River Water Quality Index Prediction using different Machine Learning Model using UK River Data

***Abstract*—River Water Quality analysis is pretty essential for all nations. Every year many rivers are getting polluted in Bangladesh. We need to analyze them yearly and divide them into types. Here in our paper, we will work with the data of the water quality of UK rivers. We have got the water quality data of 1990 to 2006 of UK rivers. We are going to analyze this data by plotting a different kind of regression and rel plotting. Also, We are going to train a model using various machine learning algorithms to predict the index of water. Finally, we will evaluate our model with different accuracy metrics to find the performance of our trained model.**

***Index Terms*—river, quality, encoder, dataset, decisition Tree, random forest**

I. Introduction

River water is a valuable resource for humanity, and it is used for a variety of functions such as drinking, traveling, bathing, aquaculture development, farming, and energy generation. As a result, a suitable level of water quality is necessary.The nature and scope of farming, industrial, and other anthropogenic activities in catchments greatly influence surface water quality in a location. Water quality control is becoming increasingly important. Predicting water quality ahead of time can considerably aid in water quality management in waterways. Various chemical measures such as chemical oxygen demand (COD), biochemical oxygen demand (BOD), temperature, dissolved oxygen (DO), pH, and conductivity can be used to assess stream water quality. Traditional techniques to water quality modeling include parameter-based statistical and deterministic models that require a great amount of knowledge about numerous hydrological sub-processes to provide the desired results. Furthermore, these models necessitate precisely calculated rate constants/coefficients for a variety of time and space-dependent hydrological, chemical, physical, and biological processes. Many of the parameters impacting water quality have a complex non-linear relationship with the input variables. As a result, traditional ways to solving this problem are no longer effective.

This paper will show how to use simple Machine learning algorithms to evaluate the river water quality of some chemical parameters. We are using UK river data, but these methods can also be used in Bangladeshi river water quality data if we get the data.

II. Literature Review

Poor quality water is expensive because resources must be allocated to upgrade water delivery infrastructure whenever an issue develops. The demand for enhanced water treatment and water quality management has been rising for these objectives to provide safe drinking water at competitive prices. To address these issues, systematic studies of raw water, disposal systems, and organizational monitoring issues are necessary [1]. Accurate forecasts of changes in water quality can significantly enhance aquaculture's efficiency.

Shafi et al. [2] utilized data from the Pakistan Council of Research in Water Resources (PCRWR) to identify water quality using support vector machines, neural networks (NNs), deep NNs, and K-nearest neighbors (KNNs). Researchers used a hybrid deep learning model to forecast water quality, including a long short-term memory (LSTM).

Solanki et al. [3] used a deep learning network model to evaluate and forecast the chemical eigenvalues of water, particularly dissolved oxygen and pH, which was claimed to produce more accurate findings than supervised learning-based approaches.

The long short-term memory (LSTM) network model was used by Liu et al. [4] to forecast drinking water quality in the Yangtze River Basin. pH, dissolved oxygen (DO), chemical oxygen demand (COD), and NH3-N were all employed in the LSTM model. The LSTM model is said to be promising for monitoring water quality.

To improve the neural networks backpropagation capacity to anticipate oxygen consumption in a lake, Yan et al. [5] proposed utilizing a genetic algorithm (GA) and a particle swarm optimization (PSO) approach. The results of the forecast were found to be more accurate.

Nikoo and Mahjouri [6] used a PSVM(Propbablistic Support Vector Machine) model in combination with a GIS technique to plan the classification and allocation of surface water in Iran. They stated that combining these two methods will result in credible data for water conservation feasibility studies. They also achieve a decent outcome using their own bespoke model.

May et al. [7] looked at the temporal history of water quality components to forecast values. They aimed to improve soft computing approaches in different water and environmental engineering in order to accurately analyze the time series of water quality parts and their inner connection.

Emamgholizadeh, et al. [8] used a multi - layer perceptron ,an adaptive neuro-fuzzy inference system, and radial basis network (RBF) for water quality components of the Karoon River. They claimed that while all of the models could predict groundwater elements, the MLP system was somewhat more accurate.

Furthermore, when compared to standard approaches, deep learning methods outperformed traditional methods in predict-

ing the WQ. Marir et al. [9] created a model for detecting unusual behavior in large-scale network traffic data. A SVM multiplayer Ensemble model was utilized for classification, while a deep learning technique was used for feature extraction. They merged these both and make a hybrid model which performs better then a normal model.

Sambito et al. [10] created an intelligent system based on the Internet of Things and a Bayesian decision network (BDN) for forecasting wastewater. The suggested method was utilized to forecast the water quality WQ of groundwater by focusing on analyses of soluble conservative contaminants such as metals, decision support systems, and auto-regressive moving averages.These paper done something different then the others which might help us in our work.

Heddam et al. [11] employed artificial neural networks to evaluate water quality components in a number of experiments. Artificial intelligence techniques, he argues, are suitable for analysis and forecasting the internal link between ground water components, as well as predicting their time series. In their research, they have used a variety of examples.

According to the literature, water quality evaluation and prediction are critical components of building water conservation initiatives, and artificial intelligence approaches have been proposed to help with this. Machine Learning and Deep Learning methods are pretty popular and gives higher accuracy then old classic methods.

III. Methodology

As our dataset is in tabular format, we need to preprocess our data before creating a model and training it. For preprocessing, we will use the following preprocessing methods [12]:
1. NULL value Identifying
2. Replace NULL values with Average MEAN
3. MIN-MAX Scaling
4. Label or One Hot Encoder

For our classification model, we have chosen the following classification models [13]:
1. Logistic Regression [14]
2. Decision Tree [15]
3. Random Forest
4. Support Vector Machine [16]

IV. Data Set

There were many datasets on the data.gov.uk website. From there, we have chosen our data from the Department of Environment, Food and Rural Affairs section. Our dataset set entitles "River Water Quality" taken from here [17].We have chosen this dataset because there are not many work done on these dataset. This dataset contains River Water Quality (Regions), including East, London, North East, East Midlands, North West, South West, South East, Wales, Yorkshire-Humber , West Midlands and RDA Summary. We are going to work on any of these area's river water quality. Our data-set is in .csv format and its a tabular data.There are 21 features(column) and 1640 individual data(row) in our data set. Our Initial Features of the dataset are shown in Fig.1.

```
RangeIndex: 1640 entries, 0 to 1639
Data columns (total 22 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   DistrictID  1640 non-null   int64
 1   District    1640 non-null   object
 2   Year        1640 non-null   int64
 3   ReportType  1640 non-null   object
 4   Akm         1640 non-null   float64
 5   Bkm         1640 non-null   float64
 6   Ckm         1640 non-null   float64
 7   Dkm         1640 non-null   float64
 8   Ekm         1640 non-null   float64
 9   Fkm         1640 non-null   float64
 10  Totalkm     1640 non-null   float64
 11  Apc         1638 non-null   float64
 12  Bpc         1638 non-null   float64
 13  Cpc         1638 non-null   float64
 14  Dpc         1638 non-null   float64
 15  Epc         1638 non-null   float64
 16  Fpc         1638 non-null   float64
 17  GOODpc      920 non-null    float64
 18  FAIRpc      920 non-null    float64
 19  POORpc      920 non-null    float64
 20  BADpc       920 non-null    float64
 21  Highpc      718 non-null    float64
dtypes: float64(18), int64(2), object(2)
```

Fig. 1. Dataset Features Information

From these features, our target class or Predicting class will be ReportType. There are four classes in our target class, and they are:
1. Chemistry
2. Nitrate
3. Phosphate
4. Biology

As we have more than two classes, we are going to make a multi-class classification model.

Fig.2 shows the total number of null values in each column. There are 720 missing values for GOODpc, FAIRpc, POORpc, BADpc, and 922 for the HIGHpc column.

Fig.3 shows the average mean value for each feature. For example, the average value for GOODpc and FAIRpc are 47.14 and 45.09. If we want to replace the null values of our dataset, we can use these mean values for that particular column.

Fig.4 shows the number of data in our four predicting classes. Our dataset has 600 data for chemistry class, 360 for nitrates and phosphate, and 320 for biology class.

Fig.5 shows that how much data has been collected From 1993 to 2006.

DistrictID	0
District	0
Year	0
ReportType	0
Akm	0
Bkm	0
Ckm	0
Dkm	0
Ekm	0
Fkm	0
Totalkm	0
Apc	2
Bpc	2
Cpc	2
Dpc	2
Epc	2
Fpc	2
GOODpc	720
FAIRpc	720
POORpc	720
BADpc	720
Highpc	922

Fig. 2. Number of Null value in dataset individual column

DistrictID	20.500000
Year	2000.000000
Akm	7.730000
Bkm	19.374878
Ckm	18.713110
Dkm	11.956402
Ekm	15.025976
Fkm	11.086524
Totalkm	83.885793
Apc	7.822833
Bpc	22.336508
Cpc	23.108913
Dpc	15.301893
Epc	18.187424
Fpc	13.242796
GOODpc	47.146413
FAIRpc	45.099565
POORpc	6.967065
BADpc	0.789130
Highpc	69.845682

Fig. 3. Average Mean Value of each column

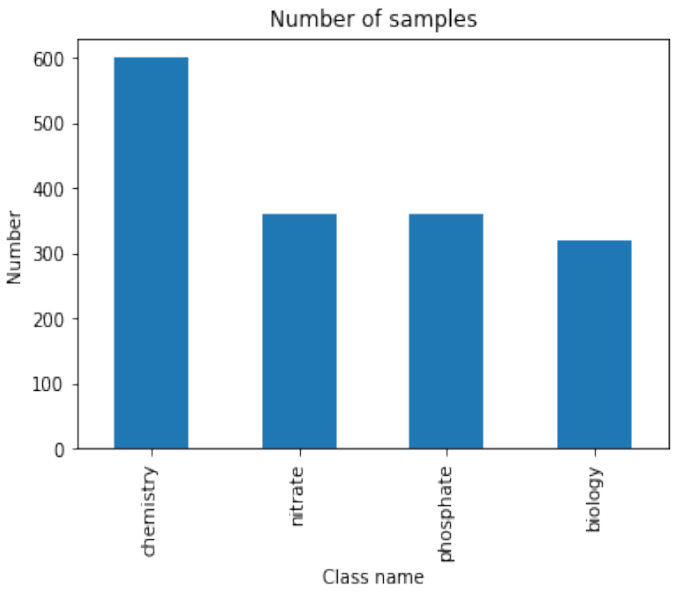

Fig. 4. Number of samples in each predicting class

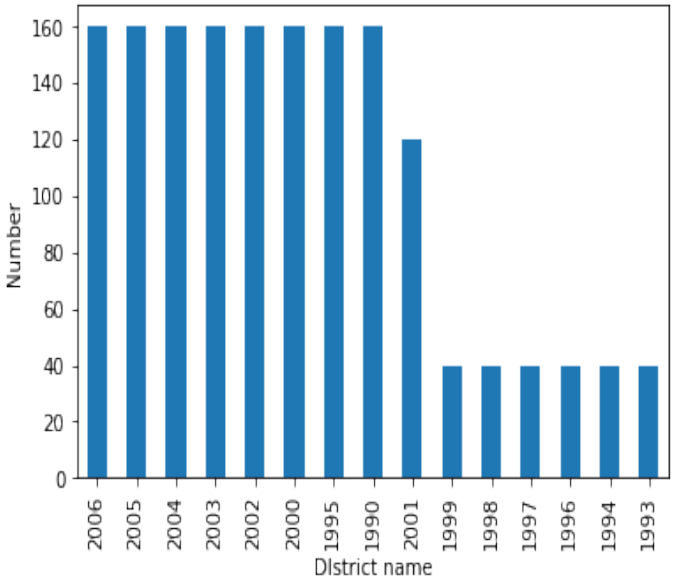

Fig. 5. Number of data by year

V. Result and Analysis

In this section, we will show the analysis of our dataset and the result we have got. Also, for predicting the river's quality, we have trained a model we random forest classifier. Fig.6 shows the re-plotting between the Akm and Bkm columns of our dataset regarding our predicting class. Here each dot point represents data, and the color is used to distinguish between the class. In Fig.7, a more clear rel-plot has been shown where each class has been plotted individually.

Fig.8 shows a regression plotting where all the data has been plotted in a 2-dimensional plane and with an average regression line of the data. Fig.9 shows a correlation matrix between our features(column). From a correlation matrix, we can analyze which features are identical to each other. If a particular feature matches another feature, we can add both of these features and make a new one. By doing this, we can reduce unnecessary features.

As we have done enough data analysis before training the model, we have done some data preprocessing, and one of

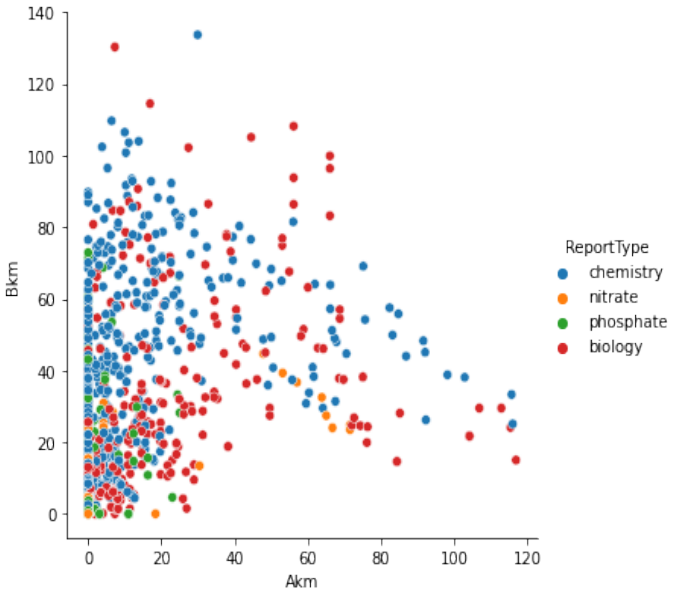

Fig. 6. Akm vs Bkm Data distribution Visualization using rel-plotting

Fig. 7. Akm vs Bkm Each class data Distribution using rel-plotting

them is null value replacement. Fig.2 shows a great amount of null value, so all the null values were replaced by the mean value of Fig.3.After that, we have split our data into 80:20. 80% for training and 20% for testing purposes. The parameters that have is used are shown in Fig.10. Using these parameters, we have trained our model.

After training, we have tested our model with our testing data and found 88% accuracy. We have also found many other accuracy matrices such as Precision, Recall, F1-Score. All the

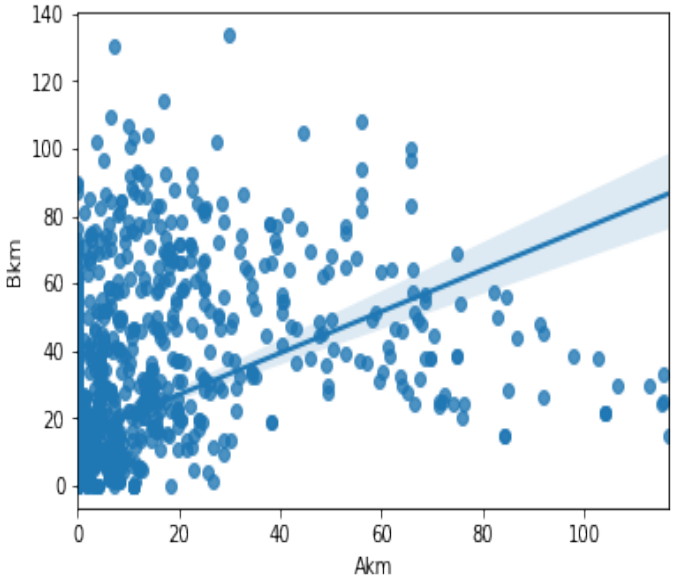

Fig. 8. Akm vs Bkm Regression Plot

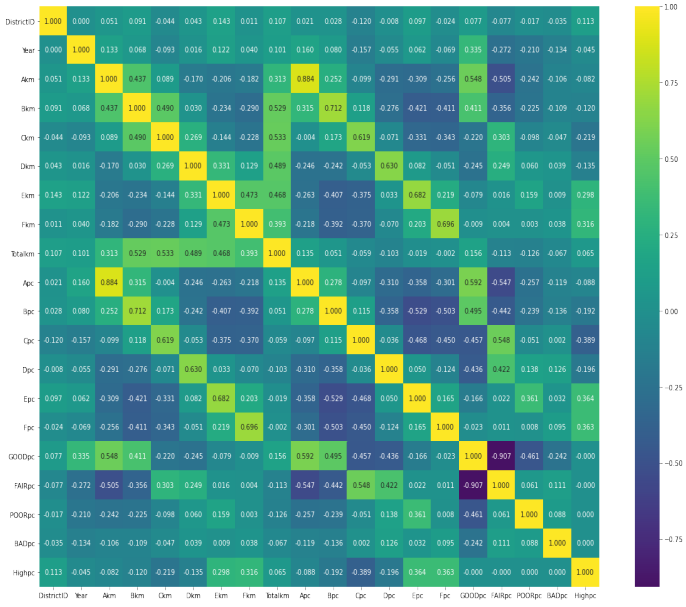

Fig. 9. Correlation Matrix of our dataset

```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=None, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=100,
                       n_jobs=None, oob_score=False, random_state=None,
                       verbose=0, warm_start=False)
```

Fig. 10. Random Forest Parameters

	precision	recall	f1-score	support
biology	0.86	0.65	0.74	78
chemistry	0.84	0.95	0.89	148
nitrate	0.97	0.89	0.92	97
phosphate	0.88	0.97	0.92	87
accuracy			0.88	410
macro avg	0.89	0.86	0.87	410
weighted avg	0.88	0.88	0.88	410

Fig. 11. Accuracy Matrices

values of the accuracy metrics are near 88%, so that means we didn't get any biased prediction. In Fig.11, all the accuracy metrics result has been shown.

CONCLUSION

REFERENCES

[1] Hu, Z.; Zhang, Y.; Zhao, Y.; Xie, M.; Zhong, J.; Tu, Z.; Liu, J. A Water Quality Prediction Method Based on the Deep LSTM Network Considering Correlation in Smart Mariculture. Sensors 2019, 19, 1420.

[2] Shafi, U.; Mumtaz, R.; Anwar, H.; Qamar, A.M.; Khurshid, H. Surface Water Pollution Detection using Internet of Things. In Proceedings of the 2018 15th International Conference on Smart Cities: Improving Quality of Life Using ICT IoT (HONET-ICT), Islamabad, Pakistan, 8–10 October 2018; pp. 92–96.

[3] A. Solanki, H. Agrawal, and K. Khare, "Predictive analysis of water quality parameters using deep learning," International Journal of Computers and Applications, vol. 125, no. 9, pp. 29–34, 2015.

[4] Liu, P.; Wang, J.; Sangaiah, A.K.; Xie, Y.; Yin, X. Analysis and Prediction of Water Quality Using LSTM Deep Neural Networks in IoT Environment. Sustainability 2019, 11, 2058.

[5] J. Yan, Z. Xu, Y. Yu, H. Xu, and K. Gao, "Application of a hybrid optimized BP network model to estimate water quality parameters of Beihai Lake in Beijing," Applied Sciences, vol. 9, no. 9, p. 1863, 2019.

[6] Nikoo, M. R. Mahjouri, N. 2013 Water quality zoning using probabilistic support vector machines and self-organizing maps. Water Resour. Manage. 27 (7), 2577–2594.

[7] May, R. J., Dandy, G. C., Maier, H. R. Nixon, J. B. 2008 Application of partial mutual information variable selection to ANN forecasting of water quality in water distribution systems. Environ. Model. Softw. 23 (10–11), 1289–1299.

[8] Emamgholizadeh, S., Kashi, H., Marofpoor, I. Zalaghi, E. 2013 Prediction of water quality parameters of Karoon River (Iran) by artificial intelligence-based models. Int. J. Environ. Sci. Technol. 11 (3), 645–656.

[9] 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.

[10] Sambito, M.; Di Cristo, C.; Freni, G.; Leopardi, A. Optimal water quality sensor positioning in urban drainage systems for illicit intrusion identification. J. Hydroinform. 2020, 22, 46–60.

[11] Heddam, S. 2016a Generalized regression neural network based approach as a new tool for predicting total dissolved gas (TDG) downstream of spillways of dams: a case study of Columbia River Basin Dams, USA. Environ. Process. 4 (1), 235–253.

[12] https://scikit-learn.org/stable/modules/preprocessing.htmlpreprocessing

[13] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[14] https://scikit-learn.org/stable/modules/linear_model.htmllogistic-regression

[15] https://scikit-learn.org/stable/modules/tree.html

[16] https://scikit-learn.org/stable/modules/svm.htmlclassification

[17] https://data.gov.uk/dataset/b33499d4-6111-4aaf-a331-82af3bc83789/river-water-quality-regions