Urban Water Quality Prediction

1.Introduction.

a.Overview:

Water occupies about 70% of the earth's surface and is one of the most important resources for a sustaining life. Rapid urbanization and and increse in industrialization have led to a deterioration of water quality at an alarming rate, resulting in various diseases. Water quality has been estimated through expensive and time-consuming lab and statistical analyses, which isn't helping to monitor water quality. The increasing consequences of poor water quality calls need for an alternative method, which should take less time and medium or low expenditure. With this all these concerns and needs , this documentation gives you a overview of how supervised machine learning algorithms can be used to estimate the water quality index (WQI), which is a singular and one of the popular index to describe the general quality of water, and the water quality class (WQC), which is a distinct class defined on the basis of the WQI. The method we are using employs several input parameters, namely, temperature, turbidity, pH and total dissolved solids etc.

So using mentioned Input parameters ,we are going to compute value of Water Quality Index and build a simple yet useful application which can be used to predict WQI value for various reasons ,which helps in further Analysis of Water quality in Urban areas.

Commonly used words/keywords: water quality prediction; supervised machine learning; ; pollution;water quality index(wqi);

b.Purpose:

As stated in overview of Introduction , main purpose of this project is to find an alternative method which should take less time and medium-to-low expenditure and mainly advancements in technology has to be properly used. We can use various machine learning algorithms and methods for data visulation,preprocessing etc.Hence it is benificial in both ways, it helps us to predict water quality using WQI and that too with the knowledge we have on alogorithms and various methods that include data.

2.Literature Survey

a.Existing problem:

Water is the most important of sources, vital for sustaining all kinds of life; however, it is in constant threat of pollution by life itself. Water is one of the most communicable mediums with a far reach. Rapid industrialization has consequently led to deterioration of water quality at an alarming rate. Poor water quality results have been known to be one of the major factors of escalation of harrowing diseases. As reported, in developing countries, 80% of the diseases are water borne diseases,which have led to 5 million deaths and 2.5 billion illnesses .The most common of these diseases of our India are diarrhea, typhoid, gastroenteritis, cryptosporidium infections, some forms of hepatitis and giardiasis intestinal worms . In India , water borne diseases, cause a GDP loss .This makes it a pressing problem, particularly in a developing country like ours.

Water quality is currently estimated through expensive and time-consuming lab and statistical analyses, which require sample collection, transport to labs, and a considerable amount of time and calculation, which is quite ineffective given water is quite a communicable medium and time is of the essence if water is polluted with disease-inducing waste.The horrific consequences of water pollution necessitate a quicker and cheaper alternative.

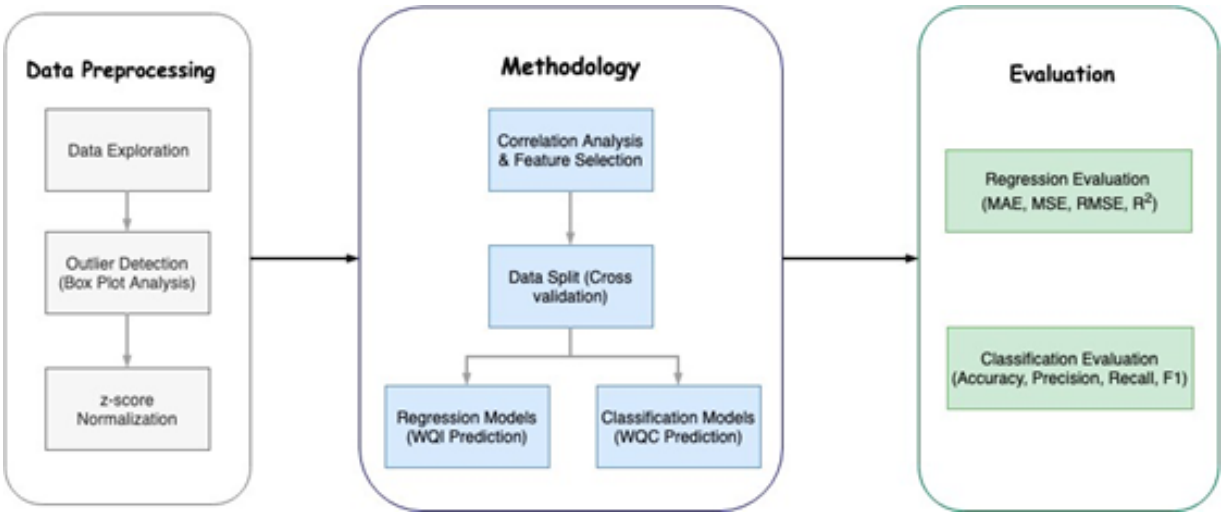
This project explores the methods that have been used to help solve problems related to water quality. Typically, conventional lab analysis and statistical analysis are used in research to help in determining water quality, while some analyses employ machine learning methodologies to assist in finding an optimized solution for the water quality problem.

In this regard, the main motivation in this study is to propose and evaluate an alternative method based on supervised machine learning for the efficient prediction of water quality in

real-time.
This project is developed on the dataset of water resources , situated in India,available online and in many open source websites.A representative set of supervised machine learning algorithms were employed on the said dataset for predicting the water quality index (WQI) for several years.

b.Proposed Solution

Most of the research either employed manual lab analysis, not estimating the water quality index standard, or used too many parameters to be efficient enough. The proposed methodology improves on these notions and the methodology being followed is depicted in Figure(given below).

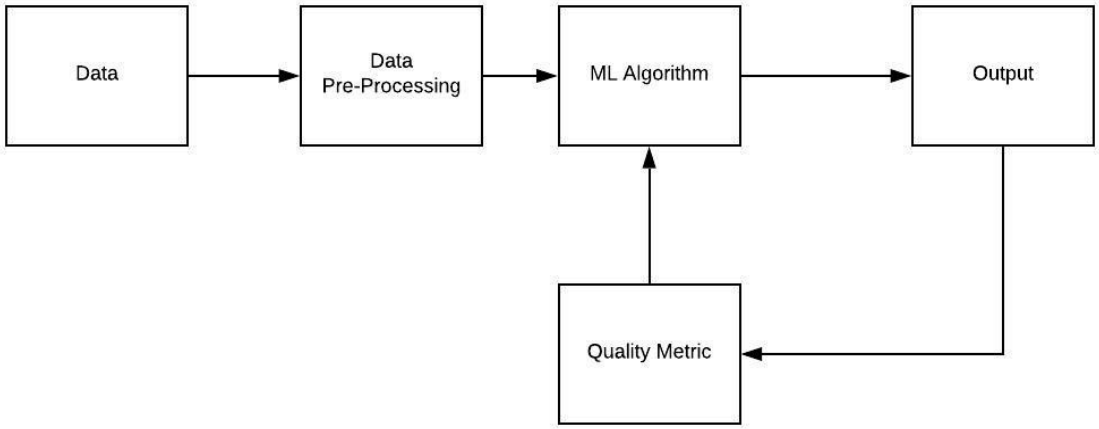


In this way this expensive and cumbersome lab analysiswith specific sensors can be avoided in further similar analyses.
A series of representative supervised prediction (regression) algorithms were tested on the dataset worked here. The complete methodology is proposed in the context of water quality

numerical analysis.

3.Theoretical Analysis

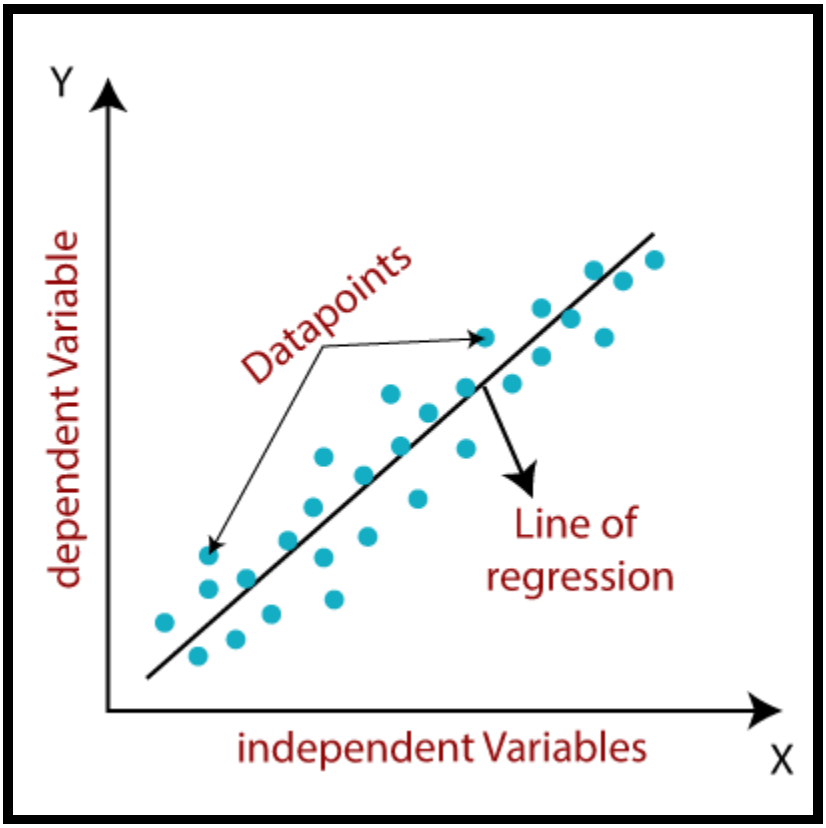
a.Block diagram:



As mentioned in a earlier section this is the block diagram ,dataprocessing,Methodology and then Evaluation. Here in this project in Linear Regression is done and evaluation metrics used are r_score and mean squared error.
More detailed documnetation of all steps in further sections!

b.Hardware/Software desinging:

Given Software desinging best represents Linear Regression analysis, it is plot of dependent variable versus independent variable plotted using data points. There is best fit line over there,finding which is main aim of any Linear regression model . Linear regression analysis is done to find best fit line and minimise sum of sqaure of errors,that is difference between actual y value and predicted value.



4.Experimental Analysis

This section gives complete description about block diagram mentioned and all steps that are performed before building our machine learning model.The data used for this research was obtained from kaggale website and it was cleaned by performing a box plot analysis, discussed

in section. After the data were cleaned, they were normalized using q-value normalization to convert them to the range of 0–100 to calculate the WQI using eight available parameters

Temp	D.O. (mg/l)	PH	CONDUCTIVITY (µmhos/cm)	B.O.D. (mg/l)	NITRATENAN N+ NITRITENANN (mg/l)	FECAL COLIFORM (MPN/100ml)	TOTAL COLIFORM (MPN/100ml)Mean

Using these eight parameters WQI value is calculated. Once the WQI was calculated, all original values can be normalised using z-score and then analysis can be done with year and calculated WQI value. The complete procedure is detailed next.

Data preprocessing

Data Collection

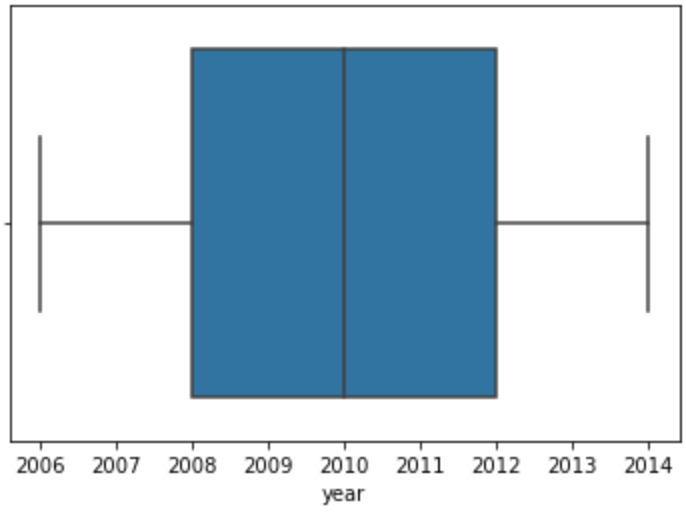
The dataset collected from kaggle website contained 1991 samples from different sources of our Indian Water resources like Lakes,rivers etc collected throughout 2006 to 2014.Given table shows parameters along with their WHO standard limits.We dont have all of these parameters but 7 to 8 of them.

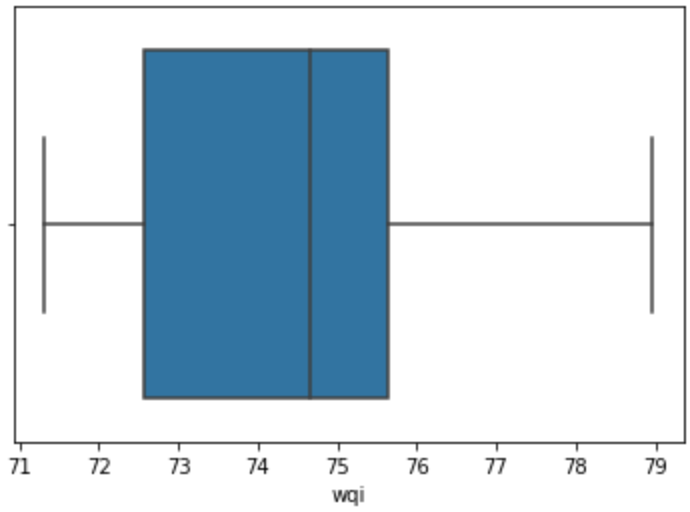
Table . Parameters along with their “WHO” standard limits

Parameter	WHO Limits
Alkalinity	500 mg/L
Appearance	Clear
Calcium	200 mg/L
Chlorides	200 mg/L
Conductance	2000 µS/cm
Fecal Coliforms	Nil Colonies/100 mL
Hardness as $CaCO_3$	500 mg/L
Nitrite as NO_2^-	<1 mg/L
pH	6.5–8.5
Temperature	◦ C
Total Dissolved Solids	1000 mg/L
Turbidity	5 NTU

Outlier detection

We chose boxplot analysis for outlier detection because most of the parameters varied and a boxplot provides insightful visualization to decide outlier detection threshold values depending upon the problem domain. Boxplot analysis of independent variable year has shown no outliers and even dependent variables "WQI" has no outliers and glimpse of boxplot which is done using seaborn is as provided.Snapshots of boxplot of both dependent and independent variables are attached below.





Water Quality Index(WQI)

Water quality index (WQI) is the singular measure that indicates the quality of water and it is calculated using various parameters that are truly reflective of the water's quality. To conventionally calculate the WQI, nine water quality parameters are used, but if we did not have all of them, we could still estimate the water quality index with at least parameters we have pH, D.O., Conductivity, B.O.D,Nitratenan N+,Fecal coliform,Total Coliform.Using these parameters and their assigned weightages, we calculated the WQI of each sample as reflected in Equation where q_{value} reflects the value of a parameter in the range of 0–100 and *w_factor* represents the weight of a particular parameter as listed in Table . WQI is fundamentally calculated by initially multiplying the q value of each parameter by its corresponding weight, adding them all up and then dividing the result by the sum of weights of the employed parameters .

P*qvalue* × *w_factor*

WQI= _____ / water factor *

Parameters weights for the WQI calculation:

Weighing Factor	Weight
pH	0.11
Temperature	0.10
Turbidity	0.08
Total Dissolved Values	0.07
Nitrates	0.10
Fecal Coliform	0.16

Z-Score Normalisation

The z-score is a conventional standardization and normalization method that represents the number of standard deviations; a raw data point is above or below the population mean. It ideally lies between −3 and +3. It normalizes the dataset to the aforementioned scale to convert all the data with varying scales to the default scale.

To normalize the data using the z-score, we subtracted the mean of the population from a raw data point and divided it by the standard deviation, which gives a score ideally varying between −3 and +3; hence, reflecting how many standard deviations a point is above or below the mean as computed by Equation, where x represents the value of a particular sample, μ represents the mean and σ represents the standard devia

$$z \text{ score} = (x - \mu) / \sigma$$

Data Analysis

After all the data processing, for data analysis, linear regression ,which is a machine learning algorithms employed to predict the WQI using the parameters. Before applying a machine learning algorithm, there are some steps, like data splitting, to prepare the data to be given as input to the actual machine learning algorithms.

Data Splitting

The last step prior to applying the machine learning model is splitting the provided data in order to train the model, test it with a certain part of the data and compute the accuracy measures to establish the model's performance. We used train_test_split from sklearn.model selection.So since data is splitted to train data and test data, now machine learning algorithm can applied.

Machine Learning Algorithm

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:

- One variable, denoted *x*, is regarded as the **predictor**, **explanatory**, or **independent** variable.

- The other variable, denoted *y*, is regarded as the **response**, **outcome**, or **dependent** variable.

Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line, while logistic and nonlinear regression models use a curved line. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.

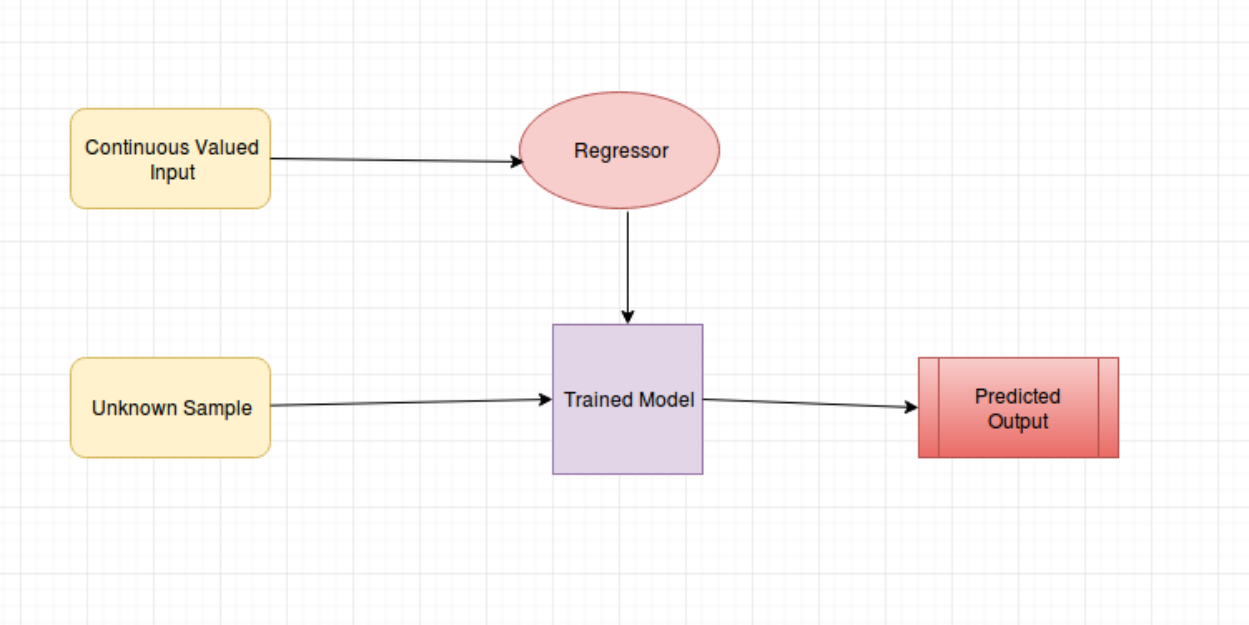
Simple linear regression is used to estimate the relationship between two quantitative variables. You can use simple linear regression when you want to know:

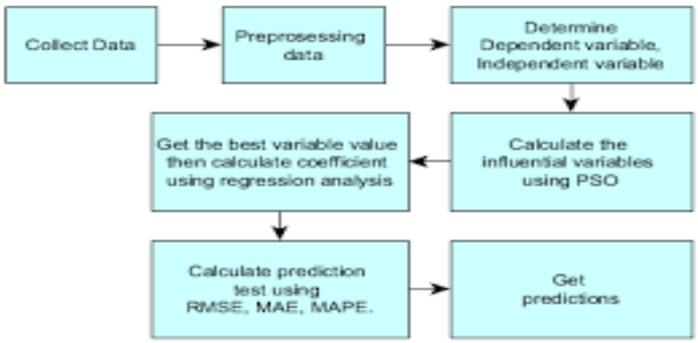
How strong the relationship is between two variables (e.g. the relationship between rainfall and soil erosion).

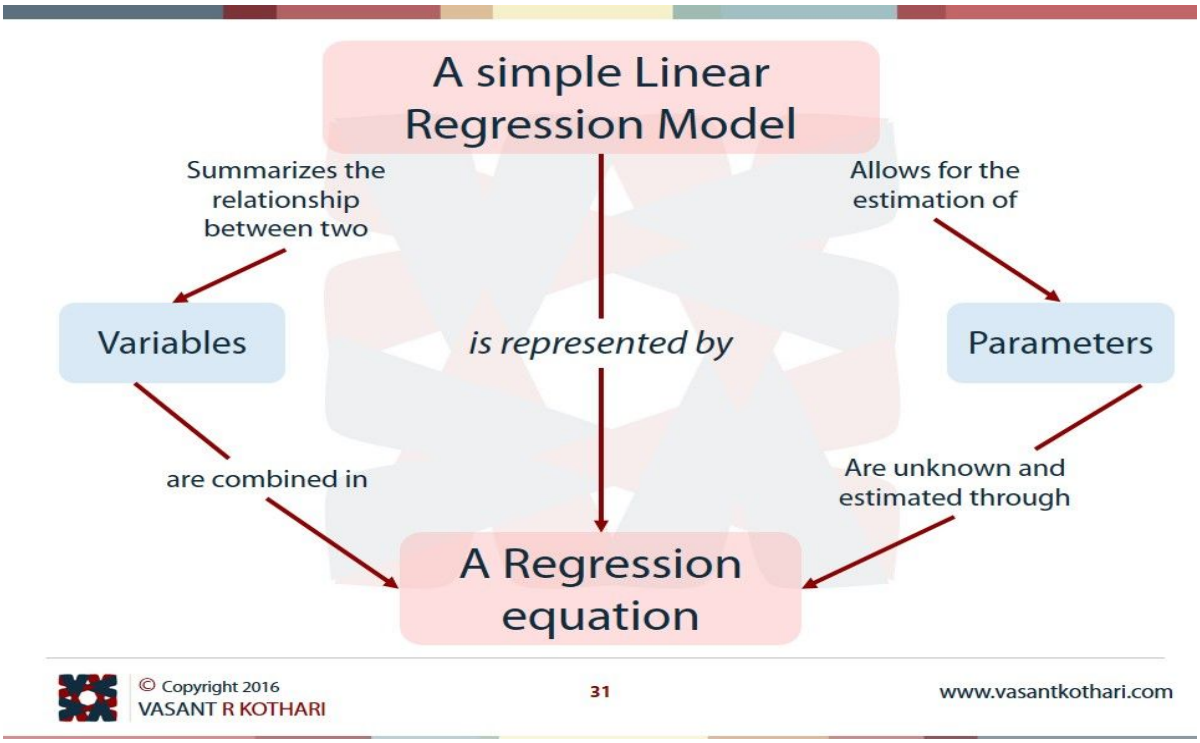
The value of the dependent variable at a certain value of the independent variable (e.g. the amount of soil erosion at a certain level of rainfall).

5.Flowchart

Various Flowcharts attached here convey the same thing , to build a linear regression model data collection,data -preprocessing,data visualisation, splitting data,model building,evalution metrics are steps that have to be followed in sequence as we did.


Continuous Valued Input
Regressor
Unknown Sample
Trained Model
Predicted Output


Collect Data
Preprocessing data
Determine Dependent variable, Independent variable
Calculate the influential variables using PSO
Get the best variable value then calculate coefficient using regression analysis
Calculate prediction test using RMSE, MAE, MAPE.
Get predictions



6.Results

Accuracy measures

1.R2 score

R-squared is the percentage of the response variable variation that is explained by a linear model. It is always between 0 and 100%. R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

2.Mean square error (MSE)

Mean square error (MSE) is the sum of squares of errors divided by the total number of predicted values. This attributes greater weight to larger errors. This is particularly useful in problems where there needs to be a larger weight for larger errors. It is measured by Equation (6), where x_{obs} is the actual value, x_{pred} is the predicted value, and n is the total number of samples considered.

MSE =P(*xobs – xpred*)^2/n

Below are the glimpses of results we obtained with our dataset

```
In [45]: from sklearn.metrics import r2_score
         r2_score(a,y_test)
         #calculating accuracy using r2_score

Out[45]: 0.8079732483547792
```

```
In [46]: from sklearn.metrics import mean_squared_error
         print('mse:%.2f'%mean_squared_error(y_test,a))

         mse:0.03
```

7.Advantages and Disadvantages

Advantages from this linear regression is that it simplified all the hectic input parameters by concatinating them to single Water quality index (WQI) and hence model is built with year as independent variable and WQI dependent variable. As we already know linear regression is about finding best fit line ,it is done even here with years ranging from 2006 to 2014. We can make predictions and get value of Water quality index (WQI) for dynamic data, that is for year other than those between 2006-2014 which is the whole purpose of this project.

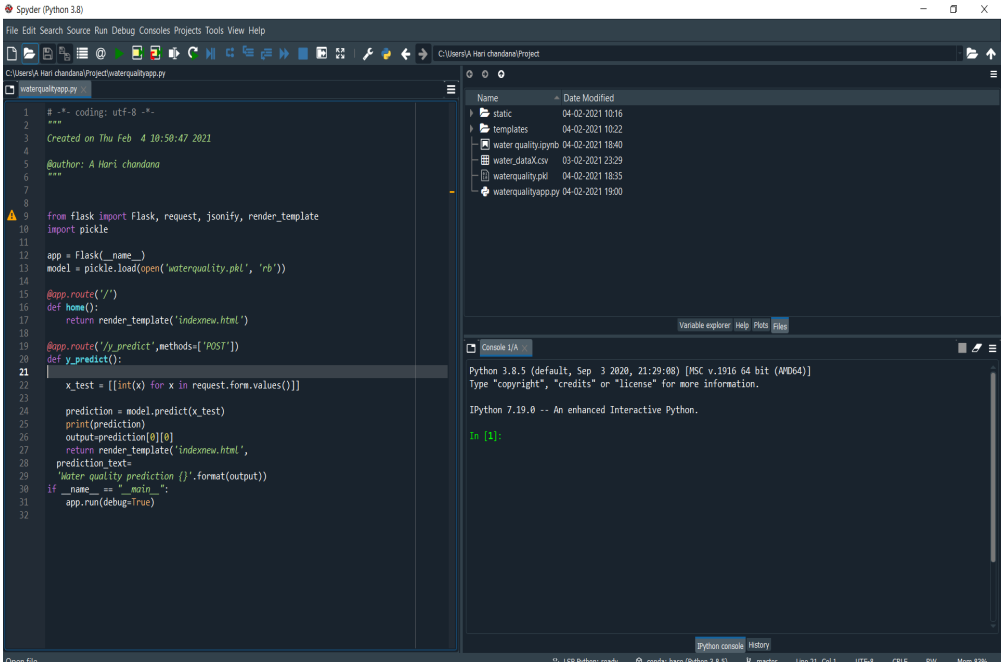
Disadvantage that we observed is that Accuracy that we got which is not appreciatably good, so instead of using Linear Regression we can shift to Multi linear regression considering all input parameters we have in our dataset.

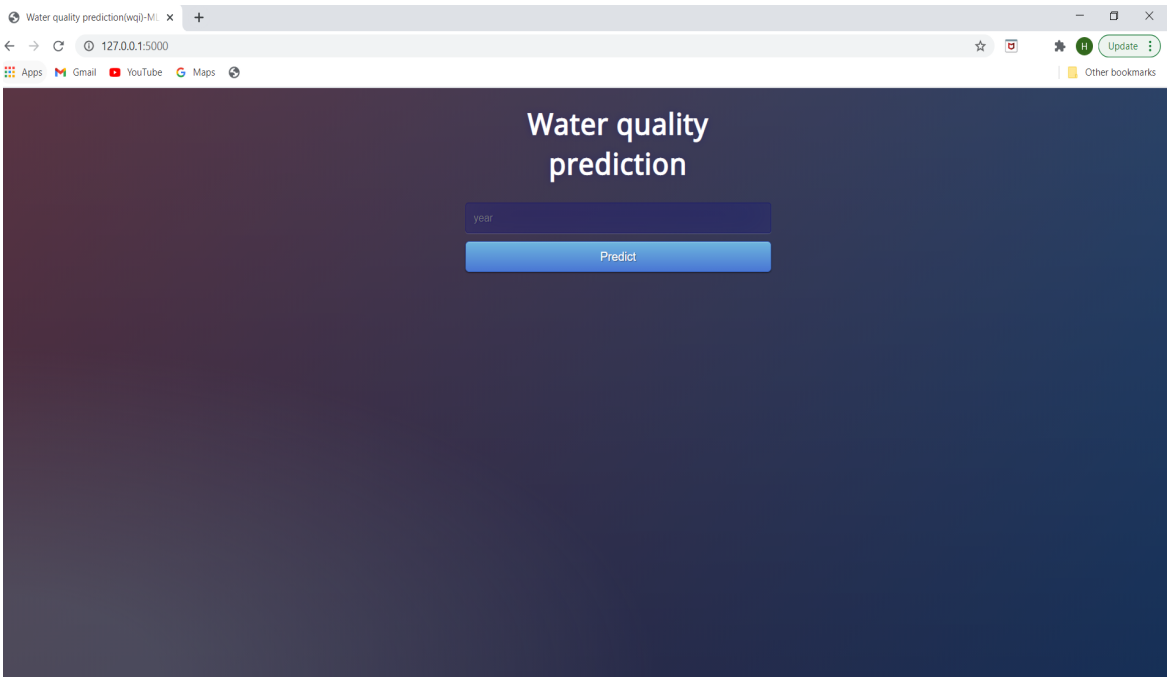
8.Applications

Major Application is this linear regression model can be used for building Application using Spyder Idle, it can be deployed in any open source service providing sites like IBM cloud, Google cloud etc.Since major aim was to compute Water quality Index for various years, it can be used by people or communities which are into Enviromental study and research.Water Quality is conventionally calculated using water quality parameters, which are acquired through time consuming lab analysis. We explored alternative methods of machine learning to estimate it and found several studies employing them. These studies used more than 6 water quality parameters to predict WQI.

9.Conclusion

Water is one of the most important and essential resources for our survival and its quality is determined through WQI. Conventionally, to test water quality, one has to go through expensive and cumbersome lab analysis. This project tried to explore an alternative method of machine learning to predict water quality using available water quality parameters. The data used to conduct the study were taken from kaggale website which is inturn taken from Indian government's official website and has 1991 samples from different water resources of India . A supervised machine learning algorithm(Linear Regression) is used to compute WQI.It showed that WQI value can be calculated while taking year as input.
Also we developed application which can predict WQI value using flask and spyder Idle,here is a glimpse of how it looks.





10.Future Scope

In future works, we propose integrating the findings of this research in a large-scale IoT-based online monitoring system using only the sensors of the required parameters. The tested algorithmswould 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 and diarrhea. In this regard, the application of a prescriptive analysis from the expected values would lead to future facilities to support decision and policy makers.

11.Bibliography

1.Improving water quality index prediction in Perak River basin Malaysia through a combination of multiple neural networks. *Int. J. River Basin Manag.* **2017**, *15*, 79–87. [CrossRef]

2.M. Artificial intelligence for the prediction of water quality index in groundwater systems. *Model. Earth Syst. Environ.* **2016**, *2*, 8. [CrossRef]

3.Dataset is taken from Kaggle website and link is here
https://www.kaggle.com/anbarivan/indian-water-quality-data

4.Flowchart of Simple linear Regression
https://medium.com/coinmonks/linear-regression-bf5141ce9ac8

12.Appendix

a.Source code

Source code along with code of Application usingour linear regression model is built is added to Github repository and link is attached here https://github.com/ahcs21/waterquality/tree/master

b.UI Output Screen

We built a application using Flask,Spyder and our Linear regression model and steps we followed are as follows,

1.Creating HTML file
2.Importing Libraries
3.Routing to HTML page.
4.Showing predction on UI
5.Run the app in Localserver

All these steps have been properly executed and we have run app on localserver,here is the UI

output screenshot.