Water Quality predicting

Programming For Data Science

2024-2025

3rd year Engineering's Degree in Data Science

Department of Applied Mathematics and Statistics

Institute of Technology of Cambodia

Lecturer: Mr. PHAUK Sokkey (Courses)

Mr. PEN Chentra (TP)

Members of Group 03	Student ID
1. NY Chantharith	e20220472
2. SOEUK Bondol	e20221592
3. SAMBATH Seakty	e20220517
4. PHAN Wayotep	e20220850
5. NGOUN Lyhorng	e20221532

Water Quality Predicting

Contents

1. Abstract

Water quality is essential for human health, agriculture, and the environment. However, monitoring and testing water quality can be expensive and time-consuming. This study explores how data analysis and machine learning can be used to predict water quality more efficiently. By using data such as pH levels, turbidity, dissolved oxygen, and other key factors, we trained a machine learning model to classify water as safe or unsafe for various uses. The results showed that machine learning can accurately predict water quality based on patterns in the data, helping to identify contamination early and improve decision-making. This approach is cost-effective and can be applied in areas with limited resources for water testing. It also supports environmental sustainability by promoting better water management practices. This work demonstrates the potential of technology to address critical challenges in water quality monitoring and protection.

2. Introduction

Water is one of the most vital natural resources on Earth, essential for the survival of all living organisms. Ensuring the quality of water is paramount, as contaminated water can lead to serious health issues, environmental degradation, and economic losses. Water quality is generally assessed by analyzing various physicochemical parameters, including pH levels, turbidity, dissolved oxygen, and the presence of harmful contaminants. These parameters help classify water into different quality levels, ranging from portable to polluted. Predicting water quality is a challenging yet critical task that can greatly benefit from advancements in data science. Traditional methods of water quality assessment often involve manual sampling and laboratory analysis, which can be time-consuming, labor-intensive, and expensive. In contrast, leveraging data science methodologies allows for the analysis of large datasets collected over time and across various regions, enabling faster and more efficient decision-making processes. The primary objective of this study is to develop a robust model that can accurately predict water quality based on historical and real-time data. This involves applying data science techniques such as data preprocessing, exploratory data analysis (EDA), and machine learning algorithms. The outcomes of such predictions can support proactive measures to ensure water safety, optimize resource allocation, and improve public health outcomes. This report provides a comprehensive framework for water quality prediction, starting from data collection and cleaning to model construction and evaluation. The dataset used in this study includes multiple parameters that influence water quality, such as total dissolved solids (TDS), electrical conductivity (EC), temperature, and chemical oxygen demand (COD). By analyzing these parameters, patterns and trends can be identified, leading to actionable insights. The field of data science plays a pivotal role in addressing environmental challenges, including water quality management. By utilizing machine learning algorithms, this study aims to classify water samples into predefined quality categories or predict specific water quality metrics. Techniques such as feature selection, data normalization, and model optimization are applied to enhance the performance of predictive models. Moreover, this report emphasizes the importance of data visualization and exploratory data analysis as essential steps in understanding the dataset and identifying relationships between variables. For example, a heatmap showing correlations among parameters can help determine which factors have the most significant impact on water quality. Such insights can guide the selection of features for machine learning models, improving their accuracy and interpretability. This study also explores the application of advanced machine learning techniques, including ensemble learning and hyperparameter tuning, to achieve optimal performance. Evaluation metrics such as accuracy, precision, recall, and F1 score are used to assess the effectiveness of the models, ensuring their reliability in real-world applications. In conclusion, predicting water quality using data science not only enhances the efficiency of water monitoring systems but also contributes to sustainable water resource management. The insights gained from this study can inform policymakers, water management

authorities, and researchers in their efforts to ensure the availability of clean and safe water for all. By integrating data science with environmental studies, this report highlights the potential for innovative solutions to some of the most pressing global challenges.

3. Data Collection

Data collection is a critical step in building a reliable water quality prediction model. For this study, data was sourced from reputable organizations and databases, including government agencies, environmental research institutes, and publicly available datasets. These sources ensure the credibility and comprehensiveness of the data used.

4. Exploring Data Analysis

Exploratory Data Analysis (EDA) is a crucial step in understanding and preparing your dataset for predictive modeling. It helps identify patterns, detect anomalies, and generate insights that inform feature selection and model development. Below is an organized guide to exploring the data for a water quality prediction project:

4.1. Data Cleaning

Data cleaning is a critical step to ensure the reliability of subsequent analyses. The cleaning process for this study involved:

- Identifying Missing Values: Analyzed the dataset to locate missing entries and employed imputation techniques (mean, median, or mode) to fill gaps. In some cases, records with excessive missing data were removed.
- Outlier Detection: Used statistical methods such as the Z-score and interquartile range (IQR) to detect and treat extreme values. Outliers were either removed or adjusted based on their potential impact.
- Standardization and Normalization: Transformed data to ensure consistency in measurement units and scales. Parameters with significantly different ranges were normalized to facilitate comparison and analysis.
- Removing Redundancies: Eliminated duplicate records and redundant features that did not contribute to the predictive analysis.

Ensure data is accurate, consistent, and ready for analysis.

	Index	pH	Iron	Nitrate	Chloride	Lead	Zinc	Color	Turbidity	Fluoride	...	Chlorine	Manganese	Total Dissolved Solids	Source	Water Temperature	Air Temperature	Month	Day	Time of Day	Target
1	2	5.443762	2.010586e-02	3.816994	230.995630	5.290000e-76	0.528280	Light Yellow	0.319956	0.423423	...	3.560224	7.007989e-02	570.054094	River	11.643467	44.891330	January	31.0	8.0	0
5	52	8.460833	1.986337e-02	8.601511	134.202428	1.170000e-256	2.600728	Faint Yellow	0.249098	0.515965	...	2.494293	5.710000e-10	13.925614	Aquifer	7.611181	82.674304	January	2.0	23.0	0
6	64	8.194406	3.387248e-03	8.344541	248.043661	2.410000e-71	1.993738	Near Colorless	0.019442	0.355717	...	3.781047	2.220000e-05	297.621227	Spring	15.786764	35.653137	January	2.0	19.0	0
8	90	5.812626	1.061910e-04	3.032464	199.084282	4.120000e-114	0.951398	Faint Yellow	1.613035	0.029181	...	2.374635	1.848483e-01	188.786881	Ground	9.818649	41.814583	January	8.0	0.0	0
9	116	6.806017	2.458747e-01	6.685902	98.370632	2.420000e-90	4.279026	Near Colorless	0.084110	0.499284	...	3.158896	6.486010e-03	479.485597	River	16.353545	74.757185	January	5.0	12.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
88211	1048529	7.786804	2.044458e-01	4.537733	218.653193	2.670000e-198	0.915345	Near Colorless	0.000383	1.564805	...	4.126209	2.840000e-10	100.008291	Spring	16.399880	77.623536	January	11.0	22.0	0
88213	1048540	7.949245	4.640000e-08	2.394735	167.133580	4.160000e-93	0.005478	Near Colorless	0.000395	0.173191	...	3.866278	2.080000e-11	463.738017	Stream	13.485630	63.273387	January	18.0	12.0	0
88214	1048554	7.228452	7.080000e-13	6.643336	145.662056	4.110000e-139	1.437379	Faint Yellow	0.434500	1.188041	...	2.193430	2.504174e-03	421.057598	River	12.740142	56.691924	January	10.0	4.0	0
88215	1048562	6.749023	2.860000e-07	5.431533	153.103341	1.130000e-69	0.064376	Near Colorless	0.897031	1.281133	...	2.319100	8.615916e-03	12.503163	Spring	18.652998	50.339735	January	17.0	13.0	0
88216	1048566	7.922532	1.910188e-02	2.239796	165.107143	6.280000e-131	3.017865	Colorless	0.547606	1.029528	...	3.162530	4.252938e-02	412.149003	Lake	14.604997	81.806878	January	20.0	21.0	0

Figure 1 : Data Cleaning

```
Column: Total Dissolved Solids - No outliers detected.
...
```

	Column	Outlier Count	Lower Bound	Upper Bound
0	pH	2292	5.286231e+00	9.640777e+00
1	Iron	14119	-7.368455e-02	1.228314e-01
2	Nitrate	2603	-1.475181e+00	1.297446e+01
3	Chloride	2473	1.963160e+01	3.333875e+02
4	Lead	21670	-3.060000e-27	5.100000e-27
5	Zinc	2516	-2.291450e+00	4.917853e+00
6	Turbidity	6053	-7.875617e-01	1.409430e+00
7	Fluoride	2751	-1.059308e+00	2.749785e+00
8	Copper	4162	-7.097893e-01	1.523244e+00
9	Sulfate	1915	-2.777821e+01	3.060582e+02
10	Conductivity	1637	-8.234693e+01	8.991140e+02
11	Chlorine	1405	1.305129e+00	5.106773e+00
12	Manganese	13721	-2.105224e-02	3.509259e-02
13	Water Temperature	3611	-7.549653e+00	4.272384e+01
14	Air Temperature	626	1.153356e+01	1.087469e+02

Figure 2 : Check Outlier

4.2. Data Visualization

Visualization played a crucial role in uncovering patterns and insights within the dataset. Key techniques included:

- Histograms: Used to examine the distribution of individual parameters, such as pH levels and turbidity, providing insights into their central tendency and variability.
- Box Plots: Highlighted the presence of outliers and variability in key water quality metrics.
- Scatter Plots: Explored relationships between pairs of variables, such as turbidity versus TDS, to identify potential correlations or trends.
- Heatmaps: Generated to visualize correlations among multiple parameters, revealing which factors were closely interrelated and could influence water quality predictions.
- Time Series Analysis: Plotted temporal trends for specific parameters, such as seasonal variations in dissolved oxygen levels or temperature fluctuations.

Understand data distributions, relationships, and trends visually.

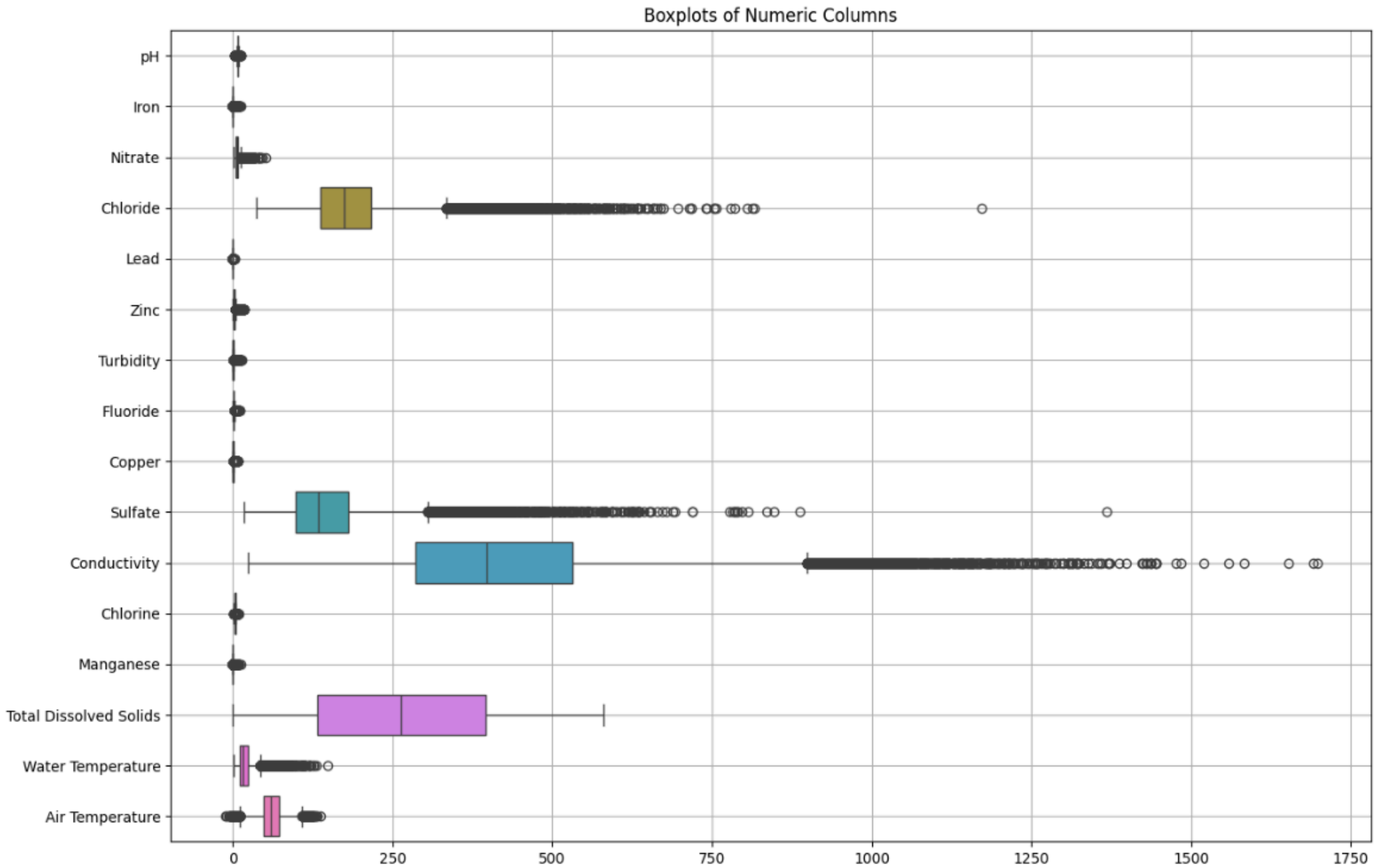

Figure 3 : Box Plots

Before we can decide which of these features need to use any engineer technique or remove we need to find the skewness of it.

Numeric columns	Apply for
pH is symmetrically distributed	no transformation applied
Iron is highly skewed	applying log transformation
Nitrate is highly skewed	applying log transformation
Chloride is highly skewed	applying log transformation
Lead is highly skewed	applying log transformation
Zinc is highly skewed	applying log transformation
Turbidity is highly skewed	applying log transformation
Fluoride is highly skewed	applying log transformation
Copper is highly skewed	applying log transformation
Sulfate is highly skewed	applying log transformation
Conductivity is mildly skewed	applying square root transformation
Chlorine is mildly skewed	applying square root transformation
Manganese is highly skewed	applying log transformation
Total Dissolved Solids is symmetrically distributed	no transformation applied
Water Temperature is highly skewed	applying log transformation
Air Temperature is symmetrically distributed	no transformation applied

```
New Skewness after transformations:
pH                        -0.100035
Iron                       4.882636
Nitrate                    0.130381
Chloride                  -0.073298
Lead                      35.232441
Zinc                       0.450475
Turbidity                  1.974282
Fluoride                   0.588614
Copper                     1.317122
Sulfate                   -0.167893
Conductivity               0.235059
Chlorine                   0.255393
Manganese                  5.164318
Total Dissolved Solids     0.052642
Water Temperature          0.094707
Air Temperature            0.009079
```

Figure 4 : Skewness for each feature

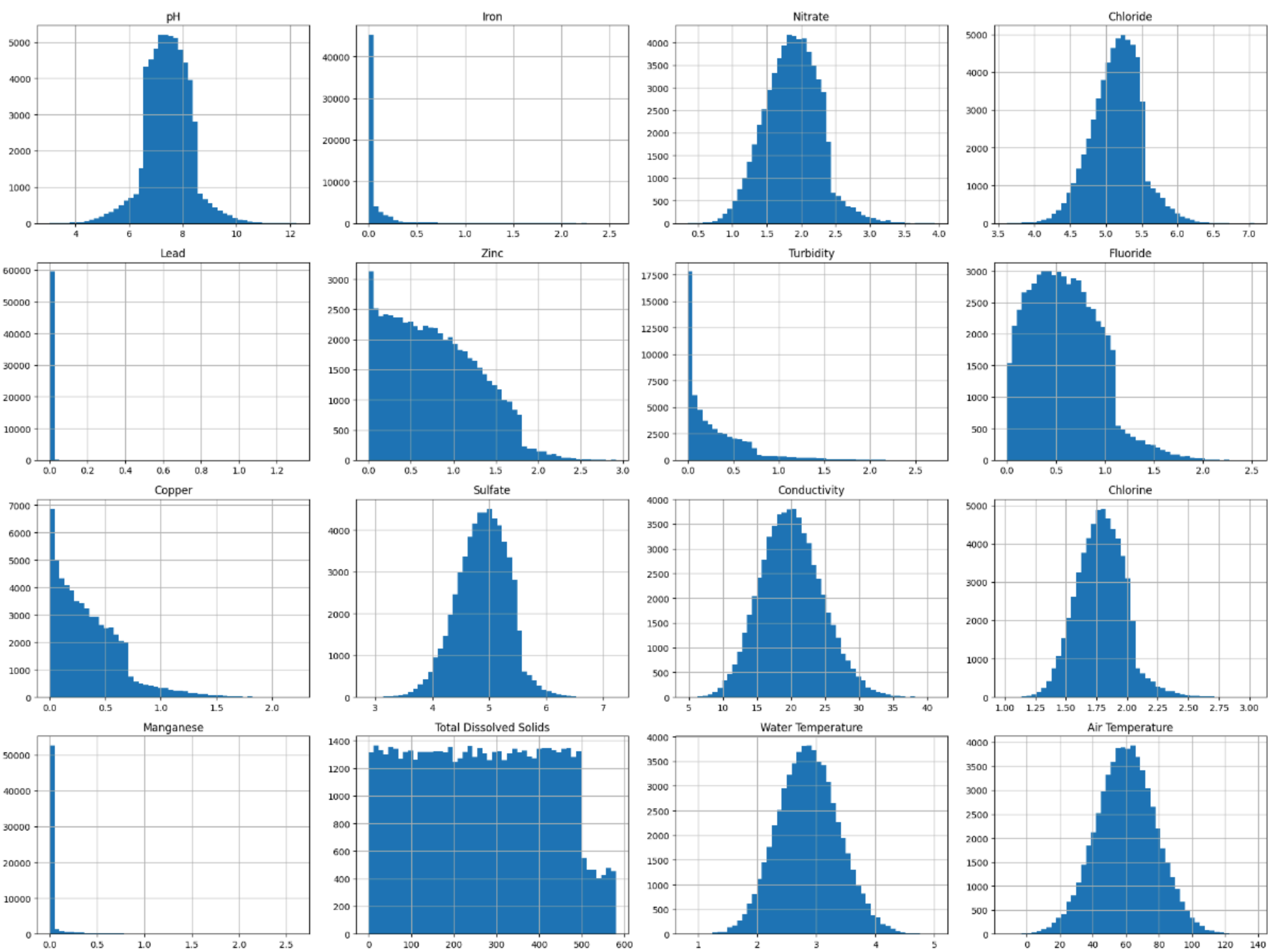

Figure 5 : Transformed Data Distribution

From the updated skewness values after applying transformations, some features are still highly skewed

- Features with High Skewness (Post-Transformation):
  - Iron: 4.88
  - Lead: 35.23
  - Turbidity: 1.97
  - Copper: 1.32
  - Manganese: 5.16
  - Despite log transformations, these features remain highly skewed.

```
New Skewness after transformations:
Iron          1.145256
Turbidity     0.259167
Copper        0.089594
Manganese     1.738080
Lead          0.000000
```

Figure 6 : High Skewness (Post-Transformation)

For features still highly skewed after log transformation, consider further adjustments:

- Apply Box-Cox transformation
- clip the extreme value of highly skew

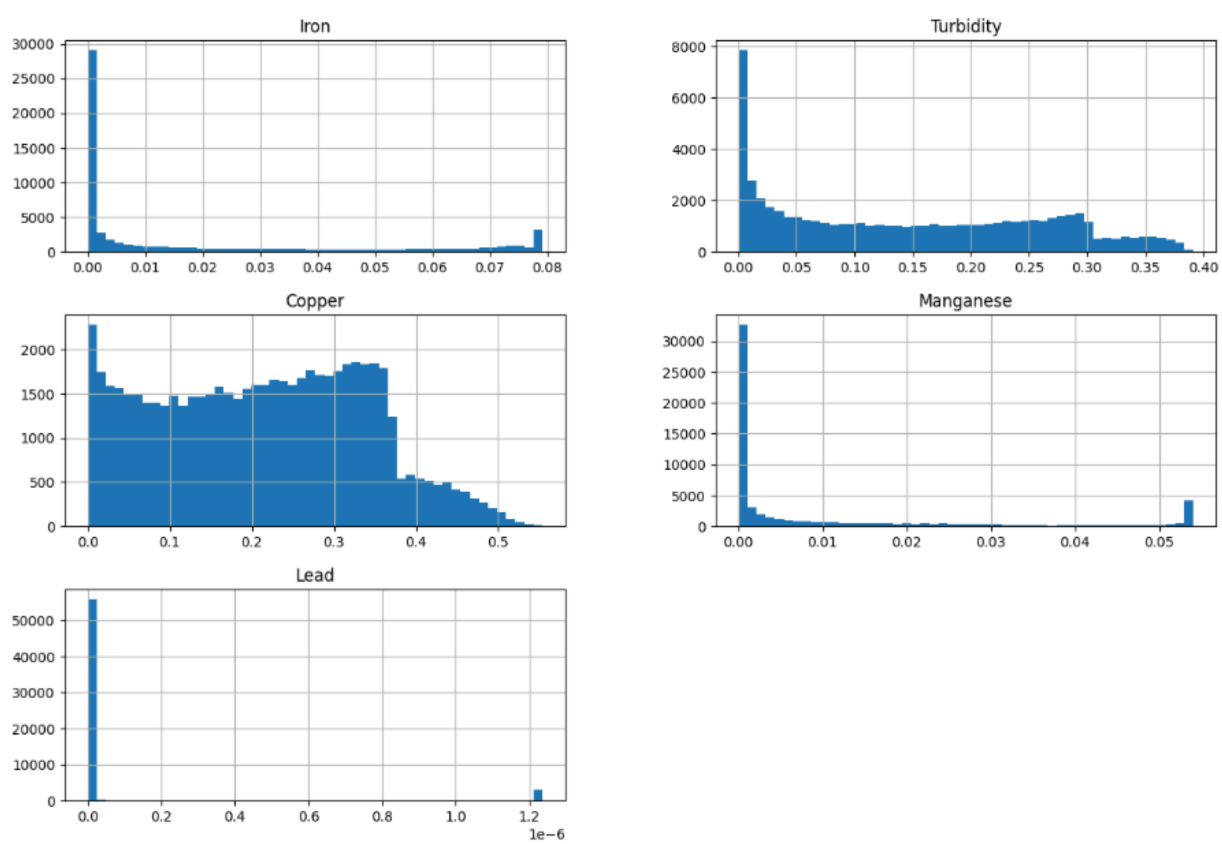

Figure 7 : Transformed Data Distribution

```python
# Boxplot for each numeric column
plt.figure(figsize=(15, 10))
sns.boxplot(data=transformed_data[numeric_cols], orient='h')
plt.title("Boxplots of Numeric Columns")
plt.grid(True)
plt.show()
```

Output:

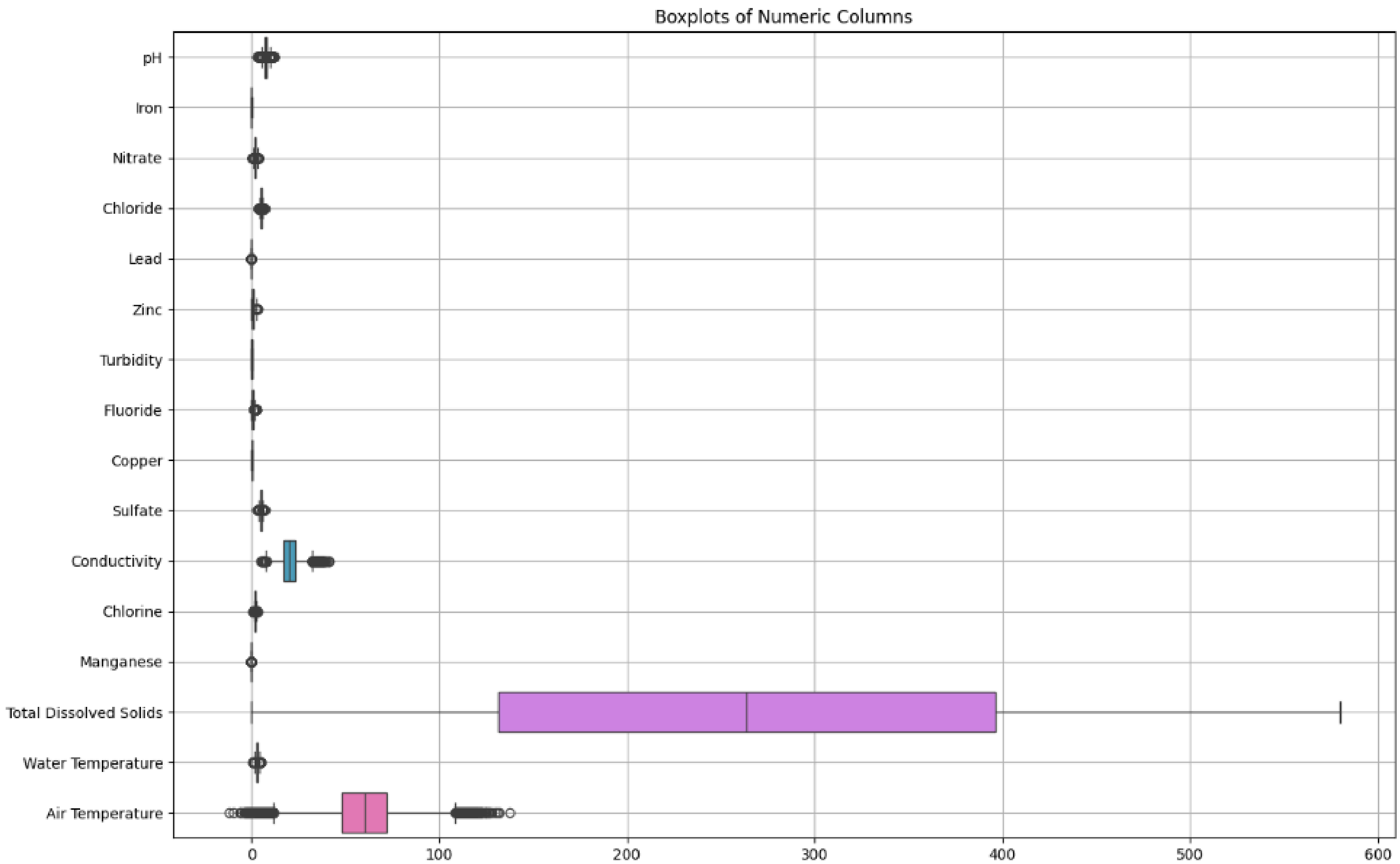

Figure 8 : Box Plots new check outlier update

	Index	pH	Iron	Nitrate	Chloride	Lead	Zinc	Color	Turbidity	Fluoride	...	Chlorine	Manganese	Total Dissolved Solids	Source	Water Temperature	Air Temperature	Month	Day	Time of Day	Target
1	2	5.443762	0.017448	1.572150	5.446719	5.290000e-76	0.424143	1	0.184143	0.353065	...	1.886856	3.797756e-02	570.054094	1	2.537141	44.891330	January	31.0	8.0	0
5	52	8.460833	0.017266	2.261920	4.906773	1.170000e-256	1.281136	2	0.158578	0.416052	...	1.579333	5.709999e-10	13.925614	2	2.153062	82.674304	January	2.0	23.0	0
6	64	8.194406	0.003305	2.234792	5.517628	2.410000e-71	1.096523	3	0.018634	0.304331	...	1.944491	2.219495e-05	297.621227	3	2.820591	35.653137	January	2.0	19.0	0
8	90	5.812626	0.000106	1.394378	5.298739	4.120000e-114	0.668546	2	0.329479	0.028763	...	1.540985	5.108170e-02	188.786881	4	2.381271	41.814583	January	8.0	0.0	0
9	116	6.806017	0.072728	2.039388	4.598857	2.420000e-90	1.663742	3	0.070714	0.404988	...	1.777328	6.074994e-03	479.485597	1	2.853797	74.757185	January	5.0	12.0	0

Figure 9 : Dataset after cleaning

4.3. Data Analysis

The exploratory data analysis (EDA) phase focused on extracting meaningful insights from the dataset:

- Descriptive Statistics: Computed measures such as mean, median, variance, and standard deviation for each parameter to summarize their characteristics.
- Correlation Analysis: Identified strong positive or negative correlations between parameters, aiding feature selection for model development. For instance, high correlation between electrical conductivity and TDS informed their relevance in the predictive model.
- Clustering: Performed unsupervised clustering (e.g., k-means) to group similar water quality samples, providing insights into natural patterns within the data.
- Anomaly Detection: Used algorithms to detect unusual observations, such as abnormally high COD levels, which could indicate pollution events.
- Hypothesis Testing: Conducted statistical tests to verify assumptions, such as whether temperature significantly affects dissolved oxygen levels.

These analyses not only provided a deeper understanding of the dataset but also guided decisions for feature selection and model development, ensuring that the most relevant and impactful variables were used in the predictive models.

	Index	pH	Iron	Nitrate	Chloride	\
count	5.979000e+04	59790.000000	59790.000000	59790.000000	59790.000000	
mean	5.239919e+05	7.455159	0.019393	1.872402	5.149811	
std	3.023917e+05	0.855354	0.027442	0.400410	0.345201	
min	2.000000e+00	3.033252	0.000000	0.351518	3.618146	
25%	2.631190e+05	6.919513	0.000008	1.599718	4.930092	
50%	5.238640e+05	7.459256	0.001981	1.878048	5.169117	
75%	7.858725e+05	8.008094	0.035323	2.145414	5.379465	
max	1.048566e+06	12.245415	0.079159	3.943200	7.067695	

	Lead	Zinc	Color	Turbidity	Fluoride	\
count	5.979000e+04	59790.000000	59790.000000	5.979000e+04	59790.000000	
mean	6.797064e-08	0.783505	3.283944	1.463524e-01	0.598558	
std	2.749762e-07	0.520838	1.378037	1.138340e-01	0.356799	
min	0.000000e+00	0.000003	1.000000	9.559997e-11	0.000025	
25%	8.440000e-123	0.342956	2.000000	3.393717e-02	0.315888	
50%	7.460000e-63	0.729778	3.000000	1.361795e-01	0.570110	
75%	1.867500e-27	1.166013	5.000000	2.470228e-01	0.842334	
max	1.235499e-06	2.916077	5.000000	3.913577e-01	2.523432	

	...	Conductivity	Chlorine	Manganese	Total Dissolved Solids	\
count	...	59790.000000	59790.000000	59790.000000	59790.000000	
mean	...	20.092081	1.787895	0.009870	264.979089	
std	...	4.525991	0.197291	0.016735	154.877198	
min	...	4.832520	1.017634	0.000000	0.020555	
25%	...	16.924792	1.653687	0.000002	131.559070	
50%	...	19.920794	1.788418	0.000550	263.795589	
75%	...	23.051229	1.919156	0.012089	396.325215	
max	...	41.196751	3.037712	0.054062	579.783416	

	Source	Water Temperature	Air Temperature	Day	\
count	59790.000000	59790.000000	59790.000000	59790.000000	
mean	4.492474	2.865586	60.127076	15.970597	
std	2.291003	0.516303	18.083839	8.941568	
min	1.000000	0.821652	-12.087380	1.000000	
25%	2.000000	2.507437	47.961045	8.000000	
50%	4.000000	2.857521	60.132301	16.000000	
75%	6.000000	3.210665	72.212583	24.000000	
max	8.000000	5.009696	137.632506	31.000000	

	Time of Day	Target
count	59790.000000	59790.000000
mean	11.491855	0.231761
std	6.953728	0.421961
min	0.000000	0.000000
25%	5.000000	0.000000
50%	11.000000	0.000000
75%	18.000000	0.000000
max	23.000000	1.000000

Figure 10 : basic information about the data

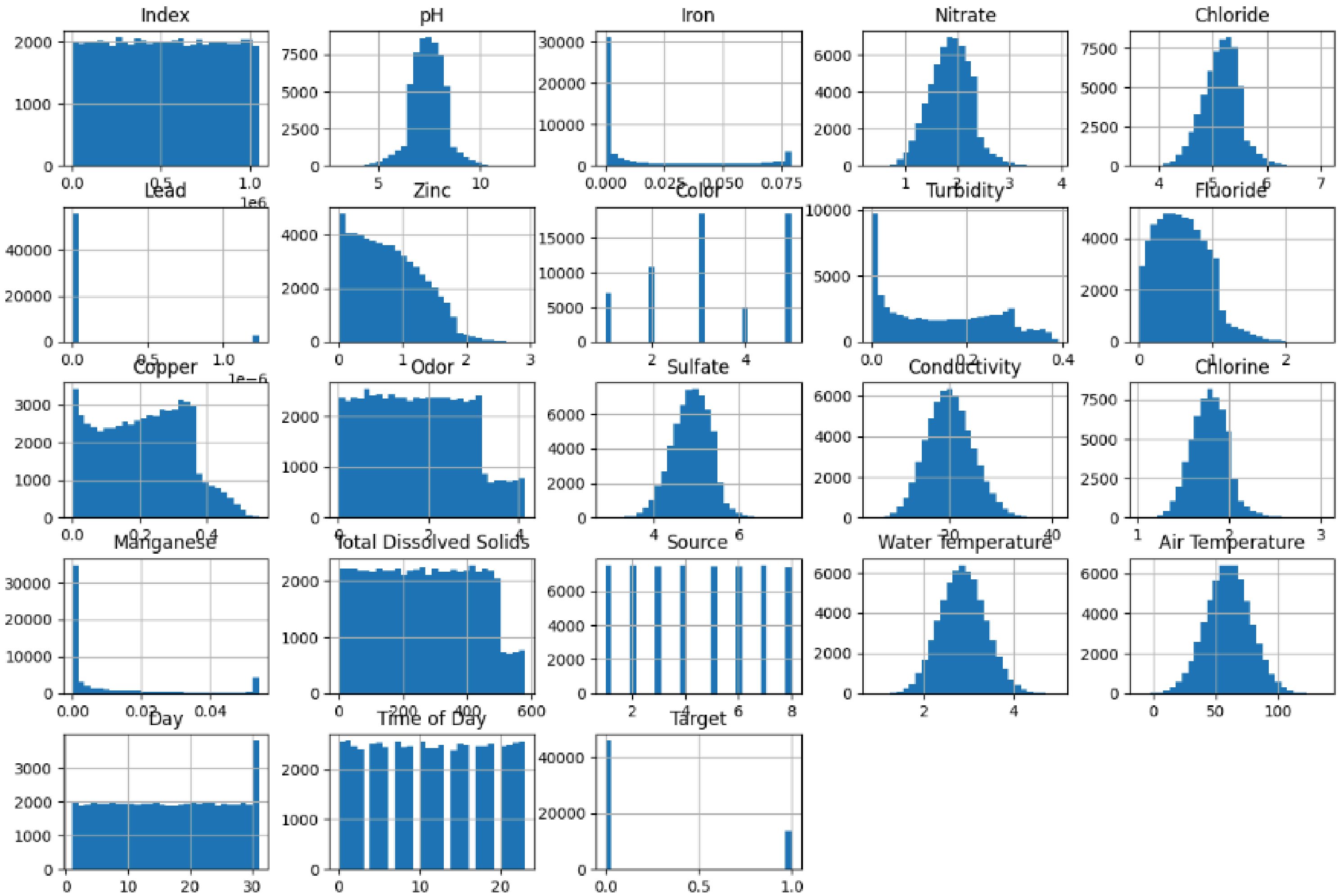

Figure 11 : Histograms for numerical columns

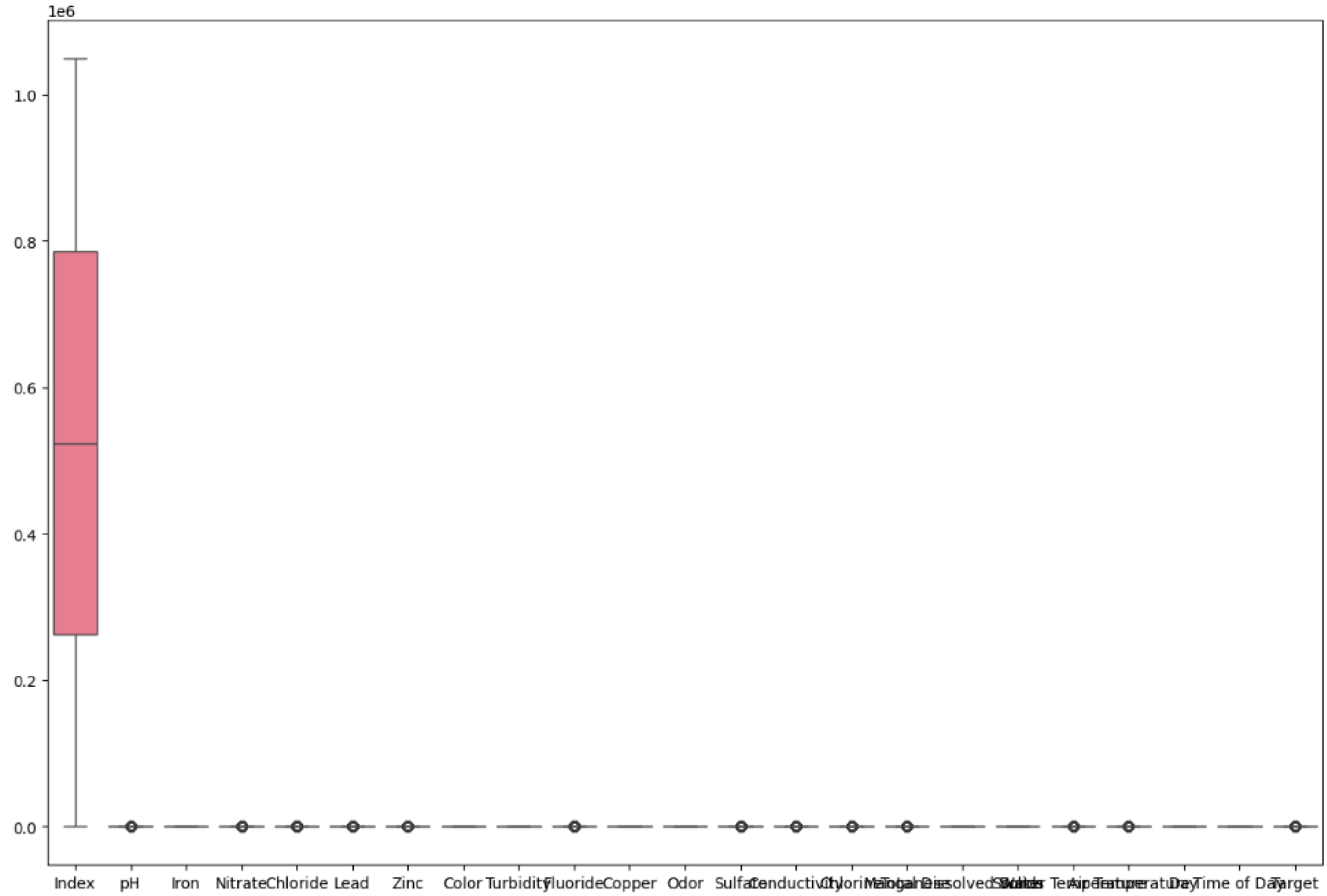

Figure 12 : Box plots for numerical columns

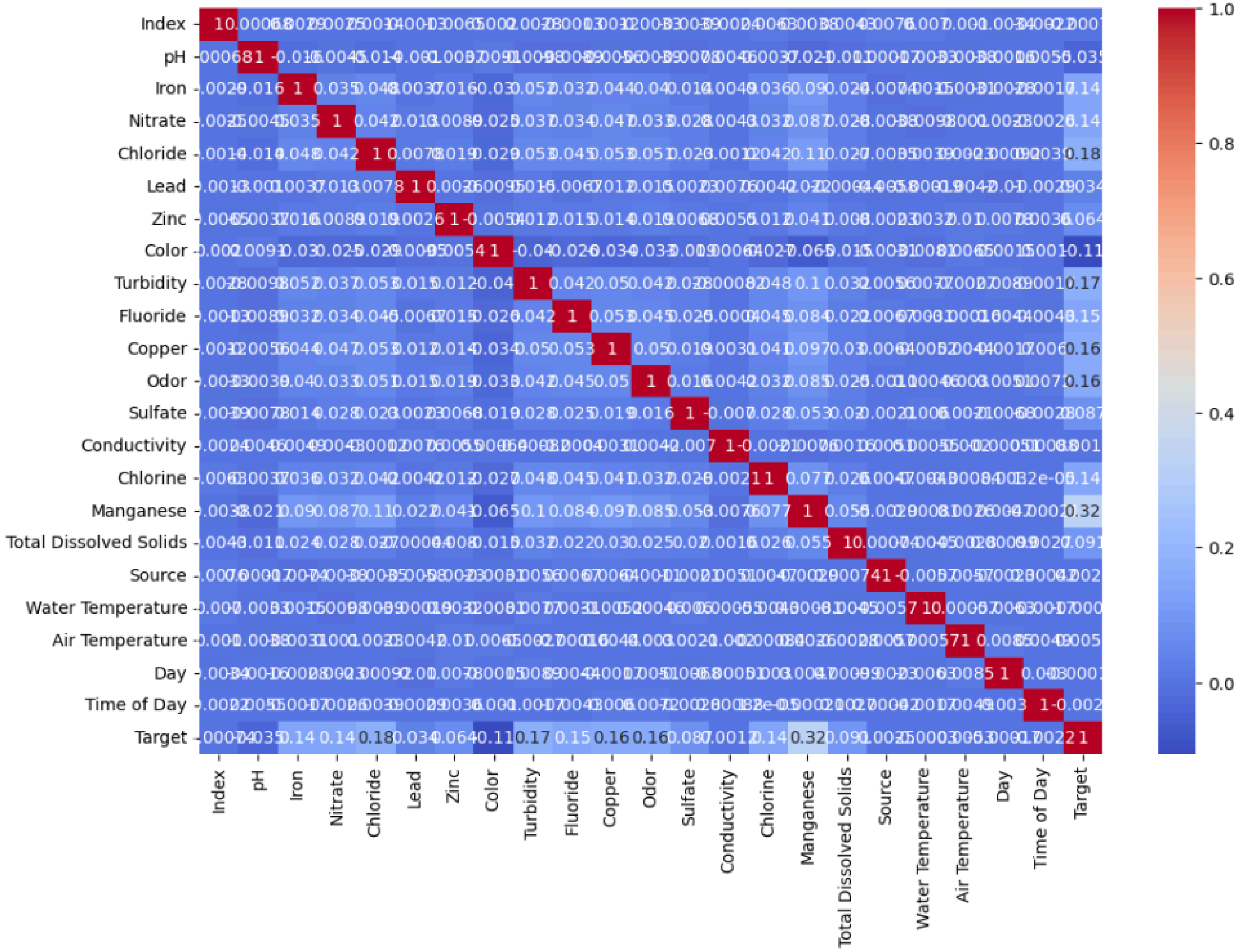

Figure 13 : correlation matrix

5. Model Construction

Model construction involves building and refining predictive models using data science techniques to analyze and predict water quality. This phase includes data processing, training, testing, application of machine learning algorithms, and evaluation of model performance. Each step ensures the development of accurate and reliable models tailored to water quality analysis.

5.1. Data Processing

Data processing is a critical step in model construction as it prepares raw data for analysis. This involves several sub-steps:

1. Data Splitting: The dataset is divided into training and testing subsets. Typically, 70-80% of the data is used for training the model, while the remaining 20-30% is reserved for testing. This ensures that the model is evaluated on unseen data.

2. Feature Selection: Not all variables in a dataset are equally important. Feature selection involves identifying and selecting the most relevant parameters (e.g., pH, dissolved oxygen) to improve model performance. Techniques such as correlation analysis or principal component analysis (PCA) are commonly used.

3. Scaling and Normalization: Many machine learning algorithms require data to be scaled or normalized to work effectively. Standardization ensures all features are on the same scale, preventing bias toward variables with larger values.

4. Handling Missing Data: Missing values are addressed using imputation techniques such as mean, median, or mode substitution, or more advanced methods like K-nearest neighbors (KNN) imputation.

5. Data Augmentation and Transformation: In some cases, synthetic data is generated to balance classes or improve diversity. Transformation techniques, such as logarithmic scaling or box-cox transformations, are applied to address skewness or non-linearities.

Efficient data processing ensures the model's robustness and reliability in predicting water quality under diverse conditions.

	Index	pH	Iron	Nitrate	Chloride	Lead	Zinc	Color	Turbidity	Fluoride	...	Chlorine	Manganese	Total Dissolved Solids	Source	Water Temperature	Air Temperature	Month	Day	Time of Day	Target
0	0	8.332988	8.350000e-05	8.605777	122.799772	3.710000e-52	3.434827	Colorless	0.022683	0.607283	...	3.708178	2.270000e-15	332.118789	NaN	NaN	43.493324	January	29.0	4.0	0
1	2	5.443762	2.010586e-02	3.816994	230.995630	5.290000e-76	0.528280	Light Yellow	0.319956	0.423423	...	3.560224	7.007989e-02	570.054094	River	11.643467	44.891330	January	31.0	8.0	0
2	8	8.238149	8.080000e-10	3.192381	143.222718	1.840000e-57	0.134371	Near Colorless	0.662611	0.316945	...	3.798676	3.508666e-02	436.317937	Spring	69.943048	92.420381	January	5.0	14.0	0
3	42	7.431496	1.635646e-03	1.539861	149.921626	6.450000e-195	2.602858	Colorless	0.477700	1.010053	...	2.667467	8.520000e-09	356.376552	Stream	6.314277	35.875578	January	20.0	20.0	0
4	45	6.618013	2.240000e-07	1.046327	137.400933	2.440000e-81	0.571151	Near Colorless	0.354691	1.241011	...	2.332552	2.292164e-03	383.813214	Ground	17.131867	83.487466	January	4.0	17.0	0
5	52	8.460833	1.986337e-02	8.601511	134.202428	1.170000e-256	2.600728	Faint Yellow	0.249098	0.515965	...	2.494293	5.710000e-10	13.925614	Aquifer	7.611181	82.674304	January	2.0	23.0	0
6	64	8.194406	3.387248e-03	8.344541	248.043661	2.410000e-71	1.993738	Near Colorless	0.019442	0.355717	...	3.781047	2.220000e-05	297.621227	Spring	15.786764	35.653137	January	2.0	19.0	0
7	81	6.735242	8.131881e-01	5.492246	117.293010	5.150000e-26	3.847109	Colorless	0.066189	1.250291	...	2.364011	1.140000e-06	88.336068	NaN	23.234504	67.804340	January	NaN	13.0	0
8	90	5.812626	1.061910e-04	3.032464	199.084282	4.120000e-114	0.951398	Faint Yellow	1.613035	0.029181	...	2.374635	1.848483e-01	188.786881	Ground	9.818649	41.814583	January	8.0	0.0	0
9	116	6.806017	2.458747e-01	6.685902	98.370632	2.420000e-90	4.279026	Near Colorless	0.084110	0.499284	...	3.158896	6.486010e-03	479.485597	River	16.353545	74.757185	January	5.0	12.0	0

Figure 14

```
Data columns (total 24 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Index                   88217 non-null  int64
 1   pH                      86519 non-null  float64
 2   Iron                    87605 non-null  float64
 3   Nitrate                 86639 non-null  float64
 4   Chloride                85602 non-null  float64
 5   Lead                    87812 non-null  float64
 6   Zinc                    85937 non-null  float64
 7   Color                   88136 non-null  object
 8   Turbidity               87464 non-null  float64
 9   Fluoride                85345 non-null  float64
 10  Copper                  85258 non-null  float64
 11  Odor                    85527 non-null  float64
 12  Sulfate                 85338 non-null  float64
 13  Conductivity            85792 non-null  float64
 14  Chlorine                87357 non-null  float64
 15  Manganese               86546 non-null  float64
 16  Total Dissolved Solids  88190 non-null  float64
 17  Source                  86952 non-null  object
 18  Water Temperature       85695 non-null  float64
 19  Air Temperature         87749 non-null  float64
 20  Month                   88217 non-null  object
 21  Day                     86733 non-null  float64
 22  Time of Day             86510 non-null  float64
 23  Target                  88217 non-null  int64
```

Figure 15 : Data Columns

In [155…
```python
# check the nan value
df.isnull().sum()
```

Out[155…
```
Index                       0
pH                       1698
Iron                      612
Nitrate                  1578
Chloride                 2615
Lead                      405
Zinc                     2280
Color                      81
Turbidity                 753
Fluoride                 2872
Copper                   2959
Odor                     2690
Sulfate                  2879
Conductivity             2425
Chlorine                  860
Manganese                1671
Total Dissolved Solids     27
Source                   1265
Water Temperature        2522
Air Temperature           468
Month                       0
Day                      1484
Time of Day              1707
Target                      0
dtype: int64
```

Check some nan value:

	Index	pH	Iron	Nitrate	Chloride	Lead	Zinc	Color	Turbidity	Fluoride	...	Chlorine	Manganese	Total Dissolved Solids	Source	Water Temperature	Air Temperature	Month	Day	Time of Day	Target
44	539	NaN	2.646770e-04	6.746861	246.210476	7.920000e-21	0.626322	Faint Yellow	0.096422	0.380465	...	2.511495	1.350000e-05	24.618199	Stream	15.183871	61.962636	January	16.0	14.0	0
47	552	NaN	6.147040e-04	9.165009	128.712155	2.140000e-190	1.272939	Near Colorless	0.554373	0.933602	...	3.305710	1.479657e-03	34.614726	Aquifer	15.378285	56.873707	January	2.0	4.0	0
90	1172	NaN	6.960000e-07	6.175325	204.131470	4.370000e-96	2.248067	Colorless	0.053603	0.329930	...	2.514912	2.006529e-02	118.763017	Stream	17.199637	60.115803	January	8.0	7.0	0
104	1273	NaN	2.201114e-02	2.733084	204.995334	3.320000e-308	0.407193	Colorless	0.023072	0.483568	...	3.002062	3.990000e-10	300.887085	River	NaN	38.900803	January	2.0	15.0	0
113	1364	NaN	5.865601e-03	7.591204	136.694140	4.620000e-150	3.372179	Faint Yellow	0.617407	0.040993	...	2.442869	8.811430e-01	364.021920	Ground	15.997626	75.401715	January	12.0	14.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
87690	1042732	NaN	6.201853e-02	5.511418	198.785867	3.070000e-67	0.281718	Faint Yellow	0.182003	0.376030	...	2.224731	9.280778e-03	254.392671	Spring	101.013444	42.088909	January	27.0	1.0	0
87699	1042809	NaN	6.440000e-05	7.313949	76.689216	6.970000e-238	0.637428	Faint Yellow	0.055254	0.389151	...	3.088984	3.327640e-04	23.910551	Well	11.883131	46.523553	January	1.0	0.0	0
88004	1046011	NaN	6.249987e-03	4.632648	178.076913	7.580000e-10	4.250372	Colorless	0.012303	0.731769	...	3.131738	4.560000e-10	285.811986	Ground	10.824329	55.961729	January	15.0	6.0	0
88105	1047181	NaN	2.018511e-01	4.568273	135.222494	5.220000e-40	0.000646	Near Colorless	0.201781	NaN	...	2.985349	8.915000e-04	209.925480	Well	53.755586	31.880301	January	15.0	22.0	0
88193	1048360	NaN	1.099714e-03	4.878426	166.353456	1.460000e-12	0.982998	Colorless	0.010817	NaN	...	2.318632	1.119621e-02	189.873480	Well	7.572892	15.871852	January	14.0	10.0	0

Figure 16 : Dataset which pH is nan

5.2. Model Training

Model training is the process of teaching a machine learning algorithm to learn patterns in the data and make predictions.

1. Selecting Algorithms: Various algorithms are chosen based on the nature of the problem (e.g., regression for continuous output, classification for categorical predictions). In this study, algorithms like Linear Regression, Random Forest, Support Vector Machines (SVM), and Gradient Boosting are used. Each algorithm has unique strengths, and their suitability depends on the dataset and prediction goals.

2. Training Process: The model learns from the training dataset by minimizing a loss function (e.g., mean squared error for regression or cross-entropy for classification). The algorithm iteratively adjusts its parameters to improve prediction accuracy.

3. Cross-Validation: To avoid overfitting, cross-validation techniques like k-fold cross-validation are applied. This divides the training data into multiple subsets, trains the model on different combinations of these subsets, and averages the results to enhance reliability.

4. Hyperparameter Tuning: Machine learning algorithms have hyperparameters that control their behavior (e.g., learning rate, tree depth). Techniques like grid search or random search are used to find the best combination of hyperparameters for optimal model performance.

The outcome of the training phase is a model that has learned to capture relationships between input features and the target variable.

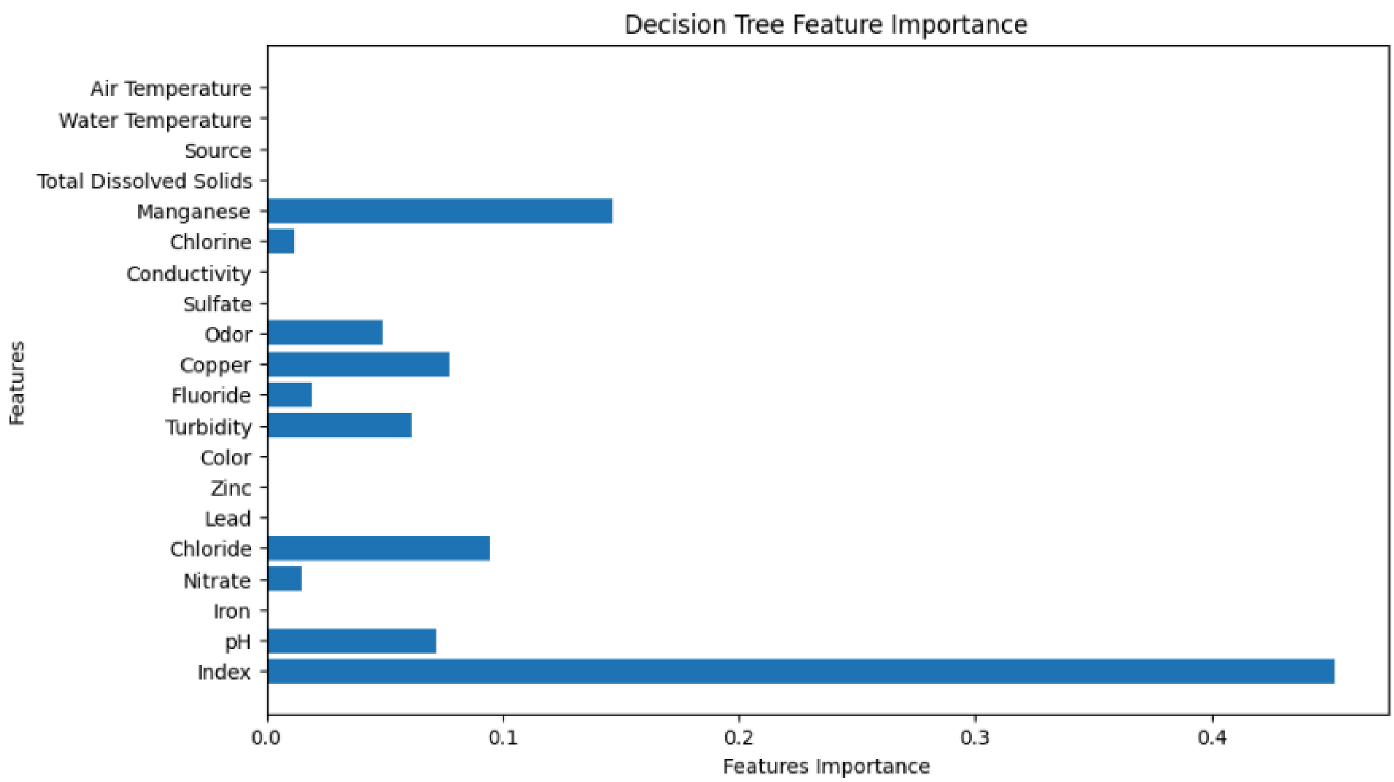

Figure 17 : Plots a horizontal bar graph

5.3. Model Testing

Testing evaluates the model's generalization ability by using the unseen testing dataset. This step determines whether the trained model can predict water quality accurately on new data.

1. Performance Metrics: Specific metrics are used based on the problem type. For regression models, metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 (coefficient of determination) are used. For classification models, accuracy, precision, recall, F1-score, and the confusion matrix are essential metrics.

2. Validation Techniques: Besides the test dataset, further validation is performed to confirm the model's robustness.or example, time-series validation can be applied if water quality data spans multiple time periods.

3. Error Analysis: Residual analysis is conducted to examine the differences between predicted and actual values. Understanding patterns in residuals helps identify areas where the model struggles and suggests improvements.

4. Performance Comparison: Models trained with different algorithms are compared to identify the best-performing one. This comparison is based on evaluation metrics and computational efficiency.

Testing ensures that the selected model is reliable and ready for deployment in real-world water quality prediction tasks.

5.4. Machine Learning

Machine learning is the core of predictive modeling in this study. It involves applying algorithms to learn patterns in data and generate predictions.

1. Supervised Learning: This study uses supervised learning methods, where the algorithm learns from labeled data. Regression algorithms like Linear Regression and Random Forest predict continuous outputs like water quality index, while classification algorithms such as SVM classify water into categories (e.g., "Good" or "Poor").

2. Ensemble Methods: Ensemble techniques, such as bagging (Random Forest) and boosting (Gradient Boosting Machines), are employed to improve accuracy. These methods combine predictions from multiple models to reduce bias and variance.

3. Feature Importance Analysis: Machine learning models like Random Forest provide insights into feature importance. This helps identify which water quality parameters contribute most to predictions, aiding in environmental decision-making.

4. Automation and Scalability: Machine learning pipelines are automated to handle large-scale datasets and enable real-time water quality predictions.

Machine learning provides a data-driven approach to water quality analysis, ensuring high accuracy and adaptability to various datasets.

```
Predicting the target for the new month data...
          pH      Iron   Nitrate  Chloride          Lead      Zinc     Color  \
0   7.329538  0.021167  2.065913  4.977253 -1.055367e-07  0.186854  4.372696
1   7.747258 -0.008455  2.162884  5.215863  2.931628e-07  0.065872  1.163674
2   8.605573  0.035558  1.914790  5.213796  5.005224e-07  1.400899  4.050580
3   7.179460  0.042511  2.186888  5.009125  5.405314e-08  1.324345  5.442422
4   6.579694  0.033453  1.799624  4.850250 -1.615790e-07 -0.144420  2.314280
5   6.691984  0.075635  1.575714  5.345413  3.633887e-07  1.084029  1.241913
6   5.638839  0.059978  1.884620  5.174778  1.929610e-07  0.038647  2.682642
7   8.176219  0.031951  2.105315  4.836622  2.088801e-07  1.581267  7.270518
8   7.672312  0.045175  1.727286  4.992094  1.150691e-07  0.571445  2.383046
9   7.209331 -0.048339  2.201826  5.083146  4.611712e-07  0.843877  5.665090
10  6.488864 -0.007191  1.953339  5.276846  3.077781e-07  1.306531  3.122901
11  7.011117  0.059479  1.670707  5.289965  2.671056e-07 -0.021895  3.065808
12  7.158199  0.011554  1.611675  5.495662  3.397956e-07  0.329131  4.409188
13  7.537428  0.010365  1.217873  4.737798 -1.086653e-07  0.618734  2.514274
14  7.493715  0.045844  2.194315  4.540424  1.277831e-07  0.802110  3.971759
15  7.289507  0.067896  2.275929  4.598263  4.780153e-07  1.400779  2.816273
16  4.597935  0.032872  2.087350  4.847013 -3.788955e-08  0.601707  4.452540
17  7.456759  0.037775  2.366636  4.691884  3.735397e-07  0.052595  2.301431
18  7.410813  0.045170  1.790767  5.880710  1.145126e-07 -0.081752  4.891877
19  7.328976  0.046451  1.838686  5.340597  3.401261e-07  0.020238 -0.232412
20  8.323205 -0.029813  3.049416  4.920702 -1.895551e-07  1.483391  4.097288
21  7.490388  0.048460  1.758187  4.889653  2.732736e-07 -0.352386  3.847570
22  8.324307  0.014546  2.639643  4.784100 -1.096413e-07  1.458175  5.562413
23  7.316484  0.018340  1.461658  4.685873 -8.654850e-08  1.181426  3.304738
24  6.855199  0.041606  1.322461  5.119171  1.856452e-07  1.337921  2.258750
25  6.984291 -0.019344  2.185354  4.848796 -1.168300e-07  0.838854  3.408369
26  8.475107  0.054451  1.603857  4.366107  1.105043e-08  0.855152  2.224943
27  7.529108  0.003312  2.613267  5.145635 -1.404622e-07 -0.099062  4.417614
28  7.262705  0.048031  1.950629  4.683104 -6.256238e-08  0.816237  2.201096
29  5.469860  0.002722  1.230072  4.989899  1.436659e-07  0.484071  0.956360
```

```
    Turbidity  Fluoride    Copper      Odor   Sulfate  Conductivity  Chlorine  \
0    0.044158  0.532573  0.383361  2.212668  5.110192     21.033366  1.518004
1    0.258458  0.429122  0.180023  1.245483  4.659790     22.457444  1.523409
2    0.083339  0.910214  0.303086  2.793179  4.530572     23.443174  1.672881
3    0.161169  0.847468  0.294598  0.727816  4.255892     17.296662  2.002069
4    0.274161  0.266269  0.069740  0.554480  4.667834     18.343283  1.081582
5    0.284023  0.581769 -0.011566  0.944871  3.848313     26.080940  1.710207
6    0.088035  0.652408  0.298197  0.539806  5.628135     25.382178  1.528400
7    0.063954  0.813996  0.044855  2.229096  5.154011     17.093170  1.787237
8    0.098146  0.697084  0.270903  3.150785  5.162471     23.981370  1.879582
9   -0.016376  0.525715  0.278591  0.928429  4.372977     20.766697  1.795392
10   0.345454 -0.149390  0.297389 -0.051700  5.039304     14.325870  2.161527
11   0.108624  1.208175  0.091582  1.433305  3.605683     17.062961  1.781414
12   0.035077 -0.435851  0.106597  1.579329  5.128576     20.929343  1.846931
13   0.281497  0.104971  0.270210  1.624403  4.752968     22.374008  1.612636
14   0.294679  0.264650  0.178627  1.222465  4.631133     18.169657  1.599957
15   0.066890  0.916500  0.269710  1.191395  4.592886     21.930780  1.283703
16   0.065775  0.352178  0.363703  1.534526  5.362064     22.443859  1.458041
17   0.194271  1.142331  0.037001  3.214221  4.593049     16.657800  1.628482
18   0.102492  0.616078  0.105450  3.114328  5.336734     14.069754  1.451762
19   0.023520  0.852363  0.326415  1.327629  4.209083     13.303246  1.916277
20   0.198843  0.556964  0.201333  4.413245  4.618052     18.799684  1.841577
21   0.037108  0.830339  0.012274  2.371143  4.607771     22.331850  1.839683
22   0.168017 -0.158562 -0.067099 -0.281892  4.573467     20.171576  1.894339
23   0.091837  0.341623 -0.045210  0.900120  4.409455     16.280457  2.000810
24   0.018207  1.080260  0.355767  1.275966  4.440412     21.938094  2.021422
25   0.282097  0.362702  0.488139  2.621407  3.913594     20.962141  1.817751
26   0.072899  0.664912 -0.072343 -0.437712  4.854889     22.750826  1.916604
27   0.027964  0.847332  0.031109  2.602789  4.580647     15.965067  1.888593
28  -0.054790  1.036242  0.296103  1.845649  5.227032     15.633298  1.807982
29   0.312651  1.145588  0.334291  1.901653  4.654860     12.383323  2.324683
```

	Manganese	Total Dissolved Solids	Source	Water Temperature	\
0	0.014870	273.638979	2.221579	3.187161	
1	0.025762	-15.766899	4.663350	3.783950	
2	0.008044	276.746809	7.934214	3.308583	
3	-0.018941	181.732187	4.694891	2.051466	
4	0.007291	531.781986	1.355708	3.050108	
5	0.005532	494.955927	3.624769	3.365758	
6	0.021109	248.732091	2.727411	2.850369	
7	0.025437	251.091318	5.641856	2.445164	
8	0.028010	334.222106	4.881238	3.256329	
9	0.041576	285.907543	5.444005	2.175464	
10	0.011492	124.247000	4.114588	2.247313	
11	-0.005141	279.106278	7.561107	2.738171	
12	-0.022570	35.700922	5.055104	1.969076	
13	0.013567	388.298226	2.878621	3.043784	
14	0.003309	259.851854	8.405734	2.480175	
15	-0.012764	189.815426	2.695836	2.564322	
16	0.005022	344.790647	4.095351	2.683886	
17	0.028942	496.709153	6.219920	2.985212	
18	0.026115	271.122253	2.618941	3.020622	
19	0.016090	417.539225	7.328787	3.496863	
20	-0.012226	163.033497	-1.214960	3.305435	
21	0.010846	136.549751	4.289706	2.878510	
22	-0.000126	95.829844	4.337756	2.378438	
23	-0.001763	400.329345	4.692570	2.189514	
24	-0.008930	167.750655	6.963857	1.732084	
25	0.013864	286.139634	8.073558	2.572162	
26	0.006430	189.847502	0.599837	2.890450	
27	0.004289	322.123093	2.803037	2.074594	
28	0.014881	59.158268	4.639728	2.641932	
29	-0.000063	218.180490	4.722882	2.642164	

	Air Temperature
0	49.874246
1	73.435764
2	29.516185
3	76.639601
4	62.583076
5	69.054484
6	56.516296
7	48.729888
8	44.075083
9	62.022947
10	53.190157
11	44.389364
12	75.690568
13	58.792133
14	66.414541
15	64.659051
16	67.309838
17	61.795331
18	51.528296
19	76.297465
20	50.543071
21	79.651686
22	63.597191
23	52.018473
24	90.803642
25	114.418138
26	80.002526
27	53.673787
28	57.759253
29	81.812122

Figure 18 : Displaying the New Data

	pH	Iron	Nitrate	Chloride	Lead	Zinc	Color	Turbidity	Fluoride	Copper	Odor	Sulfate	Conductivity	Chlorine	Manganese	Total Dissolved Solids	Source	Water Temperature	Air Temperature	Predicted Target
0	6.172770	0.050115	1.994325	5.017845	-4.994972e-07	0.635716	1.278707	0.099157	0.675492	0.252603	-0.000316	4.914718	20.373981	1.611000	0.008701	400.986692	7.372163	2.619794	37.717790	0
1	8.185389	-0.040758	2.129893	5.265945	2.516840e-08	1.319454	1.400266	0.156529	0.890780	0.278848	2.489559	5.041375	19.462671	1.753483	-0.005637	57.964403	-2.034349	2.677878	84.730976	0
2	8.542531	0.025468	2.252521	4.815399	-2.089911e-07	0.543540	2.070284	-0.064624	0.655678	0.287259	2.408626	5.391995	16.404845	2.329143	0.019376	209.819785	1.289127	3.967433	78.385414	1
3	8.027166	0.027773	1.968500	4.907549	5.732093e-07	0.780615	3.667299	0.078989	0.747390	0.160093	3.080259	5.111376	18.512175	1.826820	0.026701	355.421751	5.047142	2.931554	69.444371	0
4	7.869796	0.067055	1.852866	4.942604	-2.962675e-07	-0.035825	3.966049	0.109183	0.318674	0.320224	2.212657	4.304243	21.645689	2.019997	0.017855	70.654692	4.290045	2.195231	10.378782	0
5	7.687248	0.018320	1.993890	5.781921	-7.161772e-09	0.257288	1.650163	0.138154	0.407795	0.542315	2.386122	4.773189	18.725670	1.898175	0.005094	220.656202	4.715581	3.071810	46.896784	1
6	10.060557	-0.055163	1.682824	5.466781	-5.250373e-07	1.180413	4.879849	0.035165	0.755269	0.306328	1.219383	5.261522	21.231607	1.789674	0.029904	281.714572	5.872666	2.393224	42.799413	0
7	8.183759	0.047970	1.733823	5.119529	-1.912436e-07	0.380486	1.318070	0.224618	0.894833	0.295133	1.646735	5.241092	25.760437	1.883400	0.013119	229.164281	0.501719	3.053866	66.832112	0
8	9.209680	-0.037171	2.137397	4.512147	-5.140652e-07	0.566464	6.514960	0.258920	-0.050884	0.329331	2.166568	4.906521	23.439245	1.536864	-0.007586	20.978795	3.114487	2.550955	74.162086	0
9	7.859326	0.003642	2.188217	4.711001	4.285103e-07	0.615519	1.854264	0.195604	0.885716	0.224465	2.606576	4.735712	21.701756	1.914488	0.022719	276.022645	8.502138	3.186996	63.193734	0
10	7.046369	0.000486	1.858472	4.876597	5.996219e-07	0.948771	3.124214	-0.020816	0.492182	0.333083	2.142339	5.092510	26.202761	1.808000	-0.005637	370.733944	6.080376	3.099519	55.896502	0
11	7.117707	-0.025711	2.408986	4.920223	-2.114029e-07	0.562991	3.905986	0.192516	0.223151	0.232422	2.690632	5.115326	12.223254	1.587115	0.029398	430.689520	5.207800	1.868056	100.598699	0
12	6.412641	0.020601	1.574162	5.127554	6.482727e-07	1.039493	0.792264	0.148164	0.594377	0.004706	2.102389	5.315464	18.803190	1.876049	0.005565	451.284443	5.864067	2.878633	62.385460	0
13	7.761001	0.044632	1.502919	4.964056	4.515996e-07	0.799340	5.134070	0.232529	0.114743	0.369616	2.362777	5.390887	12.592653	1.531827	0.005030	267.870374	4.223682	3.324341	75.803301	0
14	6.845138	-0.009893	1.979185	4.535525	-9.278104e-08	0.441826	4.135461	0.298890	0.672517	0.068148	1.955174	4.427541	27.151883	1.832446	0.007566	202.745012	6.895079	3.478641	71.624838	0
15	7.851657	0.050551	1.090673	5.902577	-1.751142e-07	-0.090308	1.722014	0.030406	0.935163	0.158985	1.705262	4.840308	23.880039	1.962298	0.020055	139.444821	2.038333	3.090687	73.333223	1
16	7.970981	-0.001184	1.787153	5.047810	3.742905e-07	1.428745	6.602196	0.059905	0.453740	0.182355	2.787602	5.566288	25.772894	1.613412	0.018275	176.528900	7.065982	2.438694	60.555784	0
17	6.767661	0.038754	2.533192	5.365118	-2.059815e-07	1.408152	4.717713	0.217386	-0.385189	0.303552	3.132152	5.258481	13.640108	1.703464	0.027002	224.368409	2.247448	2.985972	61.020016	0
18	7.762365	0.045272	1.610907	5.297271	-8.388691e-08	1.769743	1.502301	0.318451	0.874940	0.096210	0.595058	5.354582	19.388384	1.953072	0.003980	-17.187048	2.212128	3.183470	73.140503	1
19	7.987629	0.027219	1.390314	5.226253	-8.390117e-08	1.822720	0.748347	0.250331	0.347489	0.256326	1.471350	4.777926	24.846810	1.615655	0.005828	338.677093	3.219181	3.083728	28.410044	0
20	6.533538	0.013143	2.231198	4.957926	9.933706e-08	0.758243	4.599285	0.273131	1.079813	0.109767	4.017265	4.847919	24.327498	1.907834	-0.004346	203.326155	5.176640	2.904177	42.266594	1
21	8.353991	0.093543	1.950719	5.640326	3.068292e-07	0.777521	3.139857	0.056705	1.164134	0.036496	0.474614	5.001397	26.475240	1.739220	-0.015418	274.464145	4.630166	2.621045	52.146833	1
22	6.981950	-0.019230	1.522072	5.540824	2.884941e-07	0.926798	3.058409	0.180790	0.301838	0.367994	0.610328	5.835721	13.723470	1.845801	0.004220	569.896566	4.047039	2.162038	53.131305	1
23	7.359258	0.033566	2.334943	5.383299	1.874116e-07	1.743016	4.021887	0.172474	0.108023	-0.030251	5.221995	5.390808	14.325149	1.977790	-0.008788	-17.931008	5.296344	3.701171	52.873657	0
24	6.322380	0.026218	2.170647	4.932863	-1.005405e-07	-0.420553	0.859056	0.118167	0.371736	0.215692	1.005118	4.957023	14.106010	1.958344	-0.001836	307.947427	6.330171	2.306516	61.880813	0
25	7.686785	0.041125	1.850124	5.251268	-1.131986e-07	0.435614	3.942333	0.405221	-0.145110	0.244046	1.030134	5.126551	19.384582	1.634078	-0.016866	384.869049	4.780060	3.387334	66.617695	1
26	8.510677	0.034571	2.643820	5.615187	-2.349435e-07	-0.045649	1.707776	0.100473	1.357379	0.276293	0.196782	5.218704	22.975401	1.400386	-0.038231	239.733246	5.021514	2.980890	65.810603	1
27	6.670273	0.006288	2.046147	5.174136	-2.089892e-07	1.032193	2.577593	0.289822	0.778737	0.346885	2.095289	4.565374	27.806196	2.053523	0.000187	121.104740	1.754324	2.777703	55.639166	0
28	8.318617	0.010648	2.119575	4.856944	-2.046859e-07	-0.004479	4.541811	-0.007332	-0.045696	0.133626	1.293101	5.376332	10.498588	1.992623	0.007333	270.703057	6.094681	2.478441	70.420659	0
29	6.807367	-0.010174	1.209353	5.001011	3.643504e-07	0.425672	4.309512	0.229247	0.351880	0.155411	1.472485	4.761928	14.444584	1.903422	0.025577	65.496974	6.394846	3.017247	81.577329	0

Figure 19 : Predicted Data

5.5. Evaluation

Model evaluation is the final step in model construction, where the effectiveness of the trained model is assessed using performance metrics.

1. Metric-Based Evaluation: Metrics like R^2, MSE, accuracy, precision, recall, and F1-score are analyzed to evaluate the model's predictive power. For multi-class classification tasks, the confusion matrix and ROC-AUC score are also considered.

2. Model Comparison: Multiple models are evaluated and compared to select the best-performing one. This involves analyzing trade-offs between accuracy, computational efficiency, and interpretability.

3. Error Diagnostics: Evaluation includes diagnosing areas where the model underperforms. For example, high residuals in specific ranges of dissolved oxygen may indicate model bias.

4. Generalization Check: The model's ability to generalize is tested using datasets from different sources or regions to ensure robustness across diverse scenarios.

5. Insights for Improvement: Evaluation results provide insights into areas for improvement, such as revisiting feature selection, adjusting hyperparameters, or trying advanced algorithms like deep learning.

Effective evaluation ensures that the model is reliable and ready for practical applications in water quality prediction.

6. Conclusion

This report highlights the application of data science in predicting water quality, showcasing how exploratory analysis, data processing, and machine learning models can solve real-world environmental problems. Key findings include the importance of parameters such as pH and dissolved oxygen in determining water quality and the effectiveness of machine learning algorithms like Random Forest and SVM in achieving accurate predictions. The study underscores the value of data-driven approaches in environmental monitoring and decision-making. Future work could involve integrating real-time monitoring systems, using advanced models like neural networks for complex relationships, and expanding the dataset to include more regions and time periods. These improvements would further enhance the reliability and applicability of water quality prediction models. Data science provides a powerful toolkit for addressing environmental challenges, promoting sustainability, and ensuring safe water resources for communities and ecosystems.

7. References

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

Code link: https://github.com/chantharith-NY/data-science-project/tree/main/data/Raw%20data

ML Testing :
https://github.com/chantharith-NY/data-science-project/blob/main/notebooks/ML%20Testing/ML_Testing_Accuracy.ipynb

https://github.com/chantharith-NY/data-science-project/blob/main/notebooks/ML%20Testing/ML_Testing_Method.ipynb

Model Training :

https://github.com/chantharith-NY/data-science-project/blob/main/notebooks/Model%20Training/KNN_Training.ipynb

https://github.com/chantharith-NY/data-science-project/blob/main/notebooks/Model%20Training/Model_Training.ipynb

https://github.com/chantharith-NY/data-science-project/blob/main/notebooks/Model%20Training/Random_Forest_Training.ipynb

Data Cleaning :
https://github.com/chantharith-NY/data-science-project/blob/main/notebooks/data%20cleaning/data_cleaning_January.ipynb

Machine Learning :
https://github.com/chantharith-NY/data-science-project/blob/main/notebooks/machine%20learning/Predicting_using_model.ipynb