CSE 545: Big Data Analytics Project Report
SDG 6 - Clean Water and Sanitation
Team Sparking Water

Murthy Kaja	Yashwanth Madaka	Kowshik Sola	Nithin Reddy
113278723	114353641	114352819	114778442

1. Introduction

SDG 6 aims to guarantee that everyone has access to safe drinking water and sanitation, emphasizing on the long-term management of water resources, wastewater, and ecosystems, as well as the importance of a supportive environment. Countries committed to thorough follow-up and review of progress toward the Goals and goals in the 2030 Agenda for Sustainable Development, using a set of global indicators**.** [1]

One of the most essential uses of water is for drinking and hygiene purposes within households. This use is captured in target 6.1, which seeks to secure safe and affordable drinking water for all. Target 6.3 sets out to improve ambient water quality, which is essential to protecting both ecosystem health and human health, by eliminating, minimizing and significantly reducing different streams of pollution into water bodies. Our work uses different countries' data to predict water quality indexes and find similar countries with similar trends like economic factors and population which are directly related to water quality.

2. Background

Hundreds of millions of people around the world still lack access to safe drinking water, sanitation, and hygiene services, which are crucial for human health and preventing the spread of the COVID-19 virus. Global water consumption has expanded at a rate more than twice that of population growth over the last century. Aside from water scarcity induced by climate change, governments and territories are experiencing increasing issues connected to water pollution, damaged water-related ecosystems, and a lack of collaboration on transboundary waters. The world is falling short of achieving Goal 6. Current rates of advancement, as well as integrated and holistic approaches to water management, must be dramatically accelerated.[2]

UN established SDG 6 in an effort to make adequate sanitation and water services available to all people by the year 2030. As many as 800 million people, or more, would require the construction of facilities to provide consistent clean water and waste removal. To succeed in their vision, the UN developed a series of targets. These targets include restoring and protecting river ecosystems throughout the world, eliminating sources of water pollution, and increasing international cooperation to bring services throughout the world.To Satisfy this goal, we use developed country's features to be implemented in developing countries.[4]

3. Data

To build our idea, we collected data from different country's data for training. Our main source of data is WATERBASE[1] and GEMSTAT[2]. Waterbase contains time series data of different nutrients, chemicals and organic matter in the lakes, rivers and groundwater bodies present all over Europe. It has comprehensive information from 1941 to 2018. Another portal called GEMSTAT is a global freshwater dataset which has similar information for almost all the countries in the world. The total aggregated dataset is around 12 GB. So our aggregated data contains features like Station id, location,

date, Iron, Nitrate, Chloride, Sodium, Dissolved_oxygen, Oxygen_saturation, Water_temperature, Total_suspended_solids, Conductivity, Phosphate, and pH are the most important features.

Continent/Country Name	From Year	To Year
Europe	1941	2018
Africa	1960	2020
India	1971	2008

Table 1. Years of data for Europe, Africa, India

4. Methods

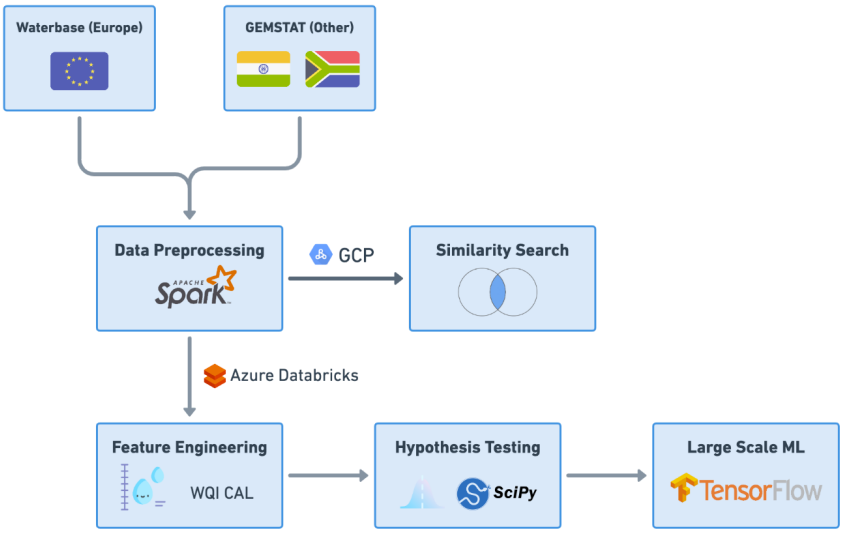

Figure 1. Data Pipeline

4.1. Data Preprocessing

As we cannot process raw data, Data Preprocessing is done for both Waterbase[3] and Gemstat[4] datasets which initially had disaggregated data in various metrics and further we have aggregated data from both datasets using techniques like Spark RDD, spark data frames and hive query language. To process this data we have deleted timestamp and filtered required columns like chemical names, station id and date from the dataset. We have then filtered rows containing a list of predetermined chemicals which are useful for calculating WQI. We also need to convert all the values to a single unit of measure for a given chemical which is achieved using spark dataframe. Missing values are then inserted by calculating the average of particular data by using reduceByKey wide transformation. After transformation, we have converted the required chemicals in the columnar format to row format for large scale machine learning. This entire process is performed using spark RDDs and later to fill the missing values in a month we have used spark data frame and hive query. Further, we have mapped the latitude and longitude of the different station IDs to the country name using GeoPy API. Further, we have joined the location data and chemical data to eliminate station id. Finally, we have aggregated the data using the reduceByKey transformation.

4.2. Feature Engineering

Following the aim to be learned and the machine learning model utilized, feature engineering is the 'skill' of creating relevant features from existing data. After exploring papers on water quality[8], we are referring to a fixed formula for the calculation of the Water Quality Index(WQI). To implement this feature we have calculated the SWQI[5] by using factors and formulated it as

$$WQI = TEMP * (BOD + TSS + DO + COND)$$

where TEMP, BOD, TSS, DO and COND represent individual index terms with different weighting factors for each parameter. Temp = water temperature index, BOD = Biological Oxygen Index, TSS = Total suspended Index, DO = Dissolved Oxygen Index, COND = Conductivity Index. WQI Range (91-100 = Excellent, 71-90 = Good, 51-70 = Average, 26-50 = Fair, 0-25 = Poor). After calculating it, the below is the heatmap of WQI indexes in the year 2008.

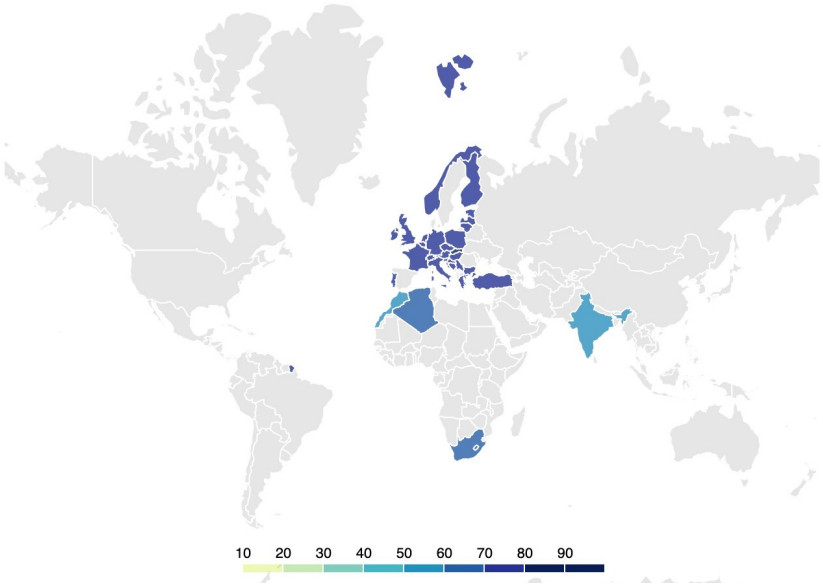

Figure 2. Heat Map of WQI ranges for Africa, Europe and India

4.3. Similarity Search

After getting the CSV data from the above-mentioned data preprocessing step, similarity search was performed to find similar countries. Grouping similar countries w.r.t their years can help policymakers and organizations take prior actions. Developing countries can follow the footsteps of developed countries present in their group and can adapt themselves to improve their water quality. Apart from the chemicals used in WQI calculation, we also considered the Environmental factors (CO2 Emissions, Air pollution), Economical (GDP (Gross Domestic Product), GNI (Gross National Income)) and Societal factors (Population, Life Expectancy, HDI (Human Development Index)) of all the countries in our dataset. Below is the pipeline flow we followed to perform similarity search.

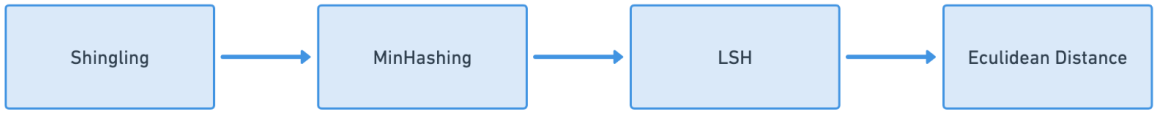

Figure 3. Similarity Search Pipeline

We took the key as (Country, Year) and our vector contains values of all chemicals, GDP, GNI, population, air pollution, and life expectancy. We used locality sensitive hashing to first calculate groups of similar countries. Later, Euclidean distance was used as the distance metric to calculate distance between the original vectors of the countries. Euclidean distance was used as a distance metric here, as we want to see how close any 2 country vectors are.

4.4. Hypothesis Testing

Feature selection is an important part of building any machine learning model. A model performs well when suited and correlated features are provided as input. As we are predicting Water Quality Index (WQI), we are finding its cosine similarity between other features like Iron, Phosphate, and other chemicals in the following table. As we are calculating the cosine similarity between each WQI and chemicals across countries, we are multiplying the p-value with Bonferroni correction (α = number of countries i.e 58)

Features	Cosine Similarity	P-Value	Bonferroni corrected P-Value
Iron	0.005157232	0.00011953	0.006932954
Nitrate	-0.074133163	0.02922253	1.69490E-08
Chloride	-0.103132809	0.00095	0.045281067
Sodium	-0.044942523	0.02678175	1.55334E-08
Dissolved_oxygen	0.606046465	1.05E-09	6.06456E-08
Water_temperature	-0.171881863	6.47E-05	0.003754424
Total_suspended_solids	-0.006519594	7.73E-05	0.004481761
Conductivity	0.017448662	0.0020221	0.017281976
Phosphate	0.062399443	9.92E-05	0.005753125
pH	0.07648714	6.40E-10	3.71104E-08
Non_ionised_ammonia	-0.099626672	0.02091294	1.2129504

Table 2. Features and their Cosine Similarity, P-Value and Bonferroni corrected P-Value

We can see that the Bonferroni corrected value for **Non_ionised_ammonia** is greater than 0.05 which is our threshold, so we should not consider these chemicals as a feature for the model. We can also see that dissolved oxygen is highly positively correlated with WQI and its p-value is less than 0.05, thus this correlation can be considered.

4.5. Large Scale ML

We implemented large scale ML for this problem using LSTM (long short-term memory) Neural Network as it can process multiple sequences of data. The LSTM model consists of 1 LSTM layer with 100 units and two dense layers with 50 units and 1 units layer as output function. The activation function used is RELU (rectified linear activation function and the optimizer is Adam optimizer as it has faster computation time and fewer parameters for tuning. For regression we have used mean absolute error as loss function with epochs = 50 and batch_size = 16 and we have used 80% split for train and 20% as test data. We used three models each for one country and achieved good MAE values for measuring performance. Below are the results and graphs indicating the models performance.

5. Results

This section demonstrates the results from time series analysis and similarity search.

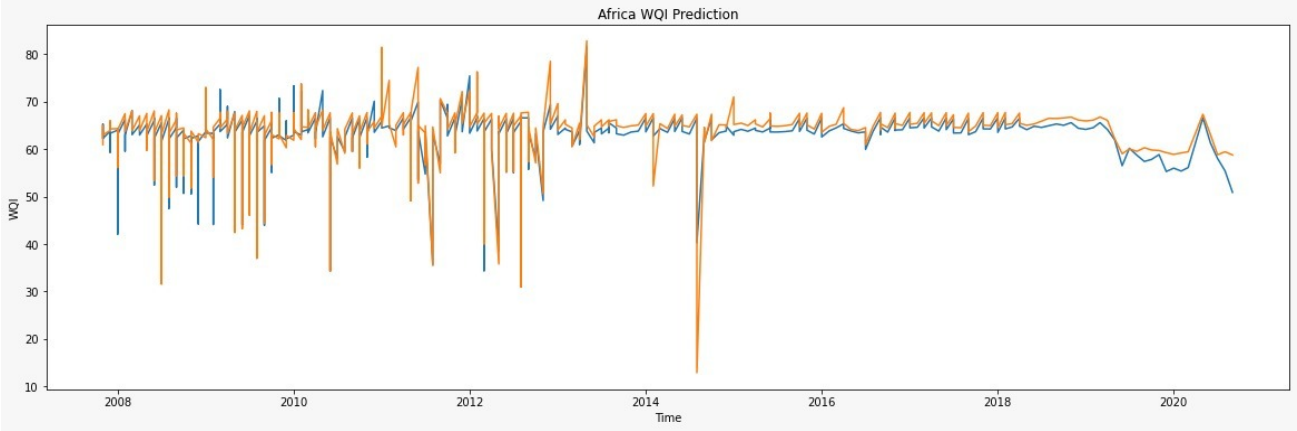

Figure 4. Africa WQI Prediction (Orange) vs Actual (Blue) from 2008 - 2020

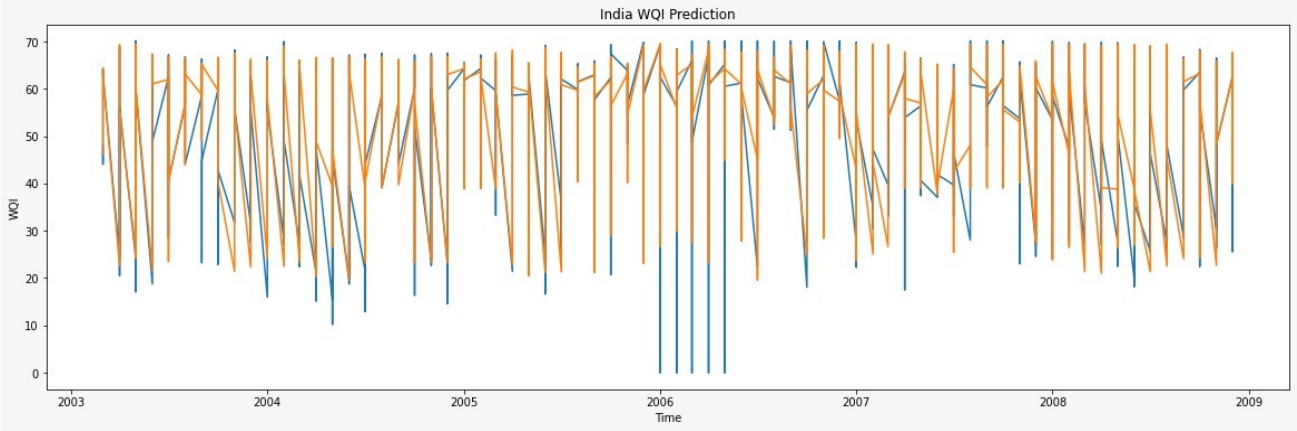

Figure 5. India WQI Prediction (Orange) vs Actual (Blue) from 2003 to 2009

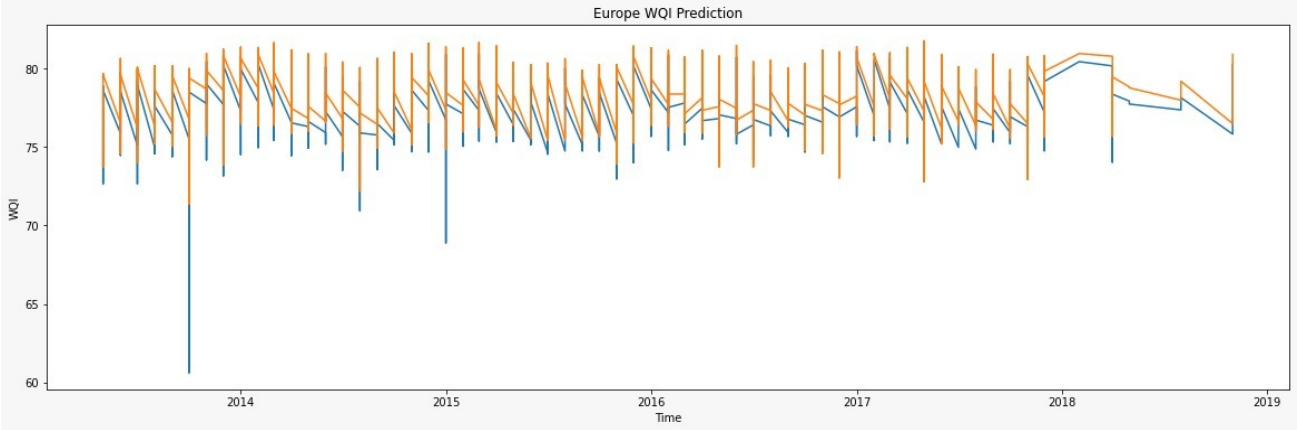

Figure 6. Europe WQI Prediction (Orange) vs Actual (Blue) from 2014 to 2019,

Figure 4, Figure 5 and Figure 6 demonstrate the successful performance of LSTM with a good match to the original WQI values. We observe the ability of capturing the spikes and this indicates the capacity of the model in capturing volatile periods.

Model	MAE Values
LSTM - Europe	0.84046
LSTM - Africa	1.05017
LSTM - India	0.88669

Table 3. Country Models and their Mean Absolute Error Values

Table 3 displays the performance of our models which were used for Europe, Africa and India. The performance metric which we used here is Mean Absolute Error (MAE).

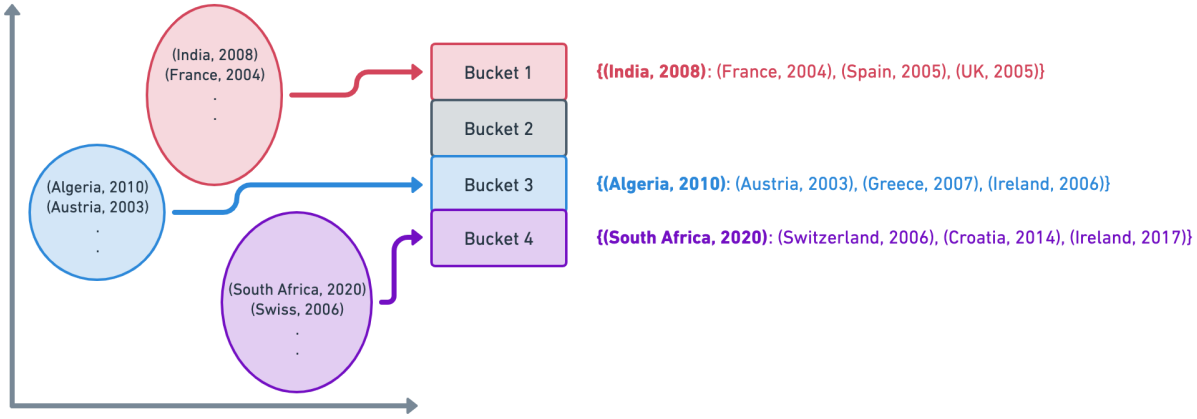

Figure 7. Similarity Search Grouping

We can see from the above clustering image that some developing countries had matched with the previous data of developed countries.

6. Conclusion

Due to their relative simplicity and readily applicable output, WQI models have been widely used for water quality assessment; nonetheless, several different versions have been developed to date. We are using different models for WQI prediction for a particular country. Using the predicted WQI values, authorities can take prior action to mitigate probable damage. For example, increase in WQI values can be because of the increase of oxygen levels at a certain location for an extended period of time and this can shoot up the concentration of algae, which pollutes the water. So the governing bodies can take action beforehand to control this algae formation. From similarity search results we can provide developed country measures that are heavily involved in maintaining good WQI levels to developing countries for their progress. This research also suggests that water in some countries should be treated before consumption. Higher values of iron, manganese, and arsenic reduces drinking water quality and awareness raising on chemical contents in drinking water at household level is required to improve public health. This study also stresses the importance of regular water quality monitoring and identifying the sources of water pollution to prevent additional contamination. This system could be implemented and optimized for assessment of different parameters representative of other types of water pollution, such as industrial effluents and drainage from agricultural areas, in addition to domestic sewage. We also can deploy the same techniques used for measuring other SDG goals like Air quality and report for progress.

7. References

[1] https://sdgs.un.org/goals/goal6
[2] https://www.sdg6data.org/
[3]Waterbase dataset - https://www.eea.europa.eu/data-and-maps/data/waterbase-water-quality-2
[4] Gemstat dataset - https://gemstat.bafg.de/applications/public.html?publicuser=PublicUser#gemstat/Stations
[5] SWQI Index Calculation - https://www.agry.purdue.edu/hydrology/projects/nexus-swm/en/Tools/WaterQualityCalculator.php
[6] Tool used for flowcharts - https://whimsical.com
[7] Icons taken from - https://www.flaticon.com/
[8] https://jhpn.biomedcentral.com/articles/10.1186/s41043-016-0041-5
[9] https://www.sciencedirect.com/science/article/pii/S1470160X20311572
[10] Aggregated Dataset - https://drive.google.com/drive/folders/1osIe34AULxTE4XrcLvT2eGJ-I9dsi4kc?usp=sharing