**Datasets:**

There are 3 type of Dataset, as it mentioned in the main manuscript. First dataset contain 5 features as follows:

* T
* PH
* EC
* PO4
* NH4

for Second dataset the NH4 column drops and for the third EC and NH4 doesn’t being used.

**Outliers:**

the data contains many outliers as it can calculated by the IQR method. The IQR method calculates the interval between quarter 2 and quarter 3 of each column.

IQR=Q3-Q2

any data greater than (1.5 \* IQR +Q3) counts as outlier. In the same way, any data lower than (1.5\* IQR – Q2) is a outlier and it should be deleted. This method implemented on both train and test data.

**Scaling:**

In the manuscript it mentioned Z Score scaling. It can be calculated by the following formula:

z = ( x - μ ) / σ

**Models:**

The models that are used in this research are as follows:

* Decision Tree
* Random Forest
* Gradient Boost Regressor
* Support Vector Regressor
* Multi Layer Perceptron
* K nearest neighbors regressor

**Results:**

|  |
| --- |
|  |
|  |
|  |

**The Best Model:**

As it can be concluded by the charts, the best model is KNN on every datasets. The KNN model works by the nearest data around the sample. The K=3 and the nearest samples calculated by the distance around the sample.

The metrics on KNN model is showed in the following table:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| dataset | mse | rmse | Relative % | mae | corr | r2 |
| ds1 | 0.00002 | 0.00414 | -0.00694 | 0.00297 | 0.97759 | 0.90168 |
| ds2 | 0.00001 | 0.00313 | -0.00461 | 0.00226 | 0.98773 | 0.93429 |
| ds3 | 0.00245 | 0.04947 | 0.01677 | 0.02382 | 0.99739 | 0.99362 |