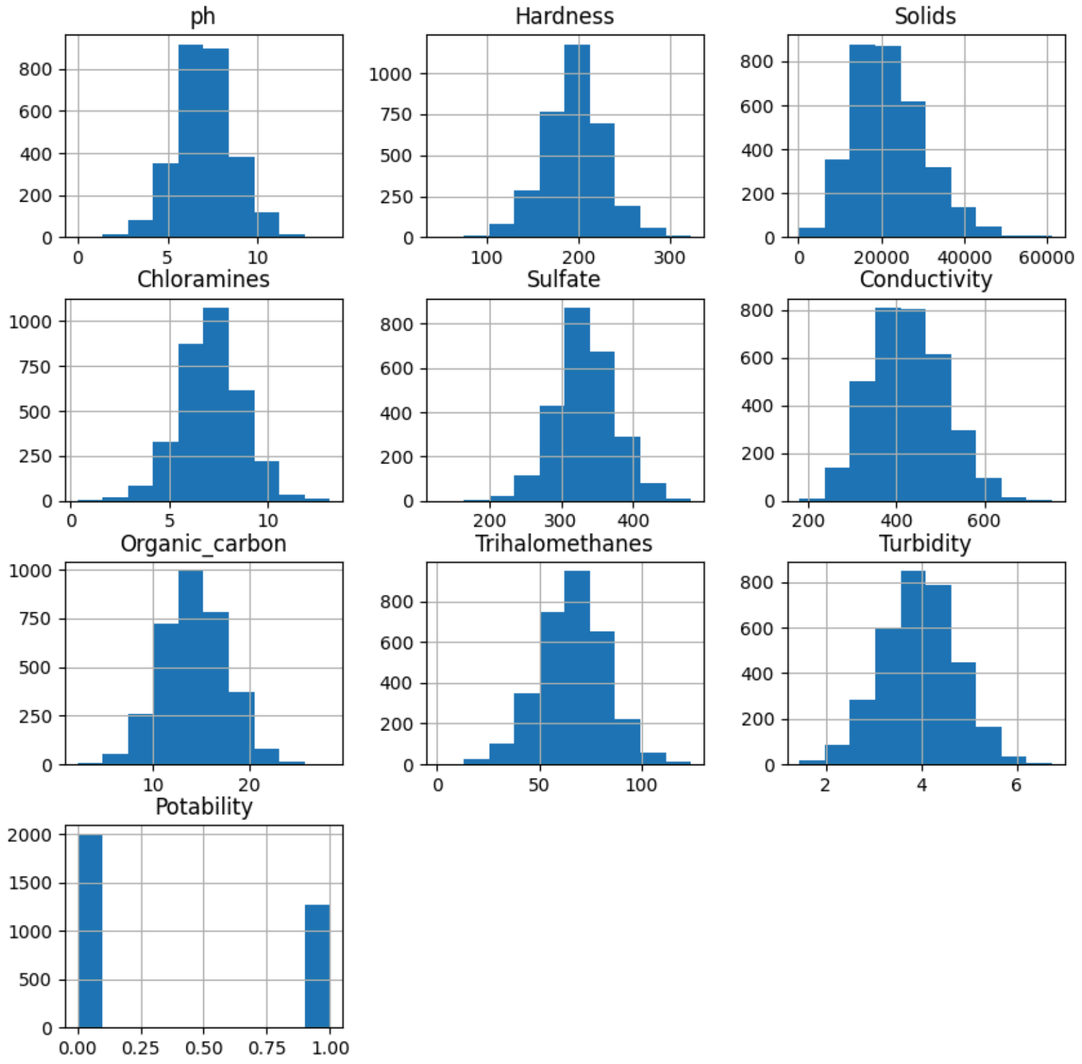
**Scatter Matrix



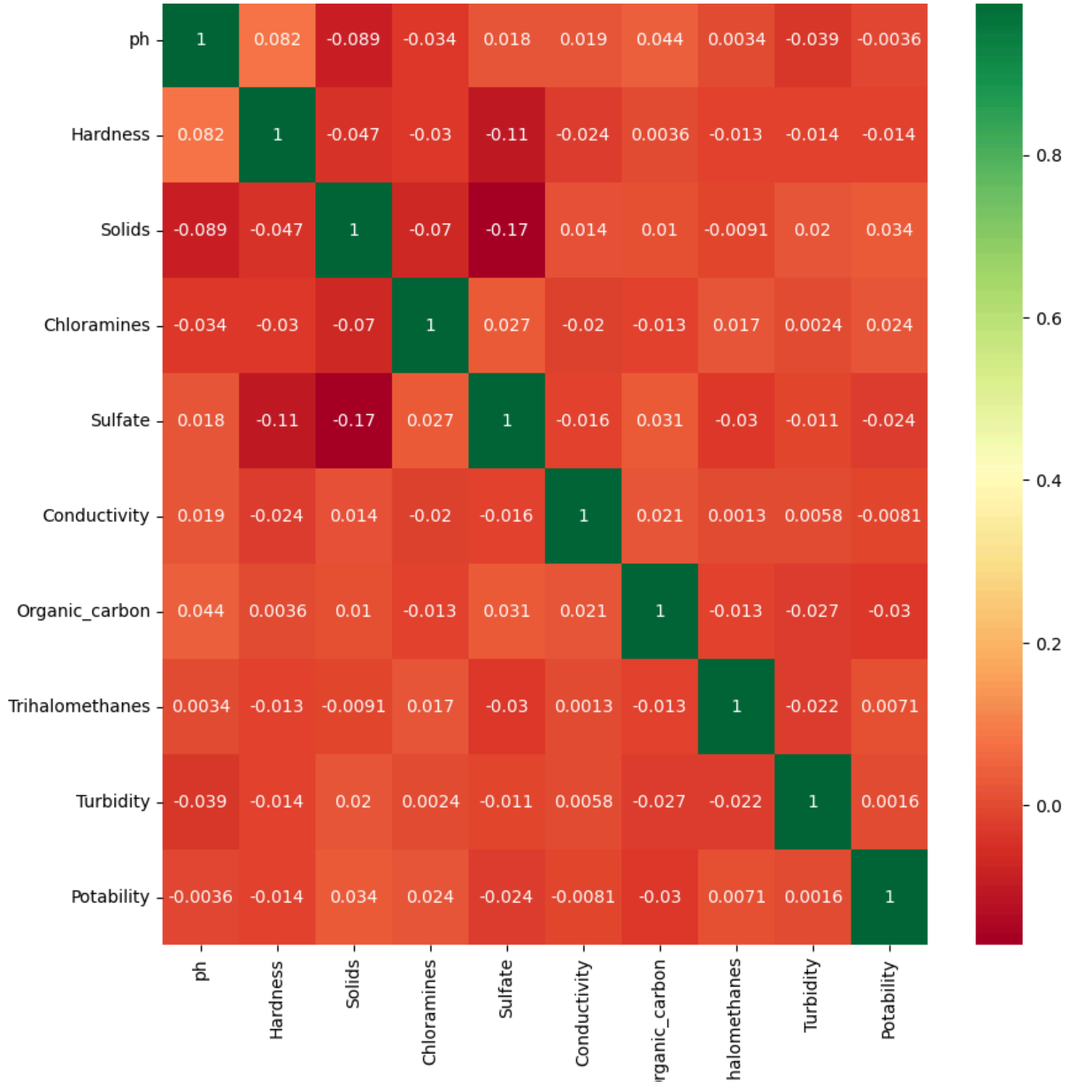
**Pairplot



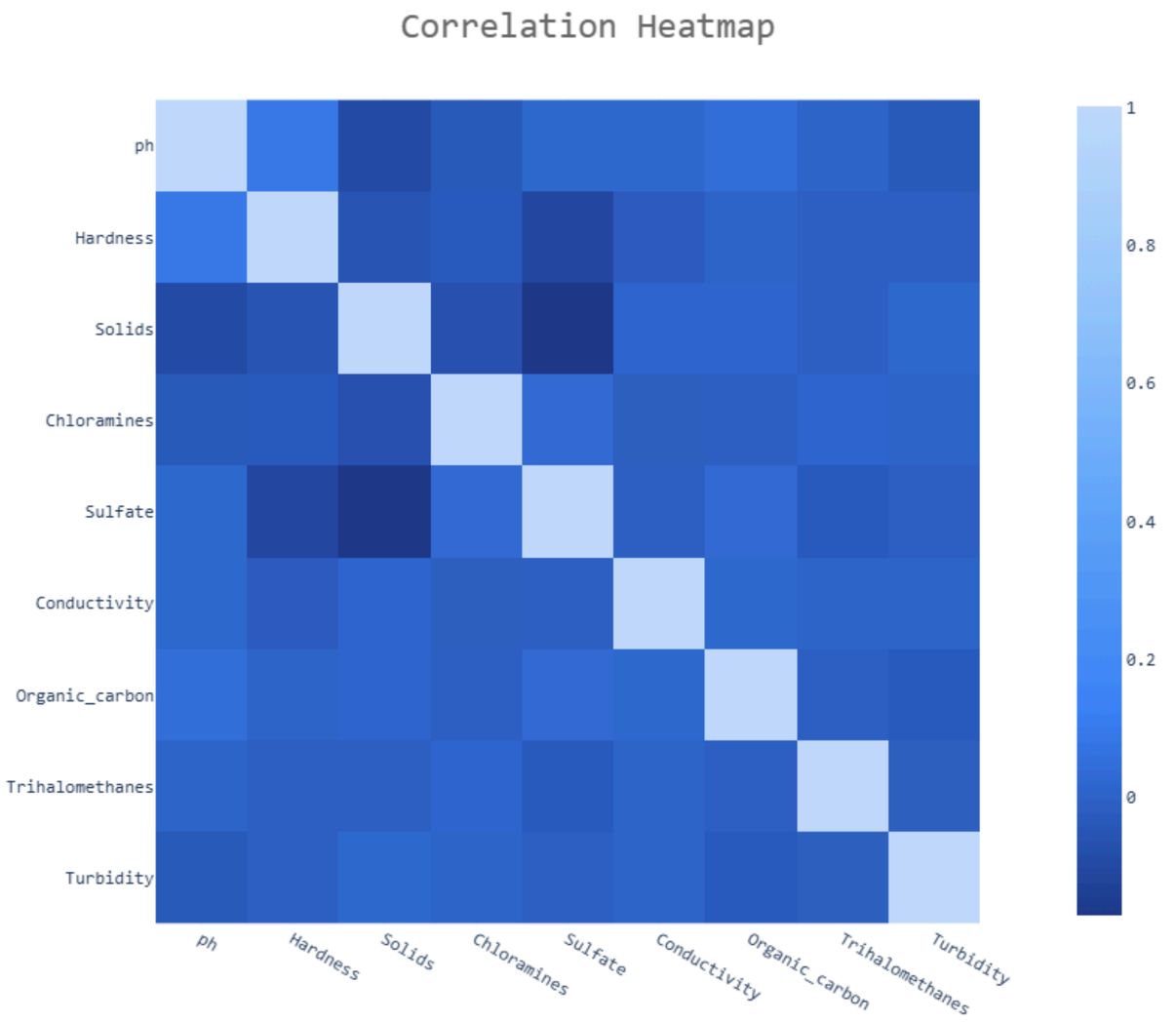
**Histogram Matrix



**Correlation Matrix

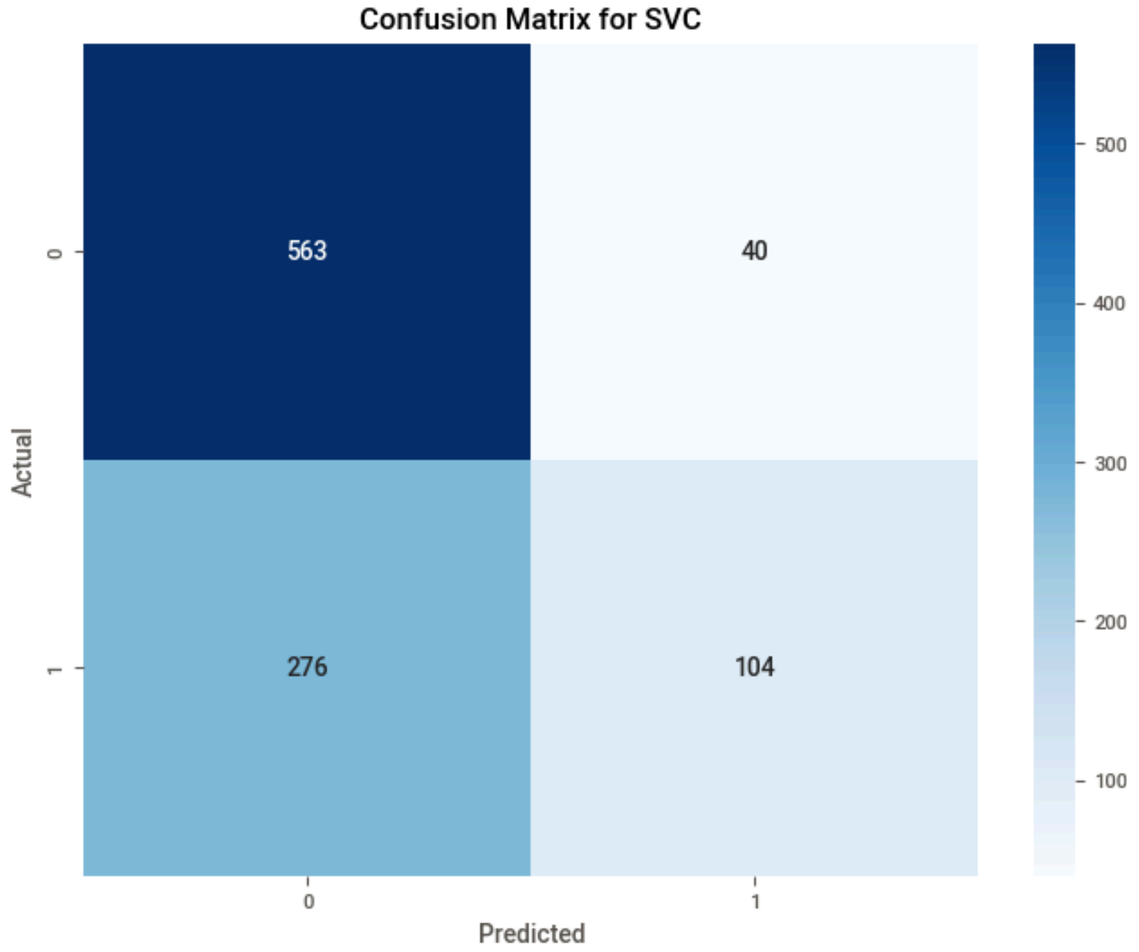


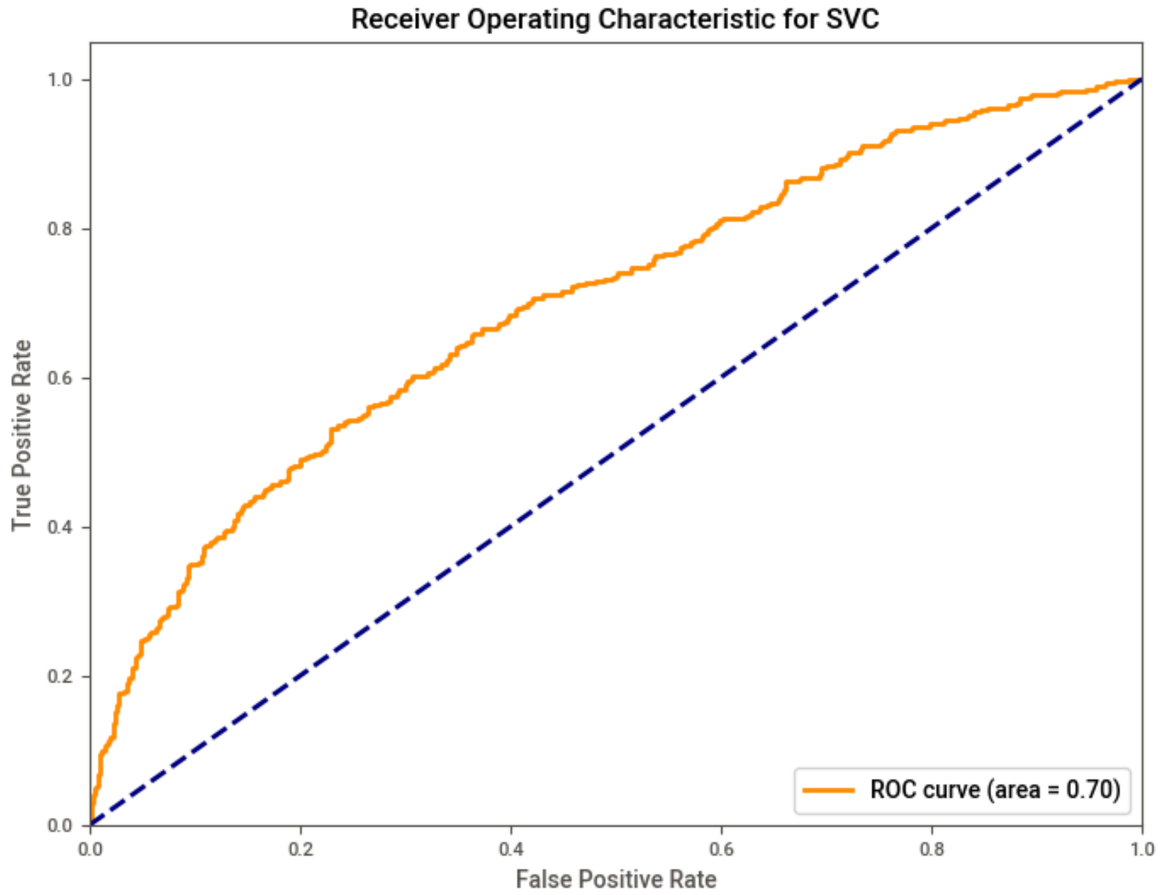
**HeatMap



** Support Vector Classifier

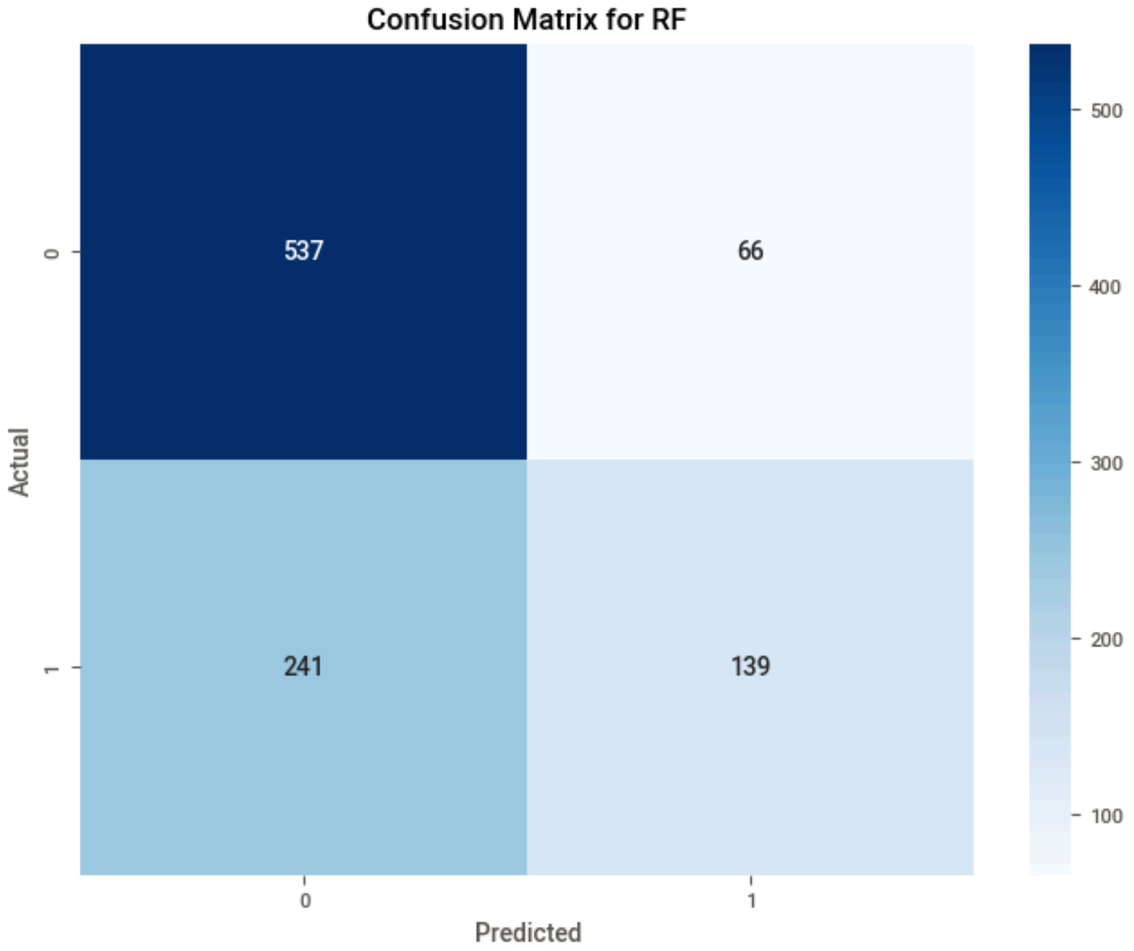
**ROC Curve; Ar(0.70)

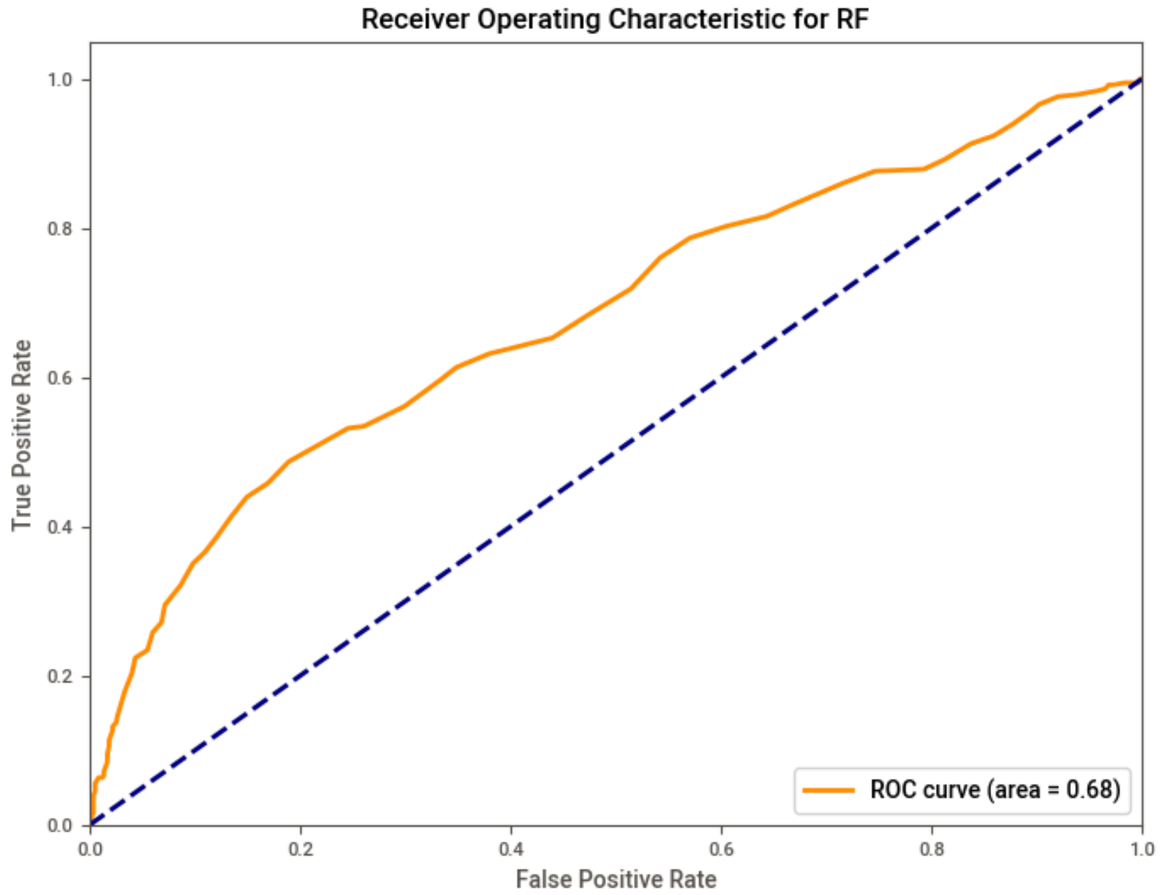




** Random Forest

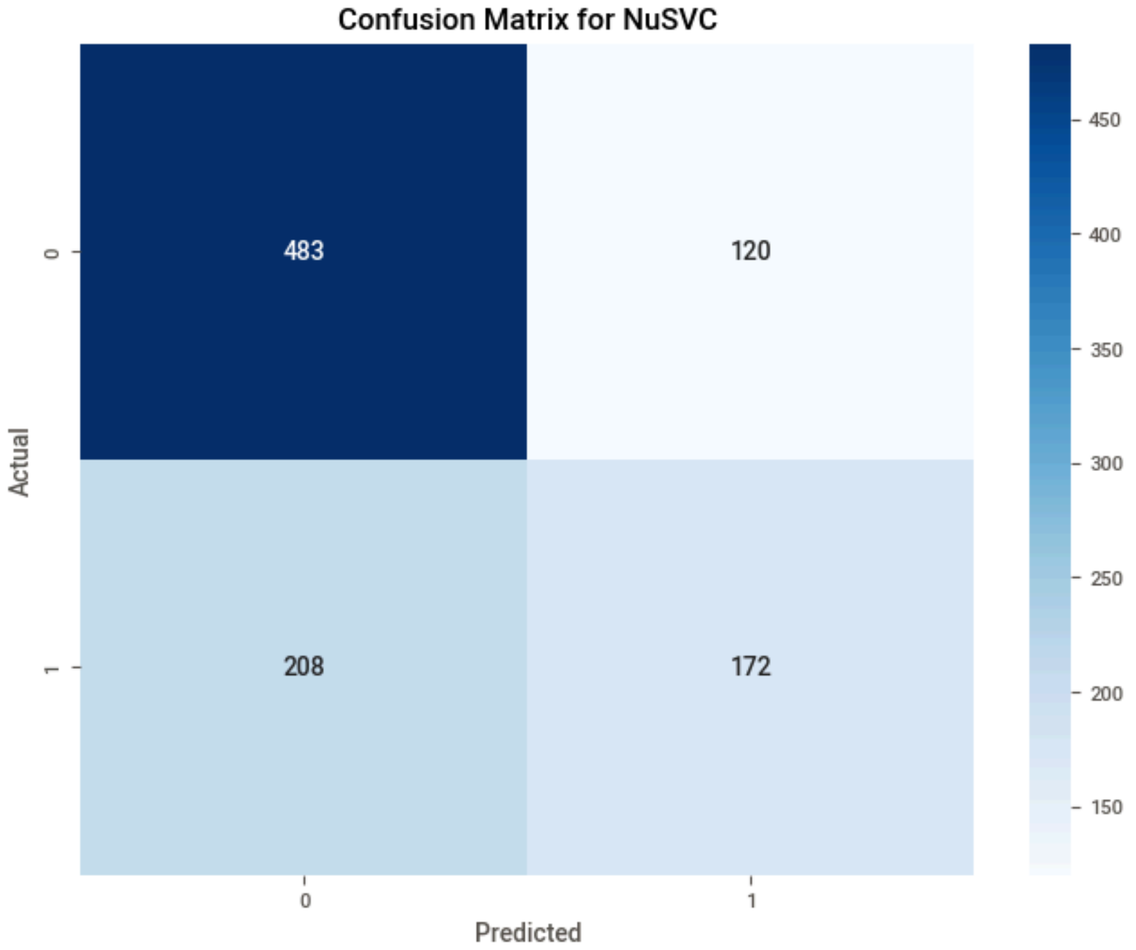
**ROC Curve; Ar(0.68)

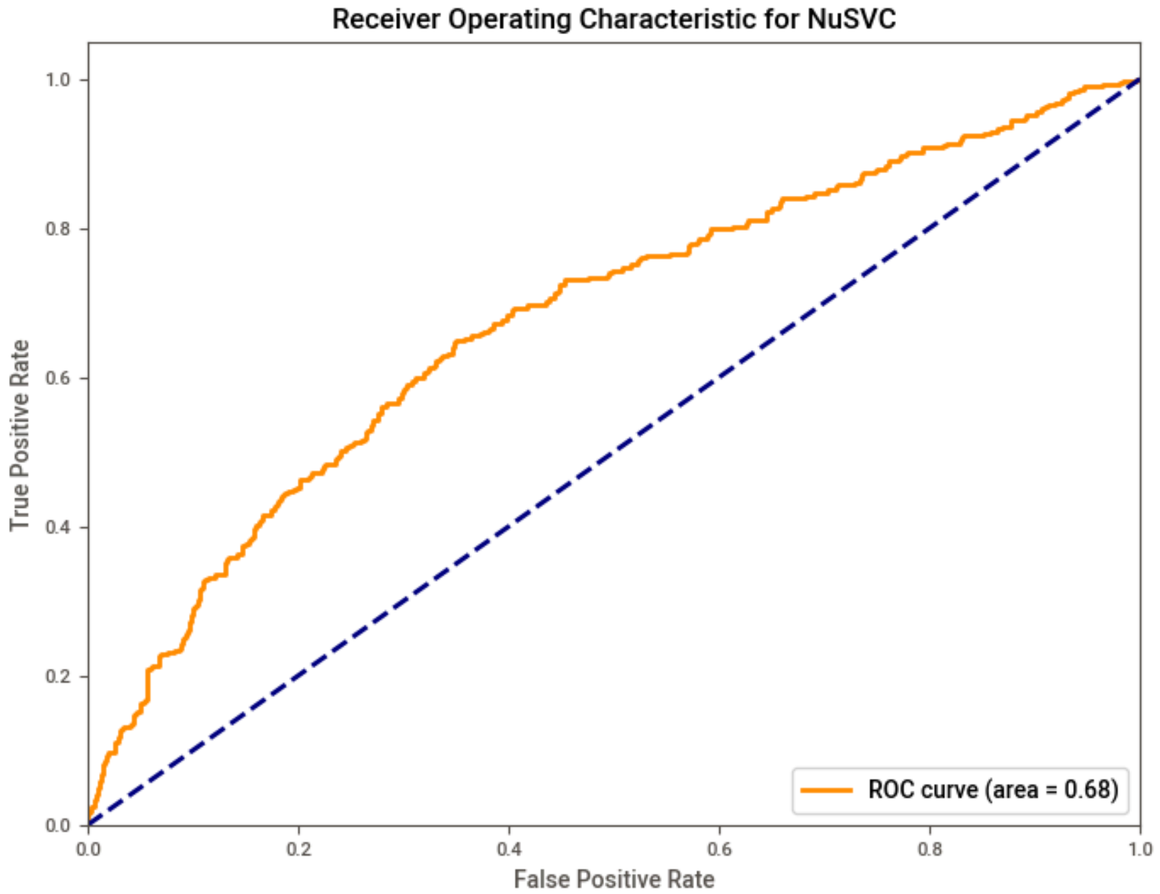




** Nu-Support Vector Classifier

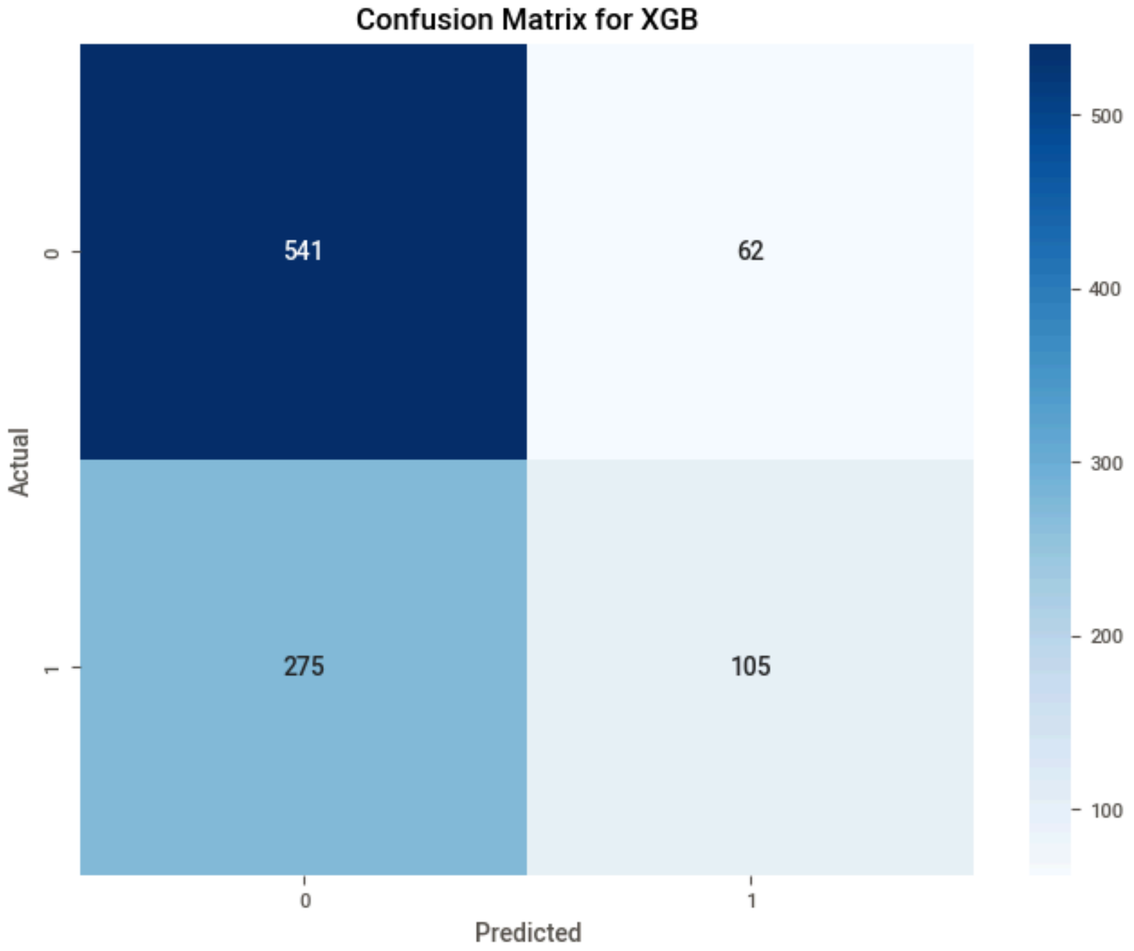
**ROC Curve; Ar(0.68)

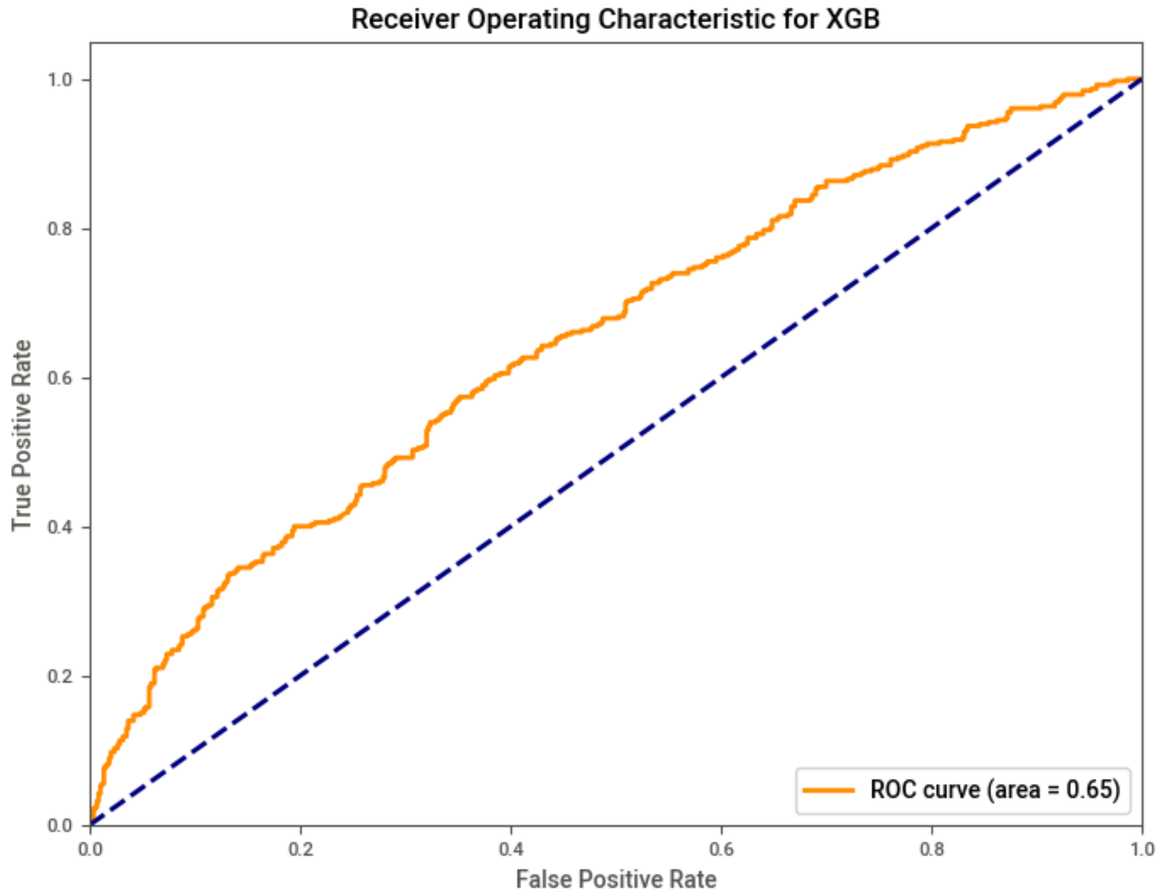




** Extreme Gradient Boosting

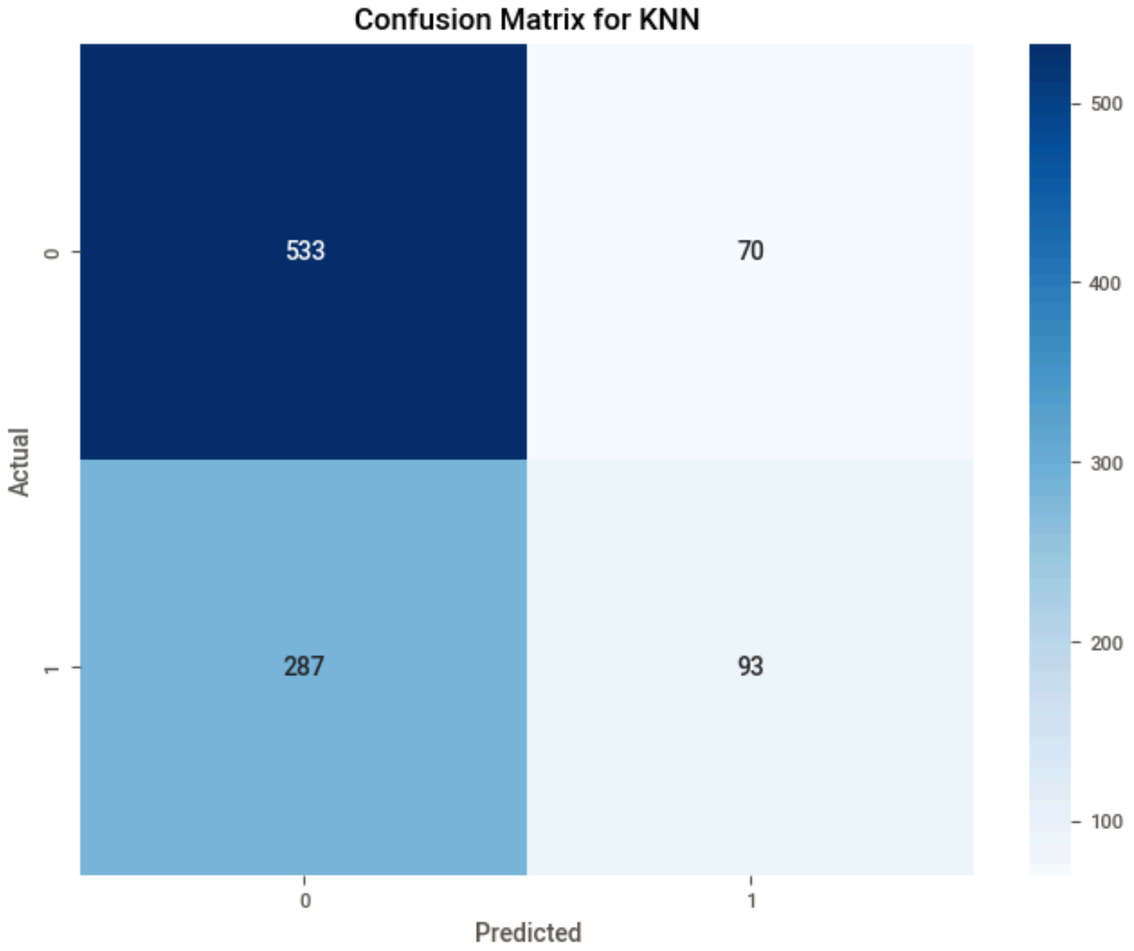
**ROC Curve; Ar(0.65)

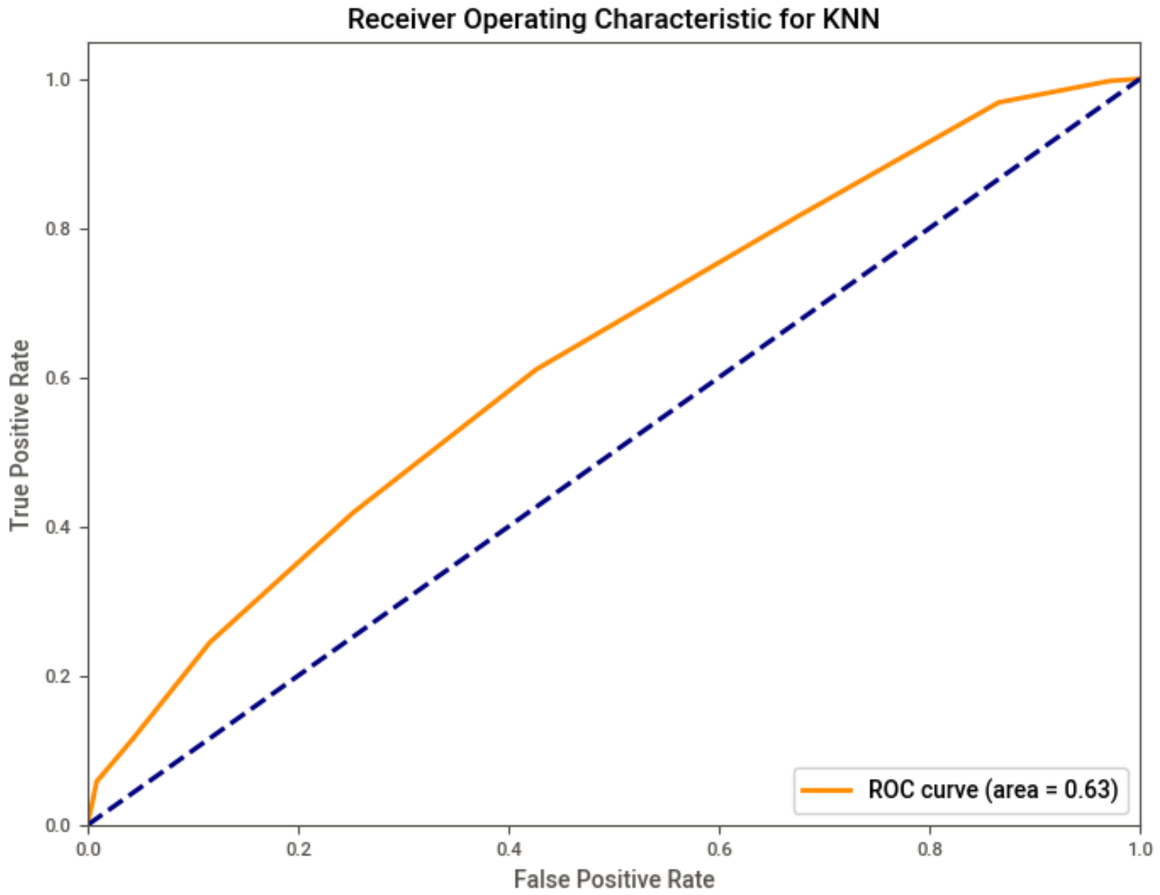




** K-Nearest Neighbors

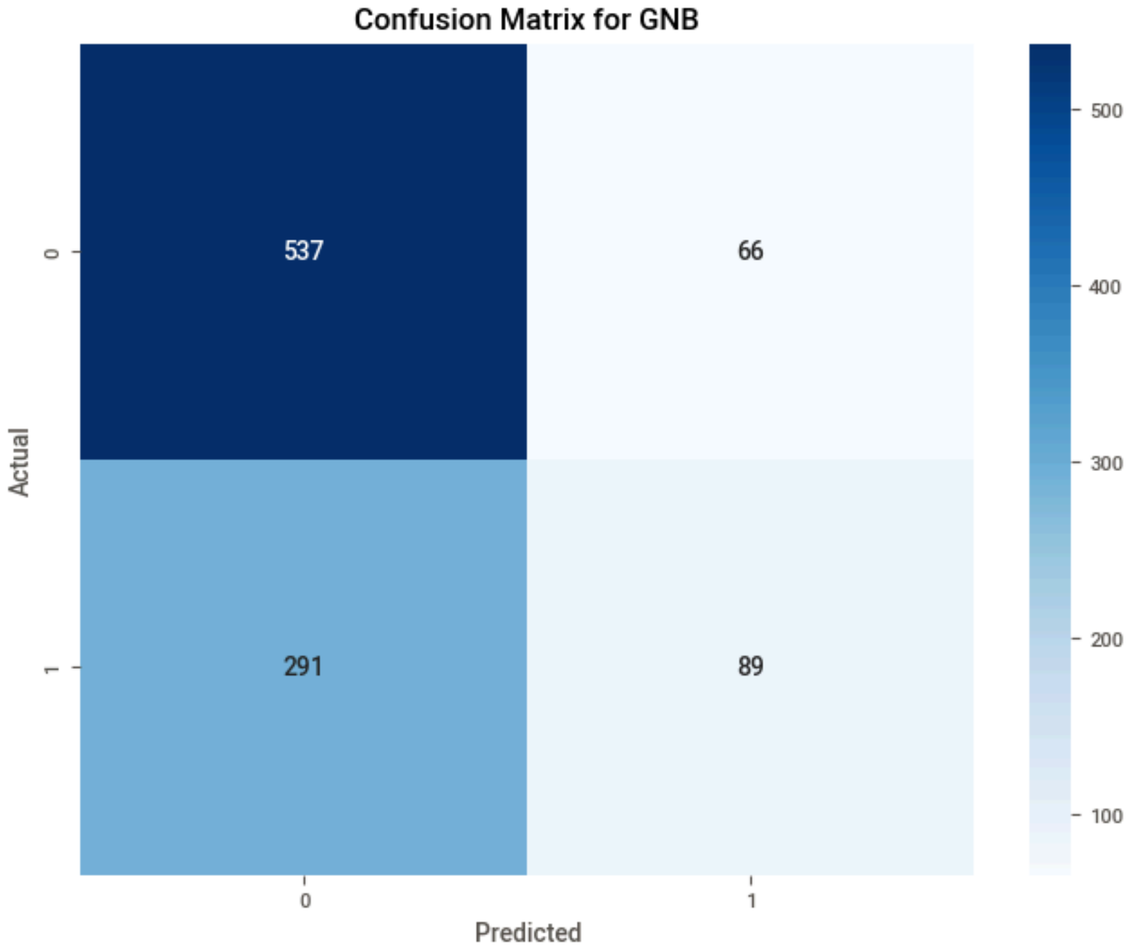
**ROC Curve; Ar(0.63)

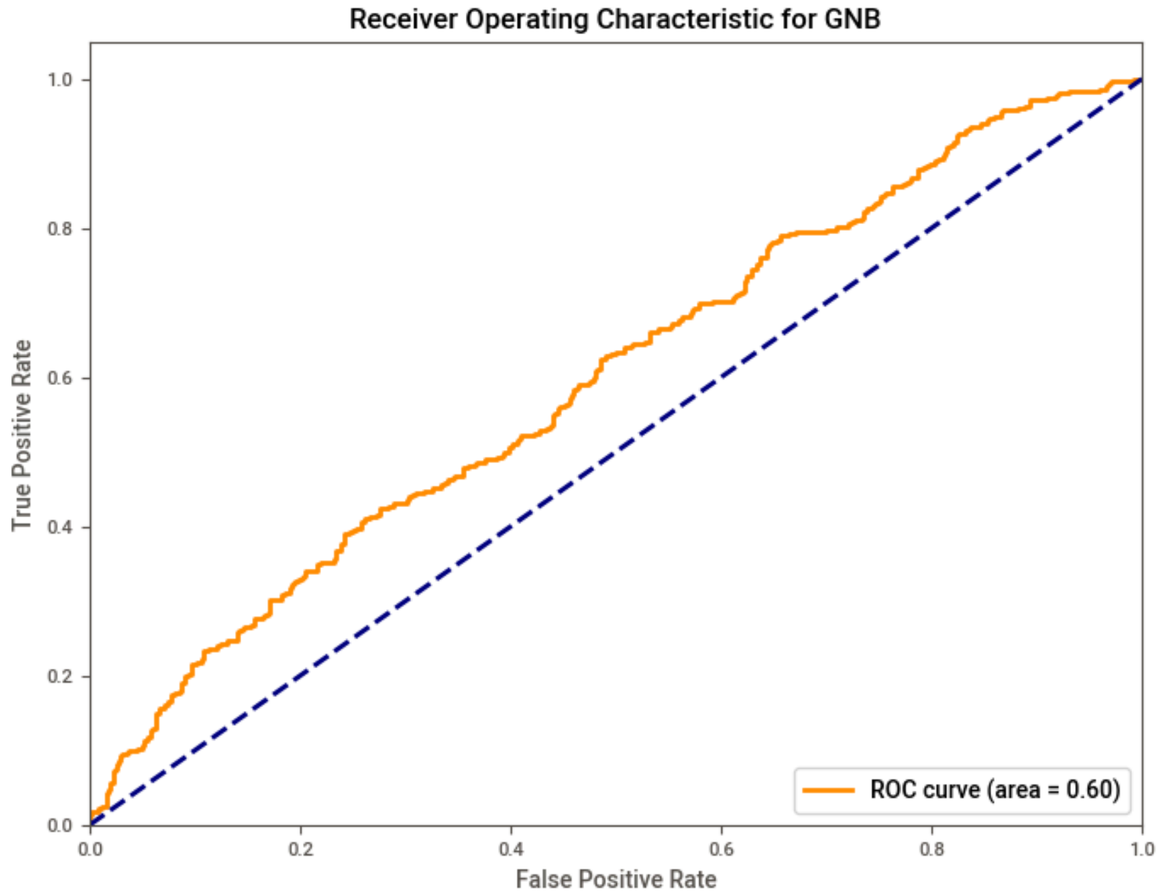




**Gaussian Naive Bayes

**ROC Curve; Ar(0.60)



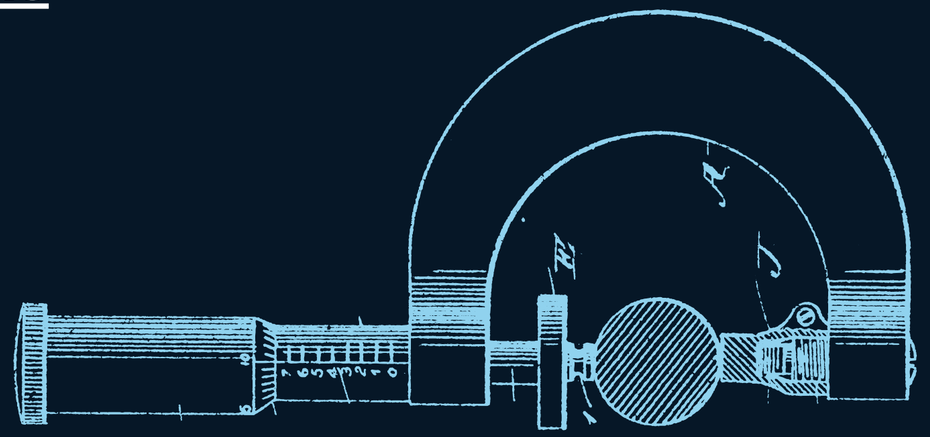


Final Results:

Classification Reports

	Model	Precision Score (%)
0	SVC	69.662959
1	RF	66.853266
2	XGB	64.887763
3	NuSVC	64.401404
4	GNB	61.137214
5	KNN	61.027607
6	DTC	57.009748
7	ADA	56.000324
8	SGDC	52.863041
9	PAC	51.877359
10	Perc	51.669882
11	NC	49.915658
12	LR	30.671414
13	Ridge	30.671414
14	BNB	30.671414

	Precision Score (%)
count	15.000000
mean	53.288031
std	13.102969
min	30.671414
25%	50.792770
50%	56.000324
75%	62.769309
max	69.662959



Classification Report for SVC:

	precision	recall	f1-score	support
0	0.67	0.93	0.78	603
1	0.72	0.27	0.40	380
accuracy			0.68	983
macro avg	0.70	0.60	0.59	983
weighted avg	0.69	0.68	0.63	983

==========

Final Results:

Accuracy Scores

	Model	Accuracy Score (%)
0	SVC	67.853510
1	RF	67.039674
2	NuSVC	66.632757
3	XGB	65.513733
4	KNN	63.682604
5	GNB	63.682604
6	LR	61.342828
7	Ridge	61.342828
8	BNB	61.342828
9	ADA	60.732452
10	DTC	60.528993
11	SGDC	53.814852
12	PAC	52.288911
13	NC	50.966429
14	Perc	50.152594

	Accuracy Score (%)
count	15.000000
mean	60.461173
std	5.928635
min	50.152594
25%	57.171923
50%	61.342828
75%	64.598169
max	67.853510

Classification Report for SVC:

	precision	recall	f1-score	support
0	0.67	0.93	0.78	603
1	0.72	0.27	0.40	380
accuracy			0.68	983
macro avg	0.70	0.60	0.59	983
weighted avg	0.69	0.68	0.63	983

==========

Conclusion

Project Outcome /Summary

1. The TDS levels *seem to contain some discrepancy* since its values are **on an average 40 folds more** than the upper limit for safe drinking water.
2. The data contains almost **equal number of acidic** and **basic pH level** *water samples*.
3. **92%** of the data was *considered* **Hard**.
4. Only **2%** of the water samples were **safe** in terms of *Chloramines levels*.
5. Only **1.8%** of the water samples were safe in terms of *Sulfate levels*.
6. **90.6%** of the water samples had **higher Carbon levels** than the *typical Carbon levels* in drinking water (***10 ppm***).
7. **76.6%** of water samples were safe for drinking in terms of *Trihalomethane levels* in water.
8. **90.4%** of the water samples were safe for drinking in terms of the *Turbidity* of water samples.
9. The ***correlation coefficients*** between the features were **very low**.
10. **Random Forest** and **SVC** worked ***the best*** to train the model.
11. The ensemble method of using the ***Voting Classifier*** on ***Stratified K-folded*** samples gave an accuracy of **>67%.**

References

Data Source /Tools & Softwares Used /Reports

Kaggle*

Tools:

- Jupyter Notebook
- Canva for Graphics
- .seaborn /.matplotlib-pyplot /.plotly for graphs & plots
- .scikit for ML Models
- SDG 06 Report

Thank You!

Team DaSci '24

// Every drop saved today, a river flows tomorrow... //

SUSTAINABLE DEVELOPMENT GOALS

IBM

CSRBOX
CSR Analytics and Collaborations