Water Quality Prediction

Jenil Doshi 202118034
Rachit Kumar Singh 202118015

April 2022

1 Abtract

Water is an important and essential element for the life on earth. Due to the growth of population and industrialization the water resources become more polluted. Waste disposal from industry, human wastes, automobile wastes, agricultural runoff from farmlands containing chemical factors, unwanted nutrients, and other wastes from point and non-point source flow to water bodies, which affects the quality of the water resources. etc. The increase in pollution influences the quantity and quality of water, which results high risk on health and other issues for human as well as for living organisms on the planet. Hence, evaluating and monitoring the quality of water, and its prediction become crucial and applicable area for research in the current scenario. In various researchers they have used traditional approaches; Now, they are using technologies like machine learning, big data analytics for evaluation and prediction of water quality. The advanced big data implementation using sensor networks and machine learning with the data related to environment, aids in building water quality prediction models. This paper analyses various prediction models developed using machine learning and big data techniques and their experimental results of water prediction and evaluation. Various challenges and issues are reviewed and possible solutions to some research issues are proposed.

2 Introduction

Water is the most significant resource of life, crucial for supporting the life of most existing creatures and human beings. Living organisms need water with enough quality to continue their lives. There are certain limits of pollutions that water species can tolerate. Exceeding these limits affects the existence of these creatures and threatens their lives.

Most ambient water bodies such as rivers, lakes, and streams have specific quality standards that indicate their quality. Moreover, water specifications for other applications/usages possess their standards. For example, irrigation water must be neither too saline nor contain toxic materials that can be transferred to plants or soil and thus destroying the ecosystems. Water quality for industrial uses also requires different properties based on the specific industrial processes. Some of the low-priced resources of fresh water, such as ground and surface wa-

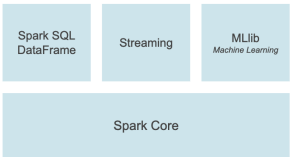
ter, are natural water resources. However, such resources can be polluted by human/industrial activities and other natural processes.

3 Dataset

The dataset used in this study is collected from certain historical locations in India. It contained 534 samples from different Indian states during the period from 2005 to 2014. The dataset has 7 significant parameters, namely, dissolved oxygen (DO), pH, conductivity, biological oxygen demand (BOD), nitrate, fecal coliform, and total coliform. Data was collected by the Indian government to ensure the quality of the supplied drinking water. This dataset was obtained from Kaggle https://www.kaggle.com/anbarivan/indi water-quality-data.

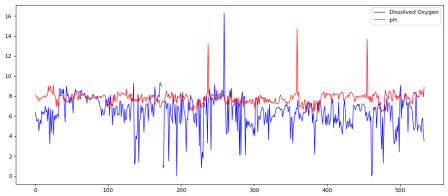
4 Pyspark

PySpark is an interface for Apache Spark in Python. It not only allows you to write Spark applications using Python APIs, but also provides the PySpark shell for interactively analyzing your data in a distributed environment. PySpark supports most of Spark's features such as Spark SQL, DataFrame, Streaming, MLlib (Machine Learning) and Spark Core.

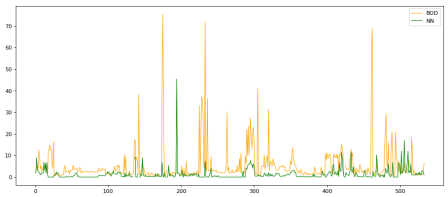


Pyspark

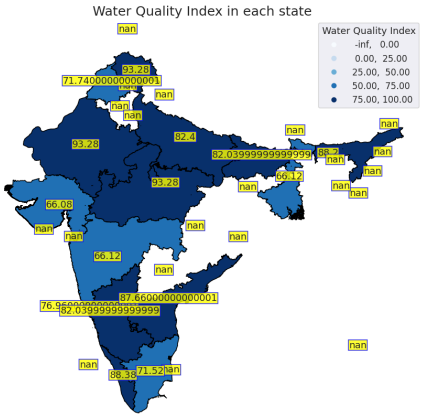
5 EDA



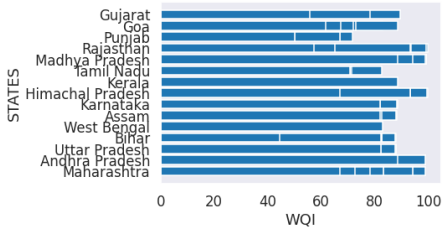
Quant present in Water(Dissolved O2)



Quantities present in Water(Ph)



Water Quality Index

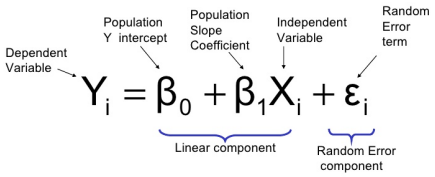


WQI State Wise

6 Linear Regression

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable. Linear regression fits a straight line or surface that minimizes the discrepancies between predicted and actual output values. There are simple linear regression calculators that use a "least squares" method to discover the best-fit line for a set of paired data. You then estimate the value of X (dependent variable) from Y (independent variable).



Linear Regression Equation

Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions. Linear regression can be applied to various areas in business and academic study.

You'll find that linear regression is used in everything from biological, behavioral, environmental and social sciences to business. Linear-regression models have become a proven way to scientifically and reliably predict the future. Because linear regression is a long-established statistical procedure, the properties of linear-regression models are well understood and can be trained very quickly.

```
[ ] predictions.select("wqi","prediction").show()
```

```
+-----------------+-----------------+
|              wqi|       prediction|
+-----------------+-----------------+
|82.03999999999999|82.00841079130217|
|             82.4|  81.759200637649|
|            66.12|67.61802046976379|
|            66.12|67.61802046976379|
|            66.12|67.61802046976379|
|            66.12|67.61802046976379|
|             82.4|  81.759200637649|
|             82.4|  81.759200637649|
|77.36000000000001|77.75891522197496|
|77.36000000000001|77.75891522197496|
|            66.12|67.61802046976379|
|82.03999999999999|82.00841079130217|
|            66.12|67.61802046976379|
|             82.4|  81.759200637649|
|             82.4|  81.759200637649|
|            66.12|67.61802046976379|
|93.82000000000001|91.25168377617582|
|            77.36|78.22462462456501|
|            82.98|83.01095558702033|
|             82.4|  81.759200637649|
+-----------------+-----------------+
only showing top 20 rows
```

Prediction of first 20 rows

```
Now we check the performance of our model.
```

```
[ ] model.stages[2].summary.r2
```

```
0.9810973586151036
```

Model Accuracy

7 Conclusion

Water is the most essential natural resource for humans to survive. But due to the ill habits of humans only, it gets impurified. And this impurified water when consumed has many ill effects on human body.

The first step to avoid the impure water is to know whether the water is contaminated or not.In this notebook we have tried to categrize the Water Quality Index of India which can be useful

to take actions to clean the water.

8 Future Work

In future, certain featrues like categorizing the water types at different levels can be used in order to dive deeper in this problem. We can create an application stating the water quality index of each state and the methods to prevent this by motivating the people to make their city, state and therefore the Country cleaner.Awareness is a big issue in India, so this can be used to create an awareness among the people.

9 References

1. Y. Khan and C. S. See, "Predicting and analyzing water quality using Machine Learning: A comprehensive model," 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT), 2016, pp. 1-6, doi: 10.1109/LISAT.2016.7494106.

2. J. P. Nair and M. S. Vijaya, "Predictive Models for River Water Quality using Machine Learning and Big Data Techniques - A Survey," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 1747-1753, doi: 10.1109/ICAIS50930.2021.9395832.

3. S. Cao, S. Wang and Y. Zhang, "Design of River Water Quality Assessment and Prediction Algorithm," 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018, pp. 901-906, doi: 10.1109/ICMLA.2018.00146.

4. S. Chatterjee, S. Sarkar, N. Dey, S. Sen, T. Goto and N. C. Debnath, "Water quality prediction: Multi objective genetic algorithm coupled artificial neural network based approach," 2017 IEEE 15th International Conference on Industrial Informatics (INDIN), 2017, pp. 963-968, doi: 10.1109/INDIN.2017.8104902.