## ABSTRACT

The major goal of this project is to use machine learning techniques to measure water quality. A potability is a numerical phrase that is used to assess the quality of a body of water. The following water quality parameters were utilised to assess the overall water quality in terms of potability in this study. ph, Hardness, Solids, Chloramine’s, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, Turbidity were the parameters. To depict the water quality, these parameters are used as a feature vector. To estimate the water quality class, the paper used two types of classification algorithms: Decision Tree (DT) and K- Nearest Neighbour (KNN). Experiments were carried out utilising a real dataset containing information from various locations around Andhra Pradesh, as well as a synthetic dataset generated at random using parameters. Based on the results of two different types of classifiers, it was discovered that the KNN classifier outperforms other classifiers. According to the findings, machine learning approaches are capable of accurately predicting the potability. Potability, Water Quality Parameters, Data Mining, and Classification are all index terms.

***Keywords***: Machine Learning, Supervised Learning, K-Nearest Neighbour (KNN), Decision Tree, Hyper Parameter Tuning, Python Programming.

## 

|  |  |  |
| --- | --- | --- |
|  | TABLE OF CONTENTS |  |
| **CHAPTER NO.** | **TITLE** | **PAGE NO** | |
|  | ABSTRACT  LIST OF FIGURES  LIST OF TABLES | Vi  Viii  ix | |
|  |  |  | |
|  |  |  | |
|  | |
|  |  |
| **1** | **INTRODUCTION** | 1 | |
| **2** | **AIM AND SCOPE**  **2.1** scope | 2 | |
|  | **2.2** data collection and creation | 2 | |
|  | **2.3** dataset created artificially | 2 | |

**2.4** set of real data 3

**2.5** scope 4

**3 FLOW CHART**

**4 EXPERIMENTAL OR MATERIALS AND METHODS**

**ALGORITHMSUSED** 5

|  |  |
| --- | --- |
|  | **4.1** methodology 6 |
|  | **4.2** classification 6 |
| **5** | **RESULTS AND DISCUSSION, PERFORMANCE** 7  **ANALYSIS** |
|  | **5.1** outputs 9 |
|  | **5.2** potability 10  **5.3** performance measures results 11 |
| **6** | **ENVIRONMENT INVESTIGATION** 12 |
| **7** | **ADVANTAGES AND DISADVANTAGES** 13 |
| **8** | **7.1** FUTURE SCOPE 14    **CONCLUSIONS** 15  **8.1**Screenshots |
|  | **8.2**SOURCE CODE 16 |

## LIST OF FIGURES

**FIGURE NO. FIGURE NAME PAGE NO.**

* 1. Flow Chart for Water Quality Prediction 4
  2. Potability Counts of Dataset 7
  3. Histogram 7
  4. Correlation Heatmap 8
  5. Finding Outliers Using Boxplot 8
  6. Scatter Plot of Hardness and Solids 9
  7. Scatter Plot of PH and potability 9

## CHAPTER 1

## INTRODUCTION

Water quality analysis is a complex topic due to the different factors that influence it. This concept is inextricably linked to the various purposes for which water is used. Different needs necessitate different standards. There is a lot of study being done on water quality prediction. Water quality is normally determined by a set of physical and chemical parameters that are closely related to the water's intended usage. The acceptable and unacceptable values for each variable must then be established. Water that meets the predetermined parameters for a specific application is considered appropriate for that application. If the water does not fulfil these requirements, it must be treated before it may be used. Water quality can be assessed using a variety of physical and chemical properties.As a result, studying the behaviour of each individual variable independently is not possible in practise to accurately describe water quality on a spatial or temporal basis. The more challenging method is to combine the values of a group of physical and chemical variables into a single value . A quality value function (usually linear) represented the equivalence between the variable and its quality level was included in the index for each variable. These functions were created using direct measurements of a substance's concentration or the value of a physical variable derived from water sample studies. The major goal of this research is to examine how machine learning algorithms may be used to predict water quality.

## CHAPTER 2

## AIM and SCOPE

**2.1 Aim**

A Water Quality Index (WQI) is a means by which water quality data is summarized for reporting to the public in a consistent manner. It is similar to the UV index or an air quality index, and it tells us, in simple terms, what the quality of drinking water is from a drinking water supply.

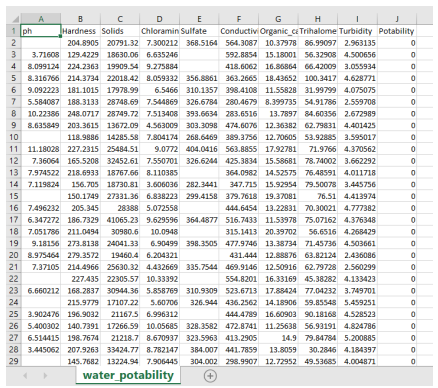
WQI scores are computed for each public water supply system that has been sampled in a sampling season. The same variables are used in the computation of the WQI for all public water supply systems and only the six most recent samples are used. However if a public water supply system is on a Boil Water Order, or it has a current contaminant exceedance, or has a THMs average above the drinking water quality guideline a WQI score is not computed.

**2.2 Data collection and creation**

Data mining techniques require domain knowledge in order to generate predictions. For water quality applications, it is vital to understand how various water quality parameters influence water quality. This information can come from a domain expert or historical data collections. For the forecasting task, two types of data sets were used: a carefully created huge synthetic data set and an available real data set. The fact that both data sets are examined on an equal number of indicator parameters is the key similarity between them. The amount of samples considered in each data set differs amongst the data sets. The number of observations in the real data set was limited. Due to the lack of big genuine data sets, a synthetic data collection was produced. The developed synthetic data set, on the other hand, captures identical relational structures and water quality parameters have the same distribution as in the real-world scenario. Ten water quality parameters were utilised to evaluate the overall water quality in terms of potability for each data set. These variables are pH and Hardness. Solids, chloramines, sulphate, conductivity, organic carbon, trihalomethanes, turbidity, and potability are all terms that can be used to describe something. The choice of parameters was influenced by the fact that they are all commonly monitored critical parameters with well-defined water quality standards. The predictive modeling described in this paper, on the other hand, is flexible enough to function with any number of parameters

**2.3 Data set created artificially**

A target data set is necessary for the use of data mining methods. If data mining is to be used to find patterns in data, the data collection should be large enough to contain these patterns as a general rule. A synthetic data collection was created to provide a realistic technique to obtaining this enormous data set. This synthetic data set was carefully produced by taking into account possible water quality parameter ranges. The benefit of using these concentration ranges was that they were developed after careful consideration of water quality standards assigned by various national and international organization’s such as the European Union (EU), the World Health Organization (WHO), the Central Pollution Control Board (CPCB), and others Scientific data was reported. Each sample reflected the occurrence of one instance of the 10 parameter concentration values under investigation. The data set that will be utilised to develop a predictive model using the classification technique must be supervised. The following step was to establish a supervised environment for the numerical data set, which was generated by assigning a label to each instance in order to forecast the water contamination level. To do this, the potability was determined for each instance of concentration values for the 10 parameters chosen.

****

DATA TO PREDICT WATER QUALITY

**2.4 Set of real data**

To analyze overall water quality in terms of potability, ten water quality factors were used for each data set. pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic carbon, Trihalomethanes, Turbidity, and Potability were among the metrics studied. The choice of parameters was influenced by the fact that they are all commonly monitored critical parameters with well-defined water quality standards. The predictive modelling described in this paper, on the other hand, is adaptable enough to function with any number of parameters.

**2.5 Scope**

During the last years, water quality has been threatened by various pollutants. Therefore, modeling and predicting water quality have become very important in **controlling water pollution**



Import

Libraries

Collect

and

Import

Dataset

Data

Pre

-

processing

Train

model

Test

and

Evaluate

model

Save Model



UI-Input



Saved Model

Prediction



UI-Output

3.

FLOWCHART



## CHAPTER 4

## EXPERIMENTAL OR MATERIALS AND METHODS;

## ALGORITHMS USED

**4.1 METHODOLOGY**

The proposed system is intended to determine potability. It is divided into two phases, one for training and the other for testing. The following procedures are carried out in both sections. Data on training pH and hardness testing data Solids, chloramines, sulphate, conductivity, organic carbon, trihalomethanes, turbidity, and potability are all terms that can be used to describe something. The data set was chosen as follows: The collection of essential parameters that affect water quality, identification of the number of data samples, and definition of the class labels for each data sample present in the data are all factors that go into selecting the water quality data set, which is a prerequisite to model construction. Ten indicator parameters make up the data sets used in this study. pH value and hardness are examples of these factors. Solids, chloramines, sulphate, conductivity, organic carbon, trihalomethanes, turbidity, and potability are all terms that can be used to describe the properties of a substance. The proposed approach, however, is not constrained by the number of parameters or the selection of parameters. A k-fold cross validation technique is employed to set the learning and testing framework in this study, corresponding to each data sample in the data set. The dataset is separated into k-disjointed sets of equal size, each with roughly the same class distribution, using this technique. This division's subsets are utilised as the test set in turn, with the remaining subsets serving as the training set. These are Decision Tree (DT) and K-Nearest Neighbour (KNN) methods. In terms of the underlying relational structure between the indicator parameters and the class label, each of these strategies takes a different approach. As a result, each technique's performance for the same data set is likely to differ. Validating the performance of different classifiers on an unknown data set: Data mining provides several metrics for validating the performance of different classifiers on an unknown data set. A repeated cross-validation procedure in the Matlab caret package was used to create the learning and testing environment. The following procedure was used to apply the classification algorithm: 1. The data set was split into two parts: training (80%) and testing (20%). (20 percent ). 2. The training set was subjected to repeated cross-validation, with the number of iterations fixed to Classifiers were trained in this manner. 3. The model's optimal parameter configuration was selected, resulting in the maximum accuracy. 4. The model was scrutinized.

**4.2 CLASSIFICATION**

To estimate river water quality class, two data mining methods were used: Decision Tree(DT) and K- Nearest Neighbour(KNN). These methods are both parametric and nonparametric classifiers, and their goal is to develop a function that maps input variables to output variables from a training dataset. Because the function's form is unknown, different algorithms make different assumptions about the function's form and how training data is learned to produce the output. The parametric learning classifier makes more confident assumptions about the data. If the assumptions for any data set are true, these classifiers will make rectification judgments. However, if the assumptions are incorrect, the same classifier performs poorly. In order to learn classification tasks, these classifiers do not rely on the quantity of the sample data set; rather, their working principles are their assumptions. This classifier is susceptible to prediction mistakes such as bias, in addition to its parametric character. When the model makes multiple assumptions, the Decision Tree yields substantial bias. Nonparametric classifiers, unlike parametric learning classifiers, do not make any assumptions about the form of the mapping function, and by not making any assumptions, they are having more accuracy. These classifiers can create any function from the training data set. The DT and KNN classifiers are included in this category. Learning techniques are used in DT, whereas the similarity principle is used in KNN. To put it another way, DT Small data sets with complete domain expertise, on the other hand, are equally advantageous for these classifiers. Instead of learning from data, the KNN classifier finds a group of k items in the training set that are the most similar to the test object. Unlike other classifiers, DT does not rely on domain expertise. To make classification decisions, it simply calculates the distance between two characteristics. Because each algorithm's mode of operation differs, a comparison of all of them is necessary to determine which one is better at approximating the underlying function for the same training and testing water quality datasets.

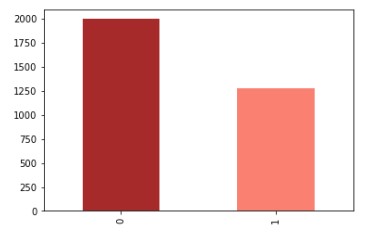
## 

## 

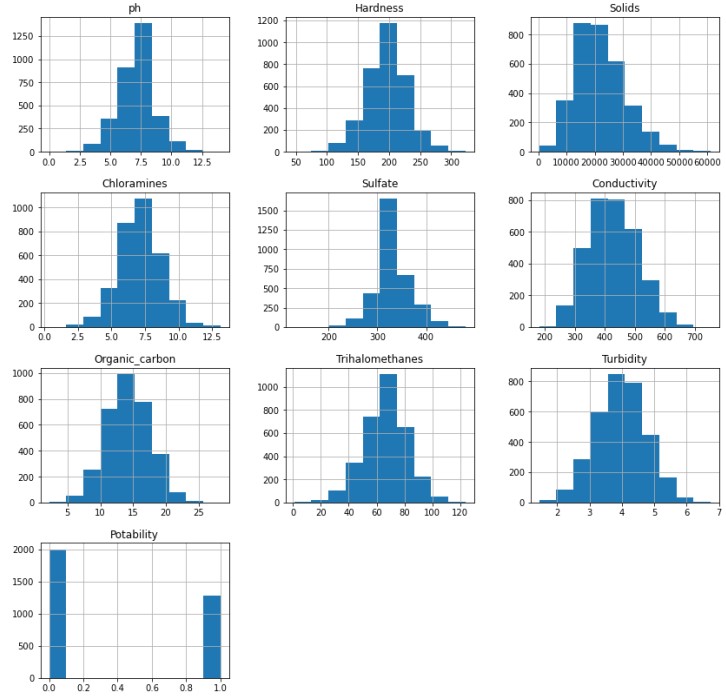
## CHAPTER 5

## RESULTS, DISCUSSION and PERFORMANCE ANALYSIS

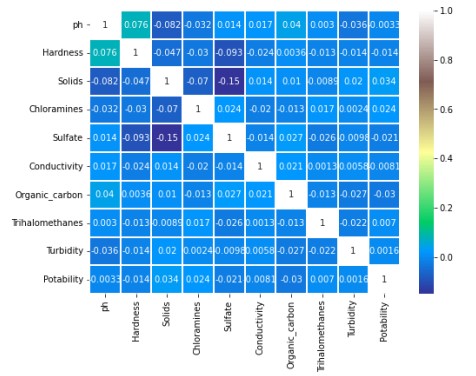
**5.1 OUTPUTS**



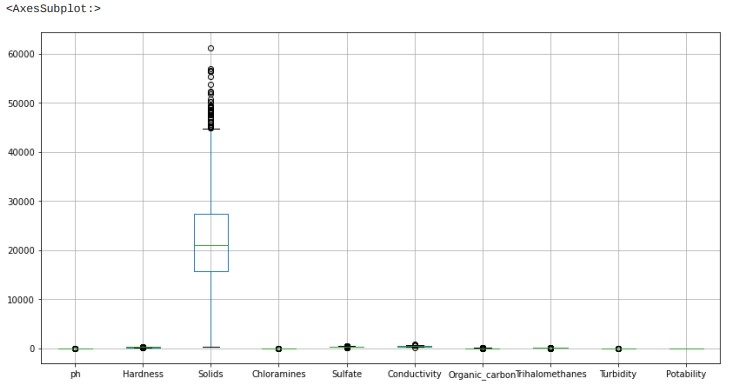
**FIG 5.1**-POTABILITY COUNTS OF DATASET



**FIG 5.1**- HISTOGRAM



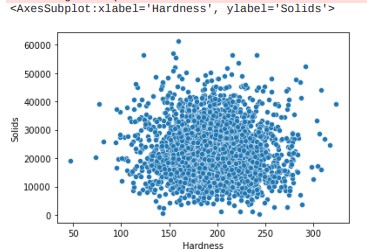
**FIG 5.1**- CORRELATION HEATMAP



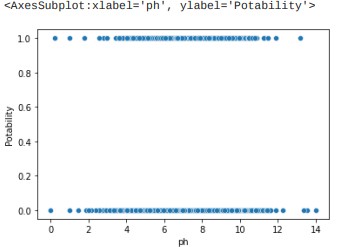
**FIG 5.1-** FINDING OUTLIERS USING BOXPLOT

**5.2 POTABILITY**

Indicates if water is safe for human consumption or not. Potable -1(Safe to drink) and Not potable - 0(Not safe to drink)

******

**Fig 5.2**- Scatter Plot of Hardness and Solids

******

**Fig 5.2**- Scatter Plot of PH and potability

**5.3 Performance Measures Results**

True Positives (TP) are when the model predicts the positive class properly. True Negatives (TN) is one of the components of a confusion matrix designed to demonstrate how classification algorithms work. Positive outcomes that the model predicted incorrectly are known as False Positives (FP). False Negatives (FN) are negative outcomes that the model predicts negative class. Accuracy is the most basic and intuitive performance metric, consisting of the ratio of successfully predicted observations to total observations. Accuracy = TP+TN/(TP+FP+FN+TN)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SN. | ALGORITHM TYPE | ACCURACY SCORE | PRECISION | RECALL | F1-SCORE |
| 1 | Decision tree | 58.9 | 0.42 | 0.38 | 0.40 |
| 2 | k-Nearest neighbour | 61.7 | 0.43 | 0.12 | 0.18 |

**TAB NO 5.3**-COMPARISON OF ALGORITHMS

This section presents the results of the classification algorithms are used to predict the water quality . Table 4.1 shows the results of the used machine learning algorithms.it is noted that the performance of KNN algorithm is superior as compared to decision tree.

## 6. ENVIRONMENT INVESTIGATION

A lot of parameters are responsible for the degradation of water quality among them the following parameters have more prominent effect:

## pH:

pH is a measure of how acidic/basic water is. The range goes from 0 to 14, with 7 being neutral. pH’s of less than 7 indicate acidity, whereas a pH of greater than 7 indicates a base.

The pH of water determines the solubility (amount that can be dissolved in the water) and biological availability (amount that can be utilized by aquatic life) of chemical constituents such as nutrients (phosphorus, nitrogen, and carbon) and heavy metals (lead, copper, cadmium, etc.).

Excessively high and low pH’s can be detrimental for the use of water. High pH causes a bitter taste, water pipes and water-using appliances become encrusted with [deposits](https://www.usgs.gov/special-topic/water-science-school/science/hardness-water), and it depresses the effectiveness of the disinfection of chlorine, thereby causing the need for additional chlorine when pH is high. Low-pH water will corrode or dissolve metals and other substances.

Pollution can change a water's pH, which in turn can harm animals and plants living in the water. For instance, [water coming out of an abandoned coal mine](https://www.usgs.gov/media/images/worlds-most-acidic-water-found-a-mine-california) can have a pH of 2, which is very acidic and would definitely affect any fish crazy enough to try to live in it! By using the logarithm scale, this mine-drainage water would be 100,000 times more acidic than neutral water -- so stay out of abandoned [mines.](https://www.usgs.gov/special-topic/water-science-school/science/mining-water-use)

## D.O (mg/L):

Unlike air, which is normally about 21 percent oxygen, water contains only a tiny fraction of a percentage of dissolved oxygen. In water it usually is expressed in milligrams per litre (mg/L), parts per million (ppm), or percent of saturation. At sea level, typical DO concentrations in 100-percent saturated fresh water will range from 7.56 mg/L (or 7.56 parts oxygen in 1,000,000 parts water) at 30 degrees Celsius to 14.62 mg/L at zero degrees Celsius.

As environmental protection is becoming an important global attention and focuses on getting a value added by product from waste which does not cause environmental pollution or eco-friendly. Chitosan is a natural biopolymer obtained as by-product from sea food industry waste, has versatile application in various fields.

Chitosan is used as a coagulant to remove turbidity and compared with other commonly used coagulant, alum (Aluminium sulphate). In present work, we also focused on parameters such as pH, B.O.D, D.O and nitrate. also studied the various characteristics of chitosan.

## 

# **CHAPTER-7**

# **ADVANTAGES AND DISADVANTAGES**

## Advantages:

* Create public and private places that harvest, clean, and recycle water, resulting in water resource, environmental and social livability benefits.
* Since it predicts the future water quality measures can be taken accordingly.
* No need of human interference in the prediction of water quality.

## Disadvantages:

* This is a supervised algorithm and hence needs to be monitored.
* The data provided to the model should be accurate.
* Lack of large dataset.

**Software Tools**

* Smart Internz
* Anaconda 3
* spyder

# 7.1 **FUTURE SCOPE**

For this water predication machine learning model, we can add few add ones to make it into advanced model. In future, we plan to deal with the water quality inference problems in the urban water distribution systems through a limited number of water quality monitor stations.

In future works, we propose integrating the ﬁndings of this project in a large-scale IoT-based online predicting system using only the sensors of the required parameters. The tested algorithms would predict the water quality immediately based on the real-time data fed from the IoT system.

The proposed IoT system would employ the parameter sensors of pH, turbidity, temperature, and TDS for parameter readings and communicate those readings using an Arduino microcontroller and ZigBee transceiver. It would identify poor quality water before it is released for consumption and alert concerned authorities. It will hopefully result in curtailment of people consuming poor quality water and consequently de-escalate harrowing diseases like typhoid.

# **8. CONCLUSION**

Few water bodies are not advisable for drinking as well as for washing or any other purposes. One of the reasons affecting the quality of water in the various states of India is that the wastewater and industrial effluents are discharged untreated directly into the river. Less than 50 per cent of the population in India has access to safely managed drinking water, Groundwater from over 30 million access points supplies 85 per cent of drinking water in rural areas and 48 per cent of water requirements in urban areas.

The Water quality prediction model can give an overview of how the upcoming future water qualities will be, so that we take precautions to overcome the situations.

If the current trend continues and then by 2025 or soon, we can expect the water to be polluted to an extent from which we might not be able to recover.

Potability determines the quality of water, which is one of the most important resources for existence. Traditionally, testing water quality required an expensive and time-consuming lab analysis. This study looked into an alternative machine learning method for predicting water quality using only a few simple water quality criteria. To estimate, a set of representative supervised machine learning algorithms was used. It would detect water of bad quality before it was released for consumption and notify the appropriate authorities It will hopefully reduce the number of individuals who drink low-quality water, lowering the risk of diseases like typhoid and diarrhea. In this case, using a prescriptive analysis based on projected values would result in future capabilities to assist decision and policy makers.

Machine learning has been widely used as a powerful tool to solve problems in the water environment because it can be applied to predict water quality, optimize water resource allocation, manage water resource shortages, etc. Despite this, several challenges remain in fully applying machine learning approaches in this field to evaluate water quality:

(1) Machine learning is usually dependent on large amounts of high-quality data. Obtaining sufficient data with high accuracy in [water treatment](https://www.sciencedirect.com/topics/earth-and-planetary-sciences/water-purification) and management systems is often difficult owing to the cost or technology limitations.

(2) As the conditions in real water treatment and management systems can be extremely complex, the current algorithms may only be applied to specific systems, which hinders the wide application of machine learning approaches.

(3) The implementation of machine learning algorithms in practical applications requires researchers to have certain professional background knowledge.

To overcome the above-mentioned challenges, the following aspects should be considered in future research and engineering practices:

(1) More advanced sensors, including soft sensors, should be developed and applied in water quality monitoring to collect sufficiently accurate data to facilitate the application of machine learning approaches.

(2) The [feasibility](https://www.sciencedirect.com/topics/earth-and-planetary-sciences/feasibility) and reliability of the algorithms should be improved, and more universal algorithms and models should be developed according to the water treatment and management requirements.

(3) Interdisciplinary talent with knowledge in different fields should be trained to develop more advanced machine learning techniques and apply them in engineering practices.

**8.2 Source Code:**

**import** numpy **as** np

**import** pandas **as** pd

**import** matplotlib.pyplot **as** plt

dataset **=**pd**.**read\_csv('water\_data.csv',encoding**=**'unicode\_escape')

dataset

dataset**.**info()

dataset**.**iloc[:,3:]**=**dataset**.**iloc[:,3:]**.**applymap(**lambda** x: pd**.**to\_numeric(x,errors**=** "coerce"))

dataset**.**info()

dataset**.**drop(['STATE','STATION CODE',"LOCATIONS","TOTAL COLIFORM (MPN/100ml)Mean"],a

print(dataset**.**isnull()**.**any())

dataset**.**describe()

plt**.**scatter(range(1991),dataset["PH"])

plt**.**scatter(range(1991),dataset["D.O. (mg/l)"])

plt**.**scatter(range(1991),dataset["CONDUCTIVITY (µmhos/cm)"])

plt**.**scatter(range(1991),dataset["B.O.D. (mg/l)"])

plt**.**scatter(range(1991),dataset["NITRATENAN N+ NITRITENANN (mg/l)"])

plt**.**scatter(range(1991),dataset["FECAL COLIFORM (MPN/100ml)"])

dataset**=**dataset[dataset["PH"]**<**14]

dataset**=**dataset[dataset["PH"]**>**4]

dataset**=**dataset[dataset["B.O.D. (mg/l)"]**<**190]

dataset**=**dataset[dataset["FECAL COLIFORM (MPN/100ml)"]**<**1000000000]

print(dataset**.**info())

dataset**.**describe()

dataset['Temp']**=**dataset['Temp']**.**replace(np**.**NaN,dataset['Temp']**.**mean())*#26.318446*

dataset['D.O. (mg/l)']**=**dataset['D.O. (mg/l)']**.**replace(np**.**NaN,dataset['D.O. (mg/l)']**.**mean())

dataset['CONDUCTIVITY (µmhos/cm)']**=**dataset['CONDUCTIVITY (µmhos/cm)']**.**replace(np**.**NaN,dataset['CONDUCTIVITY (µmhos/cm)']**.**mean())

dataset['NITRATENAN N+ NITRITENANN (mg/l)']**=**dataset['NITRATENAN N+ NITRITENANN (mg/l)']**.**replace(np**.**NaN,dataset['NITRATENAN N+ NITRITENANN (mg/l)']**.**mean())

dataset**.**info()

df**=**dataset**.**groupby(by**=**["year"],sort**=True**,as\_index**=True**)**.**mean()

df

df**.**describe()

y**=**pd**.**Series()

yy**=**pd**.**DataFrame()

y**=**df["PH"]**.**apply(**lambda** x: (0 **if** (8**>=**x**>=**7)

**else** (0.028 **if** (8.5**>=**x**>=**8) **or** (7**>=**x**>=**6.5)

**else** (0.084 **if** (9**>=**x**>=**8.8) **or** (6.5**>=**x**>=**6.3)

**else** (0.112 **if** (10**>=**x**>=**9) **or** (6.3**>=**x**>=**6)

**else** 0.14)))))

yy["PH"]**=**df["PH"]**.**apply(**lambda** x: (0 **if** (8**>=**x**>=**7)

**else** (0.028 **if** (8.5**>=**x**>=**8) **or** (7**>=**x**>=**6.5)

**else** (0.084 **if** (9**>=**x**>=**8.8) **or** (6.5**>=**x**>=**6.3)

**else** (0.112 **if** (10**>=**x**>=**9) **or** (6.3**>=**x**>=**6)

**else** 0.14)))))

yy["D.O. (mg/l)"]**=**df["D.O. (mg/l)"]**.**apply(**lambda** x: (0 **if** (8**>=**x**>=**6.5)

**else** (0.04 **if** (6.5**>=**x**>=**6)

**else** 0.2)))

y**=**y**+**yy["D.O. (mg/l)"]

yy["CONDUCTIVITY (µmhos/cm)"]**=**df["CONDUCTIVITY (µmhos/cm)"]**.**apply(**lambda** x: (0 **if** (1500**>=**x**>=**50)

**else** (0.012 **if** (2000**>=**x**>=**1500)

**else** (0.048 **if** (2500**>=**x**>=**2000)

**else** 0.06))))

y**=**y**+**yy["CONDUCTIVITY (µmhos/cm)"]

yy["B.O.D. (mg/l)"]**=**df["B.O.D. (mg/l)"]**.**apply(**lambda** x: (0 **if** (3**>=**x**>=**0)

**else** (0.024 **if** (5**>=**x**>=**3)

**else** (0.072 **if** (10**>=**x**>=**5)

**else** 0.12))))

y**=**y**+**yy["B.O.D. (mg/l)"]

yy["FECAL COLIFORM (MPN/100ml)"]**=**df["FECAL COLIFORM (MPN/100ml)"]**.**apply(**lambda** x: (0 **if** (5000**>=**x**>=**0)

**else** (0.04 **if** (10000**>=**x**>=**5000)

**else** (0.12 **if** (100000**>=**x**>=**10000)

**else** 0.2))))

y**=**y**+**yy["FECAL COLIFORM (MPN/100ml)"]

y**=**y**\***100

yy["y"]**=**y

yy

x**=**df**.**index**.**tolist()

x**=**list(map(**lambda** z:[z,],x))

x

y**=**list(y)

y

**from** sklearn.model\_selection **import** train\_test\_split

x\_train,x\_test,y\_train,y\_test**=**train\_test\_split(x,y,test\_size**=**0.2,random\_sta

**from** sklearn.linear\_model **import** LinearRegression

lr**=** LinearRegression()

lr**.**fit(x\_train,y\_train)

**from** sklearn.linear\_model **import** LinearRegression

lr**=** LinearRegression()

lr**.**fit(x\_train,y\_train)

plt**.**scatter(x\_test,y\_test)

plt**.**plot(x\_test,y\_pre,"r")

lr**.**predict([[2025]])

**from** sklearn.metrics **import** r2\_score

r2\_score(y\_test,y\_pre)

**from** sklearn.metrics **import** mean\_squared\_error

mean\_squared\_error(y\_test,y\_pre)

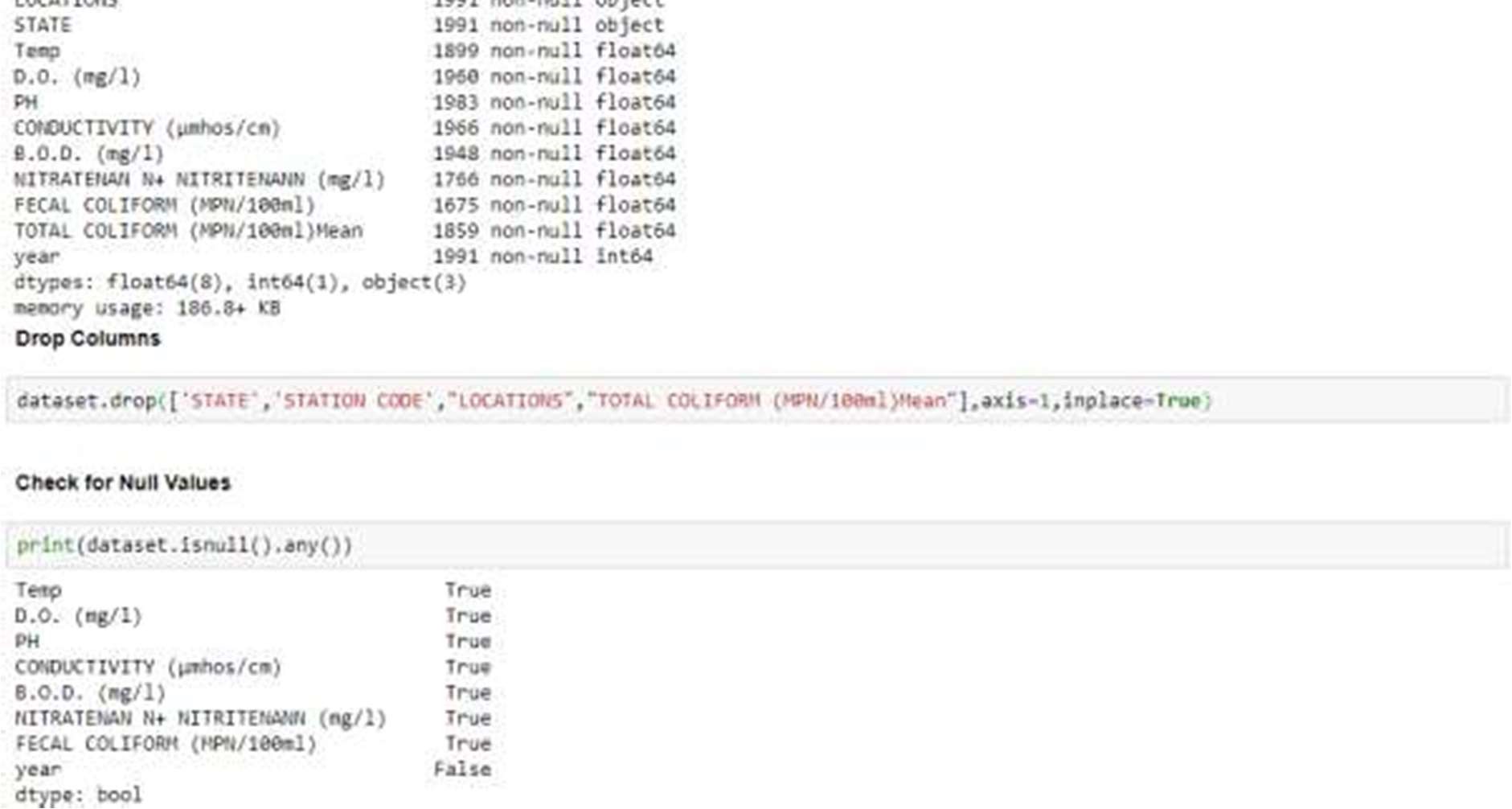
**from** joblib **import** dump

dump(lr,'model.save')

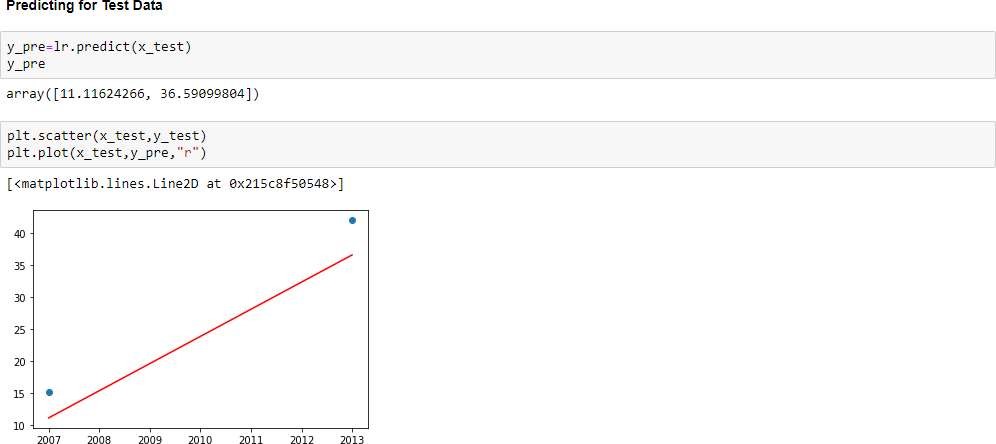
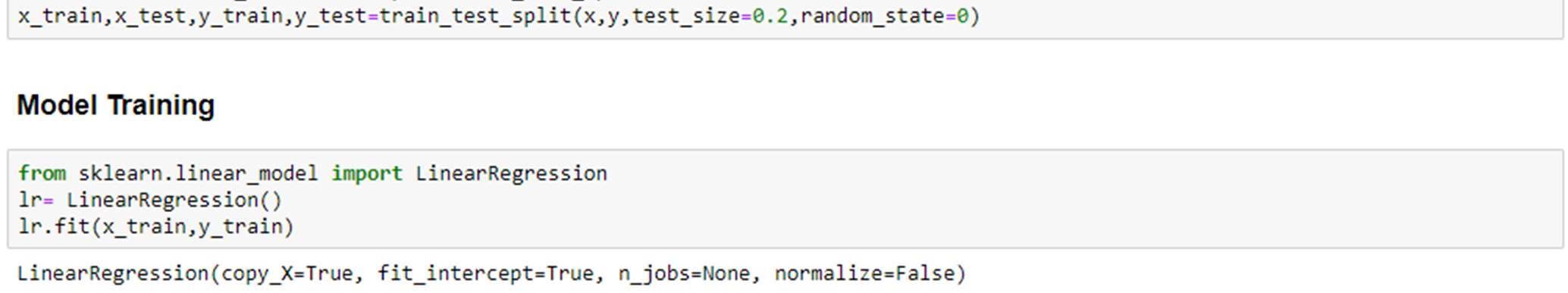
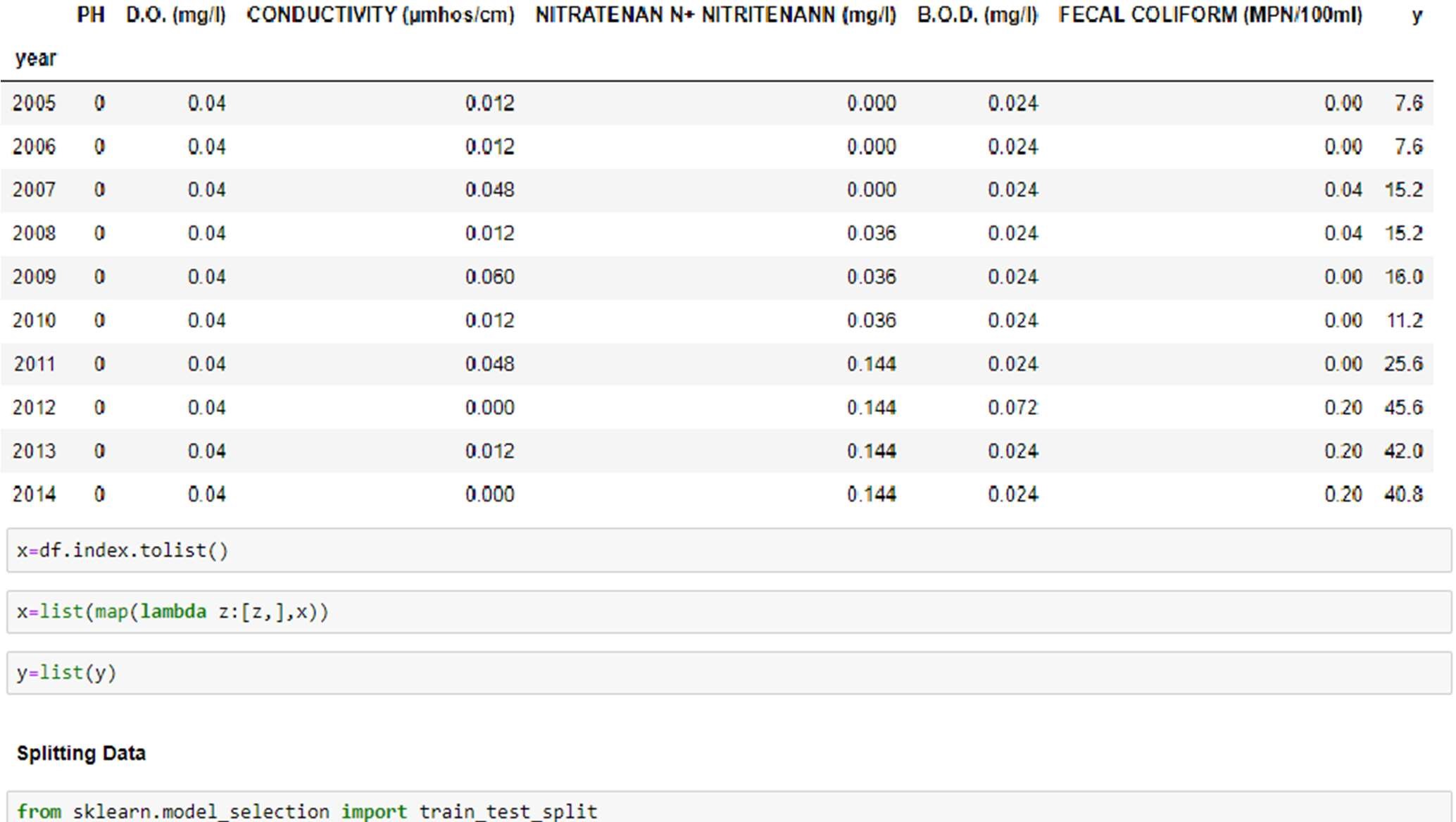
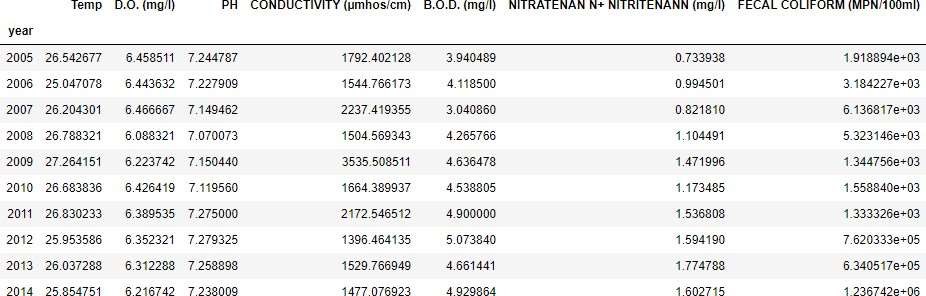
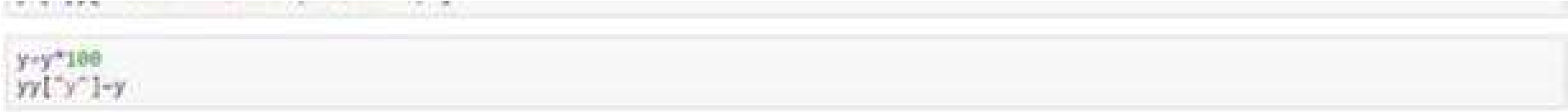
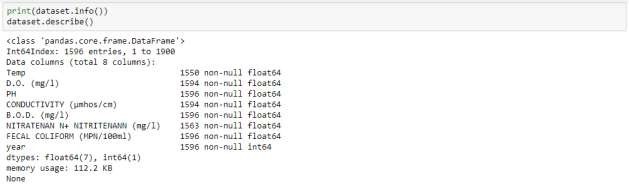
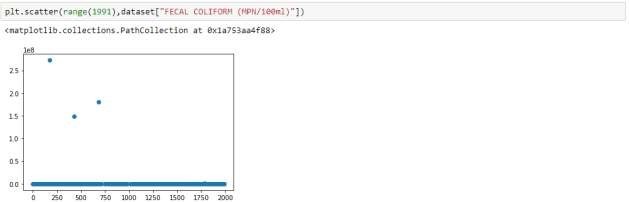
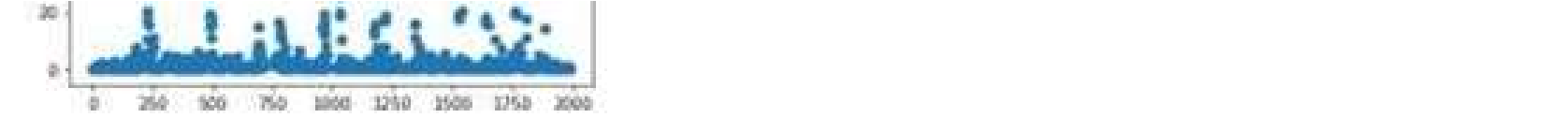
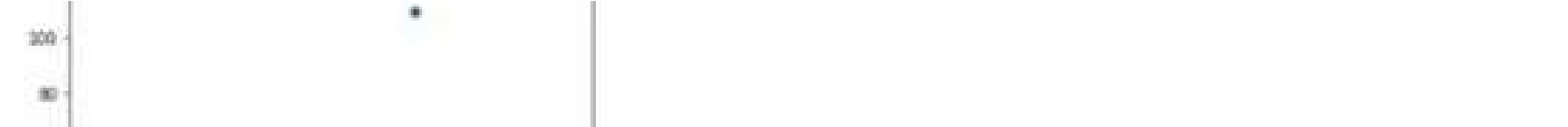
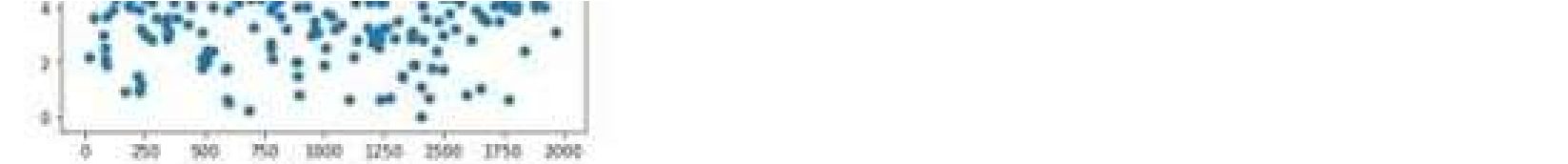
## 

## Source Code

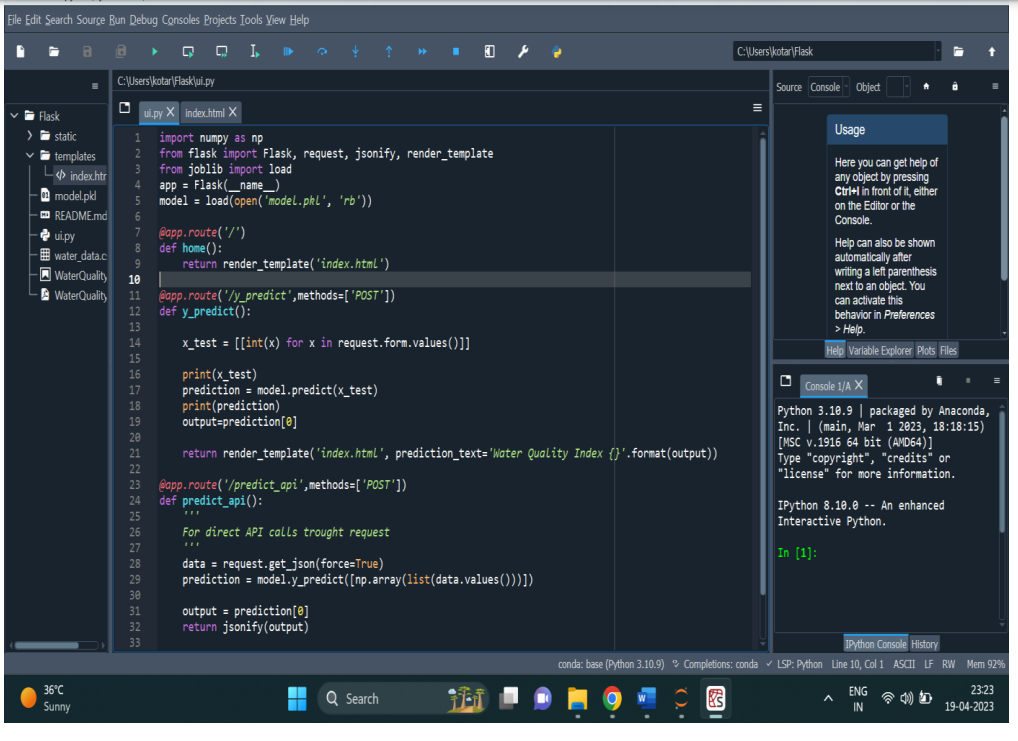
**(i) Data Pre-processing:**

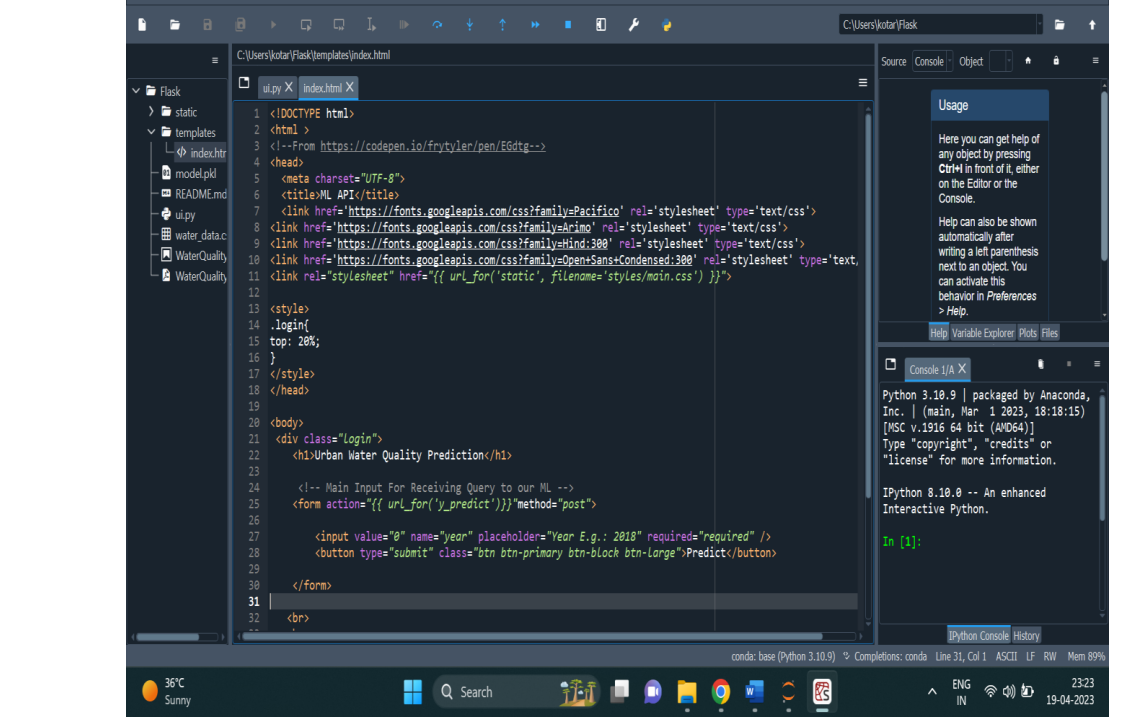




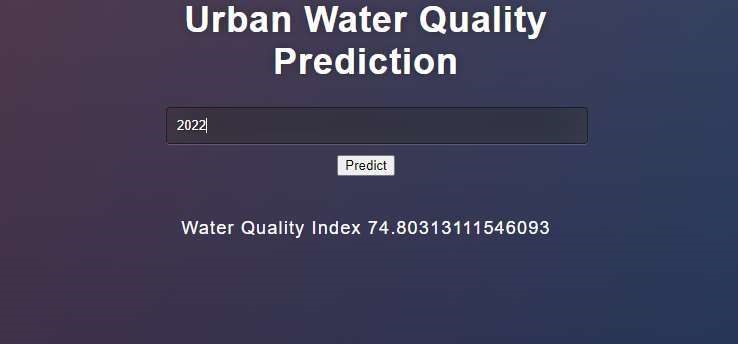


**Ui.py(spyder)**

****

**Index.html**

**UI Output screenshot**



**REFERENCES**

<https://www.kaggle.com/datasets/adityakadiwal/water-potability>

1. N. Marir, H. Wang, G. Feng, B. Li, and M. Jia, “Distributed abnormal behavior detection approach based on deep belief network and ensemble SVM using spark,” *IEEE Access*, vol. 6, pp. 59657–59671, 2018.
2. Z. M. Fadlullah, F. Tang, B. Mao, J. Liu, and N. Kato, “On intelligent traffic control for large-scale heterogeneous networks: a value matrix-based deep learning approach,” *IEEE Communications Letters*, vol. 22, no. 12, pp. 2479–2482, 2018.
3. S. Maiti and R. K. Tiwari, “A comparative study of artificial neural networks, Bayesian neural networks and adaptive neuro-fuzzy inference system in groundwater level prediction,” *Environmental Earth Sciences*, vol. 71, no. 7, pp. 3147–3160, 2014.
4. C. Min, “An improved recurrent support vector regression algorithm for water quality prediction,” *Journal of Computational Information*, vol. 12, pp. 4455–4462, 2011.
5. R. Das Kangabam, S. D. Bhoominathan, S. Kanagaraj, and M. Govindaraju, “Development of a water quality index (WQI) for the Loktak Lake in India,” *Applied Water Science*, vol. 7, no. 6, pp. 2907–2918, 2017.
6. G. Srivastava and P. Kumar, “Water quality index with missing parameters,” *International Journal of Research in Engineering and Technology*, vol. 2, no. 4, pp. 609–614, 2013.
7. S. Tyagi, B. Sharma, P. Singh, and R. Dobhal, “Water quality assessment in terms of water quality index,” *American Journal of Water Resources*, vol. 1, no. 3, pp. 34–38, 2013
8. A. A. Al-Othman, “Evaluation of the suitability of surface water from Riyadh Mainstream Saudi Arabia for a variety of uses,” *Arabian Journal of Chemistry*, vol. 12, no. 8, pp. 2104–2110, 2019.
9. T. H. H. Aldhyani, M. Alrasheedi, A. A. Alqarni, M. Y. Alzahrani, and A. M. Bamhdi, “Intelligent hybrid model to enhance time series models for predicting network traffic,” *IEEE Access*, vol. 8, pp. 130431–130451, 2020.