WQVP: An API enabled Open Data Machine Learning based Solution for Water Quality Visualization and Prediction

Pooja Lodhi, Omji Mishra, and Gagandeep Kaur
Deptt. of CSE&IT, Jaypee Institute of Information Technology
NOIDA, Uttar Pradesh
India
poolodhi@gmail.com

***Abstract*— Water is an essential component required by living bodies for their survival. In today's world, most of the water utilization is done by human beings. Due to this, there is a lot of adverse impact on water bodies. As human consumption of water increases, their pollution also increases. In order to control pollution impact and take measures to reduce water pollution, several methods have been proposed by researchers. Water Quality Index measures are one such method being adopted and used to measure harmful constituents of water. In recent times initiatives have been taken by international and national governing bodies to provide data through Open Data Initiatives that can be publicly made available. This data fetched in real time through APIs can be used for providing data analysis to naïve natives of the place with better understanding features like visualizations. Machine learning based techniques have proved to be a great tool for providing unsupervised learning in this area. We have implemented an API enabled Open Data Machine Learning based Solution for Water Quality Visualization and Prediction for Australian Rivers.**

***Index Terms*—API; Clustering; Machine Learning; Open Data Initiative; Prediction; Visualization; Water Quality**

I. INTRODUCTION

A. Open Government Data Initiative

These days Open Government Data (OGD) is gaining momentum in providing sharing of knowledge by making public data and information of governmental bodies freely available to private citizens in system processable formats so as to reuse it for mutual benefits. OGD is global movement and has its roots in the initiative started in 2009 by US President as Memorandum on Transparency and Open Government requiring providing transparency in government projects and collaborations through sharing of data by public administration and industry to private citizens. The number of countries that have agreed to provide OGD to its citizens has crossed 70, and more and more countries are understanding its need and joining it. The Indian government also has joined this initiative and provides free access to the data for development of applications etc. so as to be able to re-use the information for mutual growth of industry and government. 'Open Data' is the raw data made available by governments, the industry as well as NGOs, Scientific institutions, educational organizations, NPOs, etc. and as such is not individual's property. The growth in the field of Open Data surely asks for new tools and techniques that can support it.

B. Tools for Data Sharing

Digital transformation needs companies to look out for new tools and techniques so as to be able to support the increasing need for faster delivery of services at a large number of delivery points. Technologies like SaaS, mobile, and Internet of Things are gaining grounds in providing an increase in endpoints and thus enabling the success of Open Data initiatives.

Moreover, the frequency at which the applications and user needs are increasing asks for I.T personnel to search, innovate and develop tools that could support such tremendous upgrades. Applications Programming Interface (API)s are first and foremost in this regard. Works in the fields of Open Data are being carried out through 'Open APIs'. Open APIs provide methods so as entities can share data in trusted, timely and open format. Open APIs allow interaction between consumer and developers at one end and internal data service providers and developers at another end. It is, however, important to understand the distinction between API and Open APIs

One of the areas where open government initiatives for Open Data based on APIs is applicable is its use for providing water pollution information to natives of the place.

C. Water Pollution Control

Water is an essential component required by living bodies for their survival. Nature has provided us water in abundance yet a large portion of it is unsuitable for drinking. Generally speaking, water covers 71% of earth surface. In that only 2.5% contains freshwater in the form of lakes, rivers, and groundwater. The surface water consists of less than 0.01% as lakes and rivers [1]. According to an estimate, more than half of the world population will be facing the water-based problems by 2025. Another estimate suggests that water demand will increase by 50% by 2030. Such low concentration of fresh water reminds judicial use of this precious resource.

In today's world, most of the water utilization is done by human being be it for drinking, washing, bathing or cooking or for any industrial purpose. Due to this, there is a lot of adverse impact on water bodies. As human consumption of water increases, their pollution also increases. After the industrial revolution, humans started polluting rivers and other water bodies by dumping the toxic waste from industries into them. This led to the need for methods and

strategies that can reduce water contamination. In order to measure adulteration, several methods have been proposed by intellectuals.

So in order to focus on this serious issue, we have tried to provide a water quality measurement tool that can predict water qualities of river bodies. The analysis of the data has been done on the basis of several parameters like pH, DO, Salinity, Temperature, Turbidity, etc.

Machine learning is a technique used by researchers for automated analysis of data and building models. The iterative behavior of machine learning algorithms makes it highly useful in applications that are exposed to new real-time data by providing quick adaptability to changes. In machine learning, clustering is a process in which dataset is grouped into different clusters such that same type of data belongs to the same group. The various algorithms used for clustering are Hierarchical, K-Means, K-Medoids, DBSCAN, ANN and many more.

Through this project, we have tried to analyze the Dataset of various rivers of Australia and also used various other datasets for analysis of the reason behind the change in water quality around these rivers. The dataset has been taken from Queensland Government data [2]. The analysis of the data will have been done on the basis of several parameters like pH, DO, Salinity, Temperature, Turbidity, etc. The aim was to do a thorough analysis and visualize the dataset, and then prediction on the basis of the previous dataset was done to depict future events.

II. Related Work

In research paper [3], the authors B. Srivastava, et. al. have proposed their own dataset of water pollution collected from different sources like lab results, real-time sensors and estimates from people using mobile apps. They have taken pH, electrical conductivity, dissolved oxygen (DO) and turbidity as their measuring parameters. They have also launched *Neer Bandhu*, a mobile app for collecting pictures of water pollution [3]. In order to do the study, the authors have also released another app called *Ganga Watch* which uses public API to explore data. The dataset is also available via API named *Blue Water*. They have done an analysis of Ganga water quality during Ardh Kumbh 2016, a religious event held in Haridwar on the bank of river Ganga. They have measured the quality of water before and after the events at different ghats. They have used K-means clustering to differentiate between regions having good water quality and poor water quality. Different heat maps have been plotted, and same colors in the heat maps indicate similar water quality. The limitation of their work is that they have selected only four parameters out of more than 30 parameters recommended by CPCB[3].

In research paper [4], the authors S. Emamgholizadehet. al. have used many machine learning techniques like Multi-Layer Perceptron (MLP) model, Radial Basis Network and Adaptive Neuro-Fuzzy Inference System (ANFIS) to calculate parameters like Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD) and Chemical Oxygen Demand (COD) for an Iranian Karoon river. The authors have discussed Artificial Neural Network (ANN), and its characteristics like it work well on a large volume of data in the paper. The two kinds of ANN namely Back Propagation, and Radial Based Neural Networks have been discussed. Another algorithm named Adaptive Neuro-Fuzzy Inference System (ANFIS) is also discussed in detail. The dataset was analyzed over these models on the basis of Root Mean Square Error (RMSE), Coefficient of Determination (R2) and Mean Absolute Error (MEA). Models have analyzed over nine input variables. These were EC, pH, Turbidity, Ca, Mg, Na, PO4, NO2, NO3. For training and a testing dataset of 17 years of the river was taken. After experimentation, the computed value of DO, BOD, COD were found to be similar in both the models. The demerit of their work is that the authors have not described how they were accessing the dataset. Rather than doing analysis on real-time streaming data, all the observations ware performed on static dataset [4].

The research paper [5] provides a method for checking the water quality using the Bayesian algorithm. Here classification is done on the basis that whether water is suitable for drinking purpose or not. The data was first collected from sensors, and then it was sent to water quality predictor. The dataset consisted of 100 samples collected from 6 municipalities of Government of Tamil Nadu. Also, the implementation of the project has been done in JAVA using NetBeans IDE. The prediction results show that proposed method behaves similarly to that of results obtained from traditional methods. The shortcomings of this research work are that the dataset used for doing the experimentation is very small. Also, only five attributes have been selected for making water quality prediction. Also, no justification has been provided for the attributes that make the water unsuitable for drinking [5].

The research work of S. Y. Muhammad et al. in [6] provide an analysis and comparison of different classification models for water quality. In this paper, authors have compared classification models on the basis of different features, e.g. latitude, longitude, color, time, weather, DO, BOD, WQI, pH, Turbidity, Calcium, Iron, Lead, Chlorine, etc. that play a significant role in water quality. The experiment was done on the river Kinta River, Perak Malaysia. The dataset for the research work was taken from ESERI in University of Sultan Zainal Abidin (UniSZA), Malaysia. The content of dataset contains a record of four year from 2002 to 2006. The number of instances in the dataset was 135 and numbers of attribute were 54. From their work, they concluded that the Lazy Model using K Star Algorithm was the best classification model with highest accuracy percentage. The weakness of this work lies in the size of the dataset. Although dataset has large numbers of attributes, but the size is of only 135 instances. Also, the dataset is static and no API has been used to access the data.

The works of M. A. Dota et al. in [7] present comparative analysis of different classification algorithms on data collected from soil-contaminated water. In this experiment, work has been divided into four parts. In the first part, the scenario for water contamination by soil was created in a lab. In the second part, variables were defined on which evaluation was done. These were temperature, pH, pHmV, ORP, DO, conductivity, TDS, and Salinity. In the third part, the continuous rotation was done at specific rpm and a 1mL dosage was added at each 240sec interval and the values were recorded. In the final part, different algorithms were applied on the data collected through sensors. In the experiment, the total data was of 5100 readings. WEKA tool was to implement various algorithms. The algorithms on which data was applied were: Best-First Decision Tree

Classifier– BFTree, Functional Trees – FT, Naïve Bayes Decision Tree– NBTree, Grafted C4.5 Decision Tree– J48graft, C4.5 Decision Tree– J48 and LADTree. Two types of experiments were performed on the obtained data. In the first experiment, data was divided into two set. First set as a training set of 3400 readings and second set was test set of 1700 readings. In the second experiment, k-fold cross validation (where k is 10) was applied on the whole dataset. The model was trained with 9 training set and 1 testing set. The different classes that were used for classification vary from Excellent to Very Awful. The results of the above experiments showed that the classification proposed is rational with the category and their objects. The algorithms that better depicted the data were BFTree, J48graft and J48. The disadvantage of their work is that instead of creating soil samples in lab they could have samples from agricultural fields and ponds for more realistic experiment [7].

In the proposed work of Xiaoyun Fan et al. in [8] Principal Component Analysis (PCA) and Cluster Analysis (CA) were used to identify the features of water quality. They were also used to evaluate the water quality spatial pattern. The analysis of the water quality was done on Pearl River Delta (PRD) located in Southern China. The parameters that were used for analysis were Dissolved Oxygen (DO), Chemical Oxygen Demand (COD), Biological Oxygen Demand (BOD), Total Phosphorus (TP), Ammonia Nitrogen (NH3-N) , Mercury (Hg) and Oil. The river was divided into a large number of monitoring stations in Northern, Eastern and Western region. The monitoring stations of the Northern and Western region were divided into four clusters while monitoring stations of the Eastern region were divided into three clusters. According to the author, PCA and CA methods are useful for evaluating water quality and judicial use of water resources. The demerit of this research work is that here the size of the dataset is not specified. Also how data is collected, whether it is real-time or not is also not clear. Also, the experiment has been performed specifically during the dry season, so results for other seasons may vary.

The research work of A. Barakat et al. in [9] assesses the water quality variations of *Oum Er-Rbia* river and its tributaries. The dataset for the water quality was collected from fourteen monitoring stations for the period of 12 years. The parameters that were used for the study were Temperature, pH, Turbidity, Total Suspended Solids (TSS), Conductivity, Ammonia (NH3), Nitrate, Dissolved Oxygen (DO), Total phosphorus (TP), Biological Oxygen Demand (BOD), Chemical Oxygen Demand (COD) and Fecal Coliforms. In order to do the analysis, Pearson's correlation, PCA, and CA multivariate methods were used to identify similarities and differences between the monitoring stations. They were also used for evaluating the contribution of parameters to temporal variations and for identification of component that promotes contamination of water quality. CA basically reveals the presence of point and non-point sources of contamination. It also shows temporal variations that precipitation and water runoffs control. PCA specifically identifies the factors or sources that cause water quality degradation. The limitation of this work is that here the size of the dataset is not specified. Also how data is collected, whether it is static or streaming data is also not clear.

The proposed work of Shah C. Azhar et al. in [10] classifies the water quality using nine monitoring stations of Muda River Basin of Malaysia. The dataset of their research work is of 9 years with six water quality variables. The variables were: Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Suspended Solids (SS), pH and Ammonia Nitrogen (NH3-N). In this study, Principal Component Analysis (PCA), Cluster Analysis (CA) and Discriminant Analysis (DA) are used for doing the multivariate analysis. They have used PCA and CA for two different classes to reflect water quality features of the river. After that, DA was used for validating the classes using NH3-N as a variable. This was done by producing a Discriminant function (DF). The DF was then used for predicting the classes to which the new sample values will belong. The shortcoming that occurs in this paper is that here dataset is static and is not being accessed through API.

Y. Magara in [11] deal with the different types of water quality standards that exist and the concepts that are used for developing the standards according to the target environment. According to the author, ambient water quality standard is a very basic tool for water quality management. As an example, the author has used Japanese Ambient water quality standards to discuss the concept. The author has also described the different parameters that effect water quality standards. These are: pH, BOD, SS, DO and Total coliform bacteria. The weakness of paper is that here the author has only described ambient water quality standards as a water quality standard. For doing the classification, water quality standard should be compared with several other quality standards.

The research work of H. Effendi et al. in [12] was done for determining the water quality status of Ciambulawung River near Halimun Mountain, Indonesia. For doing the analysis, three sampling stations were made. The WQI was identified on the basis of eight parameters namely; DO, pH, BOD, Temperature, Total phosphate, Nitrate, Turbidity, Total Solids. From the research work, the authors found out that Pollution Index lies between 0-1 and Water Quality Index of the river also lies in a good range. Hence from these two attributes, the author concluded that the water quality of the river is good and the villagers along the river bank and the hydro power plant have no negative effect on their river. Here the drawback of the paper is that although author here measures Water Quality but the size of dataset size is not defined. Also, most of the experiment was performed inside laboratory instead of on site experiment.

In a research paper by Ke Gu et al. [13], the authors have proposed a heuristic recurrent air quality predictor (RAQP) for inferring air quality on the basis of factors like fine particulate matter (PM2.5). According to the authors, current meteorological factors and air pollutants have a significant impact on the air quality of next duration concentration. However, simple machine learning tools are effective in predicting air quality for short duration. But they fail to infer air quality for large time duration due to non-linear variables. To solve this problem author has given RQVP model which applies the one-hour prediction model to predict the air quality one-hour later and then estimate the air quality after few hours. According to the authors, the RQVP model proves to be superior to the traditional models. Similarly, the research work of Salah A. Sharif et al. [14], provides a study of the environment in areas near to South Baghdad Power Plants. For the research work, authors have

selected twenty one sites from inside and six from outside the power plant. These sites were chosen for sampling and testing and doing the analysis. From the analysis, the authors concluded that nearby areas of the power plant contained a significant amount of pollutants concentration including heavy metals. In the end, authors have also provided some recommendations on the basis of the interpretations derived from their research work.

After studying different research works, we selected Australian river's data for our analysis and used machine learning for prediction and visualization. In next section, we have discussed our methodology in detail.

III. Proposed Methodology

We divided our project work into two components: prediction & visualization. The first part of the project is the visualization and extracting insights from the existing dataset with the help of various APIs and visualization tools. The second part of the project deals with the design of a prediction model which can predict the future value of the different parameters of water quality measurements. These predictions will help in analyzing that how the water quality will change if the scenario remains the same.

This section describes how the application is designed. It provides a description of the different modules and their interrelation. This section also provides a detailed description of the dataset and the proposed algorithm.

Here we have described the overall architecture of the proposed model. The model is mainly divided into two components. First, part is the visualization and extracting insights from the existing dataset with the help of various APIs and visualization tools. While, the second part deals with the design of a prediction model. The prediction model is designed to predict Quality of Water on the basis of different parameters. These predictions will help in analyzing that how the water quality changes with different parameters. The aim behind designing this model is to develop a system which classifies the water quality into different categories. And this, in turn, could be used to prevent future bad quality water by taking required measures.

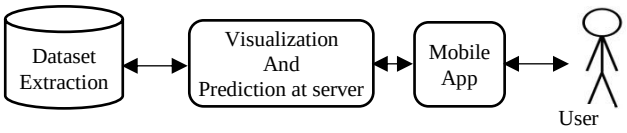


Figure 1: Overall Architecture

Figure 1 explains the overall architecture of the model. The model mainly consists of three components i.e. dataset extraction, a server which performs visualization and prediction on data and mobile application which act as a medium of interaction between user and server.

A. Phase 1: Dataset Extraction

For this project, the Australian Dataset named as Queensland Government Data is used. It has been taken from website https://data.qld.gov.au/dataset/ambient-estuarine-water-quality-monitoring-data-1993-to-2013.

Table 1
Water Quality Measuring Sensors in Different Rivers

Sensor No.	Location (latitude, longitude)	River Associated
S0	-23.5316, 150.83022	Fitzroy River
S1	-23.87388889, 151.1916667	Calliope River and Anabranch
S2	-23.958, 151.35955	Boyne River and South Trees Inlet
S3	-24.54226, 151.90452	Baffle Creek
S4	-24.7178 , 152.17464	Kolan River
S5	-25.2752 , 152.909	Great Sandy Straits and Hervey Bay
S6	-24.77166667, 152.3802778	Burnett River
S7	-25.26583333, 152.5688889	Burrum River
S8	-25.2093 , 152.49536	Isis River
S9	-25.90209, 153.02067	Tin Can Inlet and Snapper Creek
S10	-25.16821, 152.5335	Gregory River
S11	-25.45805556, 152.8822222	Mary River

The data contains records from 1993 to 2012. It includes datasets from 12 different rivers of Australia. The data contains records from 1993 to 2012. Table 1 provides information regarding sensor location in different rivers while figure 1 provides the location of the sensor on the map. The final dataset was created by combining the data of all the rivers. This combined dataset contained 74,886 records. It had 23 parameters in total with a combination of numeric, string and nominal values. The dataset had location name as the nominal parameter, while the latitude and longitudes as numeric parameters. The dataset also had a parameter called *Secchi* depth which was of numeric type. It was measured with unit meters. This depth shows the readings of sensors at different depth in the river.

Figure 2: Location of various sensors

There were various other parameters but we majorly focused on Dissolved Oxygen Concentration (DCO), pH, Salinity, Temperature and Turbidity. These were most commonly used parameters to determine the water quality. All these parameters were present in numeric form and no categorical data was necessary. Table 2 explains the parameters used in detail.

Table 2
Description of Parameters

Parameters	Units	Description	Range
Dissolved Oxygen concentration(DOC)	mg/L	It is the amount of oxygen dissolved in water in the form of microscopic bubbles and is available for aquatic life.	DOC > 5 : safe DOC < 2 : not safe
pH	H^+	This is used to measure the hydrogen ion (H+) concentration in a solution. It is a measure of the acidity or alkalinity of a	6.7 - 8.5 : safe

		solution.	
Salinity	PSU	It is the saltiness of a solution or amount of salt dissolved in the solution. It is an essential component to analyze the chemistry of water bodies.	Salinity < 0.05: fresh water Salinity > 30 : saline water
Temperature	°C	Temperature of river water is a physical parameter used to measure water quality.	15°C - 30°C : safe
Turbidity	NTU	It is the degree to which the water loses transparency because of the presence of suspended particulates in water. More total suspended solids in the water, the murkier it seems and the higher the turbidity.	Turbidity < 5: fine

B. *Phase II: Visualization and Prediction*

After dataset extraction, the data was used to gather insights and to design the prediction model.

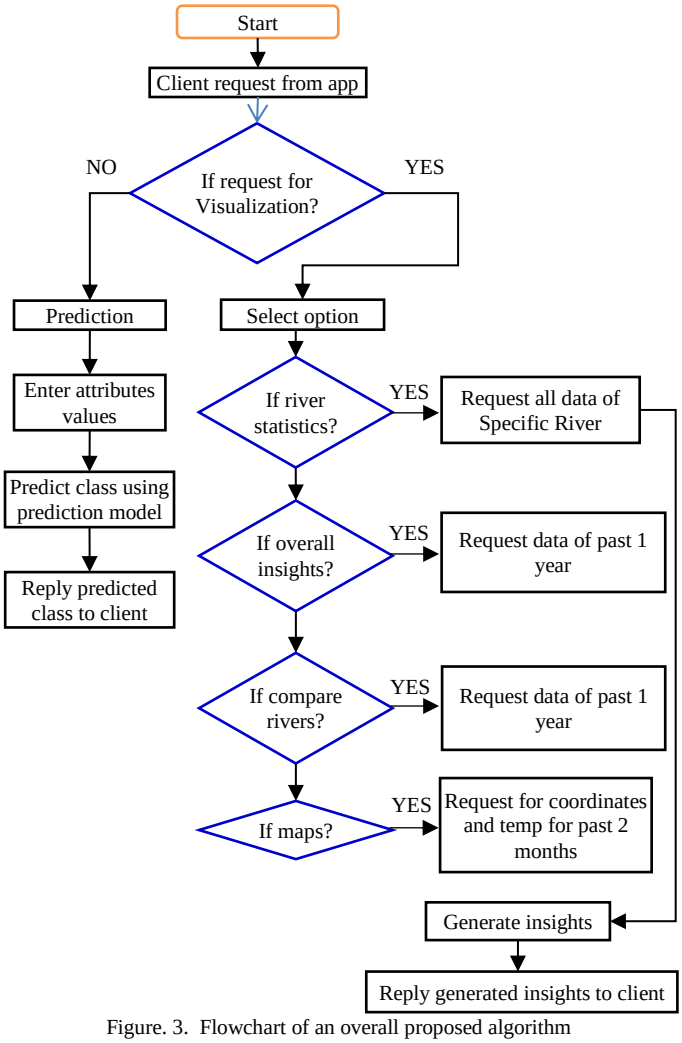


Figure. 3. Flowchart of an overall proposed algorithm

In this section, we have explained the flow of the application. Figure 3 shows that user has the option to request for either visualization or prediction. Further, the figure has explained how according to the various options the application will process.

1) *Visualization model*

Through literature review, it was identified that there are a various parameter that can be used to extract insights form the data. In this paper, we have done four types of analysis on the data and the visualized the results. Insights that we gathered shows what all we can interpret from the dataset. Firstly, we have analyzed how the different parameters have varied with time on each site. This analysis helps to understand how the level of various parameters has changed with time on different sites. This can be further linked with other parameters present in the environment during that time period. This could help in the production of new results and understanding there dependency on each other.

```
var dom1 = document.getElementById("container_ph");
var myChart1 = echarts.init(dom1);
var app1 = {};
option1 = null;
app1.title = 'ph';

option1 = {
    angleAxis: {
        type: 'category',
        data: ["0","1","2","3","4","5","6","7","8","9","10","11"],
        z: 10
    },
    series: [{
        type: 'bar',
        data: [<%=(String)request.getAttribute("ph")%>],
        coordinateSystem: 'polar',
        name: 'pH',
        stack: 'a'
    },{
        type: 'bar',
        data: [<%=(String)request.getAttribute("doc")%>],
        coordinateSystem: 'polar',
        name: 'DOC',
        stack: 'a'
    },{
        type: 'bar',
        data: [<%=(String)request.getAttribute("turb")%>],
        coordinateSystem: 'polar',
        name: 'turbudity',
        stack: 'a'
    },{
        type: 'bar',
        data: [<%=(String)request.getAttribute("sal")%>],
        coordinateSystem: 'polar',
        name: 'salinity',
        stack: 'a'
    }],
    legend: {
        show: true,
        padding: [10, 5],
        data: ['pH', 'DOC', 'salinity', 'turbudity']
    }

};

if (option1 && typeof option1 === "object") {
    myChart1.setOption(option1, true);
}
```

Figure 4: Code snippet

Secondly, we have identified how rivers can be differentiated on the basis of the amount of a parameter present on a site. Here, we have identified the number of sites which are safe or unsafe according to different parameters. Thirdly, a comparative study on different sites was performed. And lastly, the variation of the temperature on different sites was used to categorize the sites cool, normal or hot. To perform visualization, the data is extracted from the website with the help of predefined APIs. The relevant information is extracted from it and results are displayed to the user in terms of pie charts and graphs.

2) *Prediction model*

In this paper, we have designed a classification model to classify the quality of river water. Here we have provided a solution to the problem of labeling the unlabeled data. This

solution is for the scenario when we have a set of class labeled and unlabeled dataset. In the dataset, labeling can be done with the help of the different properties of attributes present. In our proposed work we have first clustered the data then this clustered data is used to create a decision tree. Later this tree was used to determine the accurate class label for each cluster depending on the properties of the attributes. Figure 5 shows the flowchart related to the prediction model.

To create this model, we firstly preprocessed the dataset. Initially, the size of the data was very large in terms of dimension, as shown in Figure 6. So, we applied feature selection on it which reduced the data to 5 dimensions i.e. Dissolved Oxygen Concentration, pH, Salinity, Temperature and Turbidity. Then a filter was applied to remove redundancy from the data.

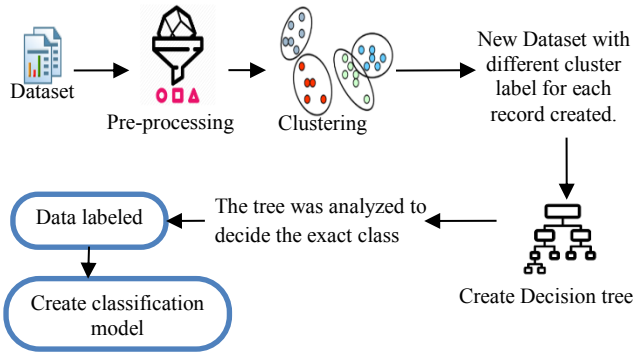


Figure 5: Flowchart of the proposed prediction algorithm

As shown in Figure 7, this reduced the size of the dataset to 68,486.

1:LOCATION_ID_BSA	2:LOCATION_NAME	3:Lat_GDA94	4:Long_GDA94	5:SURVEY_DATE	6:Time_HHMMSS	7:Depth_m	8:Chlorophyll-a (ug/L)	9:Secchi depth (metres)	10:Ammonia as N (mg/L)	11:Nitrogen Oxides as N (mg/L)	12:Total
Numeric	Nominal	Numeric	Numeric	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	
1.34E12	BAFFLE CREEK (2...	-24.54226	151.90452	06-11-12	155400.0	0.2	1.4	2.4	0.005	0.002	
1.34E12	BAFFLE CREEK (2...	-24.54226	151.90452	06-11-12	155400.0	2.0					
1.34E12	BAFFLE CREEK (2...	-24.54226	151.90452	06-11-12	155400.0	4.0					
1.34E12	BAFFLE CREEK (2...	-24.54226	151.90452	04-12-12	153400.0	0.2	2.6	2.15	0.005	0.002	
1.34E12	BAFFLE CREEK (2...	-24.54226	151.90452	04-12-12	153400.0	2.0					
1.34E12	BAFFLE CREEK (2...	-24.54226	151.90452	04-12-12	153400.0	4.0					
1.34E12	BAFFLE CREEK (2...	-24.54226	151.90452	04-12-12	153400.0	4.5					
1.34E12	BAFFLE CREEK (2...	-24.54226	151.90452	22-01-13	82700.0	0.2	5.1	1.7	0.005	0.002	
1.34E12	BAFFLE CREEK (2...	-24.54226	151.90452	22-01-13	82700.0	2.0					
1.34E12	BAFFLE CREEK (2...	-24.54226	151.90452	22-01-13	82700.0	4.0					
1.34E12	BAFFLE CREEK (2...	-24.54226	151.90452	22-01-13	82700.0	5.5					
1.34E12	BAFFLE CREEK (2...	-24.54226	151.90452	19-02-13	102800.0	0.2	1.6	0.7	0.009	0.018	
1.34E12	BAFFLE CREEK (2...	-24.54226	151.90452	19-02-13	102800.0	2.0					
1.34E12	BAFFLE CREEK (2...	-24.54226	151.90452	19-02-13	102800.0	4.0					
1.34E12	BAFFLE CREEK (2...	-24.54226	151.90452	19-02-13	102800.0	6.0					
1.34E12	BAFFLE CREEK (2...	-24.54226	151.90452	19-03-13	152100.0	0.2	2.6	0.9	0.008	0.05	

Figure 6: Snapshot of the available dataset

Finally, the missing values in the dataset were replaced with the mean values and final dataset as shown in Figure 7 was prepared.

No.	1: Dissolved Oxygen concentration (mg/L)	2: pH (Unit)	3: Salinity (PSU)	4: Temperature (deg C)	5: **Turbidity (NTU)**
	Numeric	Numeric	Numeric	Numeric	Numeric
1	6.4	7.6	16.4	22.88	75.0
2	6.3	7.64	19.0	22.75	95.0
3	7.2	7.79	9.92	20.48	35.0
4	7.0	7.75	10.8	20.35	110.0
5	7.0	7.74	10.9	20.33	137.0
6	7.7	7.9	24.6	17.3	28.0
7	7.6	7.9	25.1	17.2	41.0
8	8.2	7.8	25.1	17.1	56.0
9	7.8	7.94	27.8	19.56	13.0
10	7.7	7.94	28.2	19.46	14.0
11	7.6	7.93	28.4	19.44	15.0
12	6.9	8.11	34.5	23.17	6.0

Figure 7: Snapshot of the reduced dataset

Secondly, the clustering was performed. As our data does not contain any labeled class to describe the quality of river so, we applied the clustering model to partition the data into various clusters. Classification is a part of supervised learning and to perform supervised learning labeled training data is needed. So in order to get labeled data clustering was performed. We have used canopy clustering to cluster the dataset. It was used because it speedup clustering operations by reducing the number of comparisons. The data was partitioned into seven different clusters after application of clustering. Figure 8 shows the dataset produced after performing clustering.

No.	1: Dissolved Oxygen concentration (mg/L)	2: pH (Unit)	3: Salinity (PSU)	4: Temperature (deg C)	5: Turbidity (NTU)	6: **Cluster**
	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal
75	6.2	8.12	37.3	27.6	3.0	cluster1
76	5.7	8.14	36.9	27.2	10.0	cluster1
77	5.7	8.15	36.9	27.2	10.0	cluster1
78	5.8	8.15	37.0	27.2	9.0	cluster1
79	6.1	7.72	14.1	28.2	6.0	cluster0
80	6.3	7.93	20.3	28.0	6.0	cluster0
81	6.7	8.06	24.2	28.1	7.0	cluster0
82	6.8	8.11	25.8	28.2	9.0	cluster1
83	5.4	8.05	15.4	25.7	8.0	cluster0
84	5.5	8.17	18.0	25.7	7.0	cluster0
85	5.6	8.25	19.7	25.7	9.0	cluster0
86	6.2	8.38	30.9	23.2	11.0	cluster1

Figure 8: Clustering

After clustering, we got a new dataset which contains cluster number as one of its attributes. Now the task was to provide correct label name to each cluster. To decide the correct class label the decision tree was created. To create the decision tree, J48 algorithm was used. A decision tree was created by computing the information gain of all attributes. The attribute with the highest information gain (Salinity) resulted in the first division of the tree. Second highest information gain of Temperature was used to break the tree to next level. Similarly, other attributes were used till we reached the stage where no attribute was left. As shown in Figure 9, a snapshot of a partial tree, the internal nodes of the tree were attributes Salinity, Temperature, pH, Turbidity and Dissolved Oxygen (DO). The branches were possible values, and terminal nodes were final value of our dependent attribute.

Table 3
The range of values for ranking DOC

DOC	Rank
<2	Bad
2-5	Good
>5	Very good

Since our dataset was divided into seven clusters (Fair, Good, Very Good, Poor, Marginal, Worst, Excellent), therefore we classified it into seven class labels. To decide the label of the cluster we used the values as shown in Table 3. Table 3 enlists the range for deciding rank based on Dissolved Oxygen Concentration 'DOC' values. There are three divisions for DOC as 'bad', 'good' or 'very good'.

```
|   Temperature (deg C) <= 20.63
|   |   |   |   |   Salinity (PSU) <= 8.51: cluster2 (2641.91/7.28)
|   |   |   |   |   Salinity (PSU) > 8.51
|   |   |   |   |   |   Temperature (deg C) <= 19.82
|   |   |   |   |   |   |   Dissolved Oxygen concentration (mg/L) <= 12.3: cluster2 (172.69/1.29)
|   |   |   |   |   |   |   Dissolved Oxygen concentration (mg/L) > 12.3: cluster6 (3.0/1.0)
|   |   |   |   |   |   Temperature (deg C) > 19.82
|   |   |   |   |   |   |   Dissolved Oxygen concentration (mg/L) <= 7.2: cluster0 (16.0/2.09)
|   |   |   |   |   |   |   Dissolved Oxygen concentration (mg/L) > 7.2
|   |   |   |   |   |   |   |   Salinity (PSU) <= 9.9: cluster2 (19.1/2.1)
|   |   |   |   |   |   |   |   Salinity (PSU) > 9.9
|   |   |   |   |   |   |   |   |   Dissolved Oxygen concentration (mg/L) <= 7.8: cluster0 (4.02/0.02)
|   |   |   |   |   |   |   |   |   Dissolved Oxygen concentration (mg/L) > 7.8: cluster2 (2.01/0.01)
|   |   |   |   Temperature (deg C) > 20.63
|   |   |   |   |   Salinity (PSU) <= 1.75
|   |   |   |   |   |   pH (Unit) <= 6.72
|   |   |   |   |   |   |   Temperature (deg C) <= 23.13: cluster2 (30.0/1.0)
|   |   |   |   |   |   |   Temperature (deg C) > 23.13: cluster4 (11.0/0.0)
|   |   |   |   |   |   pH (Unit) > 6.72
|   |   |   |   |   |   |   Temperature (deg C) <= 22.89: cluster2 (601.2/6.45)
|   |   |   |   |   |   |   Temperature (deg C) > 22.89
|   |   |   |   |   |   |   |   Dissolved Oxygen concentration (mg/L) <= 6.4
|   |   |   |   |   |   |   |   |   pH (Unit) <= 7.55
|   |   |   |   |   |   |   |   |   |   Temperature (deg C) <= 23.38
|   |   |   |   |   |   |   |   |   |   |   Turbidity (NTU) <= 55.5
|   |   |   |   |   |   |   |   |   |   |   |   Dissolved Oxygen concentration (mg/L) <= 6.2: cluster0
(3.0/0.0)
|   |   |   |   |   |   |   |   |   |   |   Dissolved Oxygen concentration (mg/L) > 6.2: cluster2
(2.0/0.0)
|   |   |   |   |   |   |   |   |   |   |   Turbidity (NTU) > 55.5: cluster2 (32.0/1.0)
|   |   |   |   |   |   |   |   |   |   Temperature (deg C) > 23.38: cluster0 (4.0/1.0)
|   |   |   |   |   |   |   |   |   pH (Unit) > 7.55
|   |   |   |   |   |   |   |   |   |   Temperature (deg C) <= 22.98: cluster2 (4.05/0.05)
|   |   |   |   |   |   |   |   |   |   Temperature (deg C) > 22.98: cluster0 (21.16/1.16)
```

Figure 9: Subtree form created a Decision tree

For example, if DOC was less than 2 for the majority of the traversals, then class was assigned as 'Bad'. As shown in Figure 10, we traversed the tree bottom up for all leaf nodes of a particular cluster and decided its class based on majority score. As an example, as shown in Figure 10 while traversing three leaf nodes if we encountered DOC 'very good', Temperature 'safe' and salinity 'safe' then we labeled the cluster as 'Good'.

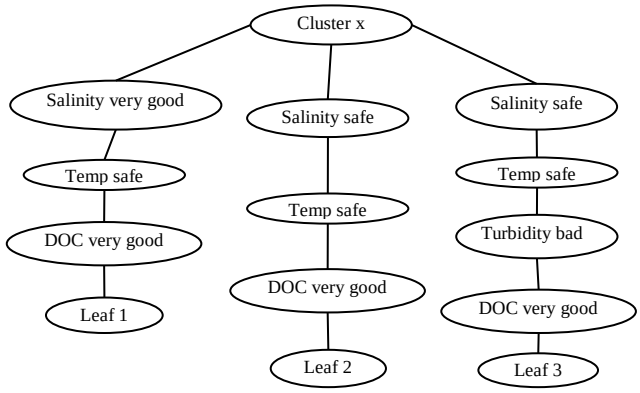


Figure 10: Cluster Class Illustration

Similarly, all clusters were labeled as given in Table 4.

Table 4
Cluster Mapping to Class Label

Clusters	Ranking	Class label
Cluster 0	4	Fair
Cluster 1	3	Good
Cluster 2	2	Very Good
Cluster 3	6	Poor
Cluster 4	5	Marginal
Cluster 5	7	Worst
Cluster 6	1	Excellent

Max, min and average score of all the five attributes were also used in finalizing the cluster. Therefore, combined with a decision tree and score the cluster labels were computed. Table 4 shows the mapping of various clusters with a class label.

As a result, we got a labeled dataset which could be used for classification. A decision tree was used for classification of data which was later used for testing. The classification model generated provided a very good accuracy.

Table 5
Minimum, Maximum & Average Scores for Different Attributes

Parameter	Maximum Value	Average	Minimum Value
DO	19.3	6.574	0.2
pH	9.26	7.884	5.24
Salinity	44.3	25.269	0.023
Temperature	36.43	24.09	13.99
Turbidity	1211	-0.4	53.329

The prediction model was tested to identify how much accurate it was. For this the data was split into 80:20 ratios, where 80% of the data was used for training and 20% was used for testing. The total number of instances in testing was 13,697. Through Table 4 it is clear how different cluster are mapped to the desired class label, while Table 5 shows the average, minimum and maximum values of different parameters. For prediction model testing weak tool was used.

Table 6
Correctness Scores of Proposed Model

	No of samples	Correctness percentage
Correctly Classified Instances	13587	99.1969 %
Incorrectly Classified Instances	110	0.8031 %

The mapped data was fed to the classification model. The model generated was tested using WEKA tool. The model had an accuracy of 99.19% which was very good. The results are tabulated in Table 6 and Table 7. Table 7 shows the confusion matrix of the proposed model. Through it we can interpret that none of the instances is misclassified as cluster 6. Further, we can also interpret that cluster 0, 1 and 2 have more probability to get misclassified. Graphs in Figure 11 shows the detailed accuracy of different classes whereas Table 8 has tabulated the same.

Table 7
Confusion Matrix

a	b	c	d	e	f	g	←Classified as
3039	20	22	2	1	0	3	**a = cluster0**
28	9201	0	0	0	0	0	**b = cluster1**
20	0	901	1	1	0	0	**c = cluster2**
2	0	295	1	1	0	0	**d = cluster3**
4	0	1	2	126	0	0	**e = cluster4**
0	0	0	0	0	12	0	**f = cluster5**
0	2	0	0	0	0	13	**g = cluster6**

Table 8
Accuracy of Different Parameters

	Fair	Good	Very Good	Poor	Marginal	Worst	Excellent
TP rate	0.984	0.997	0.976	0.99	0.947	1	0.867
FP rate	0.005	0.005	0.005	0.002	0	0	0
Precision	0.983	0.998	0.975	0.983	0.997	1	0.813

IV. RESULTS & DISCUSSIONS

This section contains information about system requirements. Further, it has explained implementation process of the application in addition to different

technologies used for implementation.

1) System Requirements

The server was designed on a single machine consisting of Intel core i5 and 2.3GHz processor. The RAM of the system was 8GB, and the system contains the main memory of 500 GB. OS of the system was Windows 7. The client was an android phone.

The proposed work was implemented using NetBeans ide in java language. The server was designed using servlet programming. The client side was an android application. This application sends a request to the server, then receives a reply and displays it to the user. To implement the prediction model the Weka libraries were used in java. The correctness of the designed model was examined with the help of Weka tool. The visual representation of data various was done using Echarts and amCharts data visualization libraries.

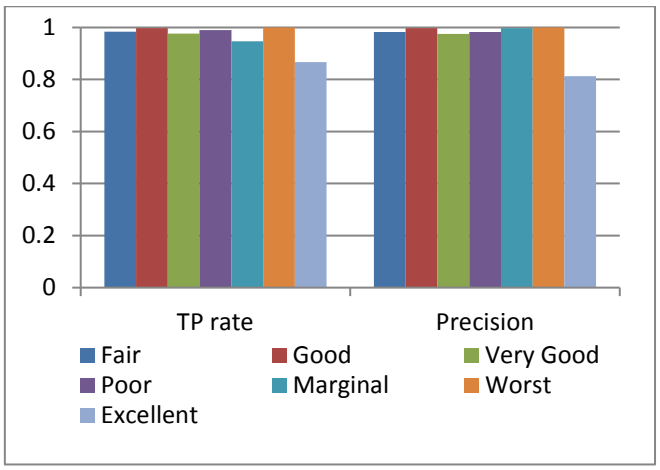


Figure 11: Accuracy of a different class

2) Visualization

Visualization plays an integral part in providing better interpretation of the results. A normal person with no expertise in cumbersome machine learning algorithms can learn more if the results are displayed as visuals to him. We, therefore, provided an app for better analysis to natives of the place. Figure 12 shows the first page or the main page of the application. This page has visualization and prediction option for the user to choose from.

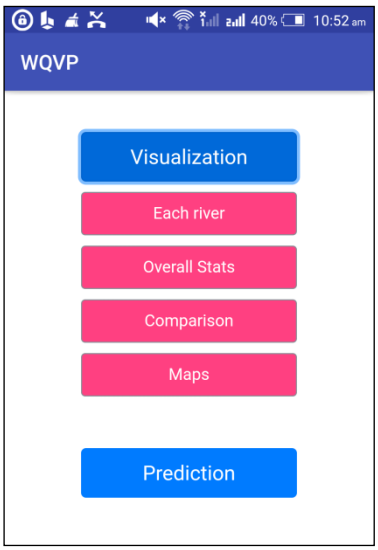
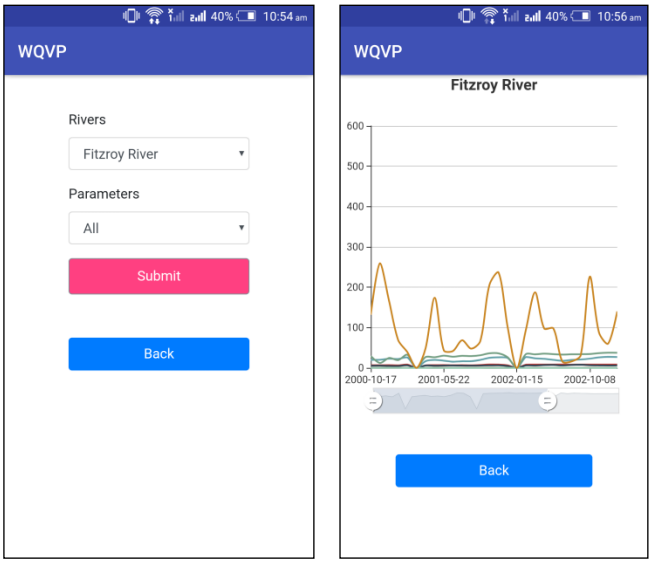


Figure 12: Front Page of the WQVP App

The user can choose visualization and ask for prediction by clicking visualization and prediction buttons respectively. The data analysis can be received for each river individually, or comparison can be asked through 'Each River' and 'Comparison' buttons. Option for 'maps' has also been provided.

The visualization option produces various insights from the dataset according to user requirements. Figure13 explains various insights that can be extracted from the data available. To get details of various water pollutants name of the river was selected from the drop down menu and choice of parameters could be given in 'parameters' box. As shown in Figure 13(a) the name of river selected was 'Fitzroy' and 'all' parameters were selected. Based on the input values received values were shown to the user as shown in Figure 13(b). Through this timeline graph, one could analyze the everyday change in turbidity, salinity, pH, total Nitrogen, total phosphorus and dissolved oxygen content in water. It provides insights on various parameters from 1993 to 2012.



(a) (b)

Figure 13: Variations in different parameters with time for different rivers

Figure 13(b) contains insights about different sites. Through literature review significance of various parameter of water was understood. Thus we have generated insights about the health and productivity risk of various sites.

Here, we have visualized the number of sites which are suitable for different scenarios on the basis of different parameters. For example, sites were categorized as safe and unsafe on the basis of pH. These statistics show that *x* sites are safe and *y* sites are unsafe in recent time. Similarly, on the basis of the turbidity, the water of various sites can be categorized as suitable for drinking, suitable for the fishery, safe for indigenous fishes and dirty. Sites were distinguished as unsafe, safe, good and perfect on the basis of the content of dissolved oxygen. On the basis of salinity, they were differentiated as below 30, between 30 to 40 and above 40. It is because the water below 30 is brackish and fresh water while above it is saline.

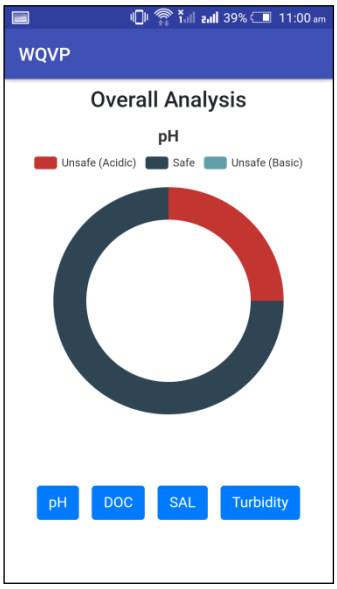


Figure 14: pH scores for Australian Rivers

Figure 14 shows overall 'pH' analysis. For given three parameters 25% were found to be acidic 'brown color', and remaining 75% were found to be safe 'dark blue color' based on pH levels. Here turbidity, salinity, pH, and dissolved oxygen of all the 13 sites are compared. The comparison was made on the basis information of past one year about different sites.

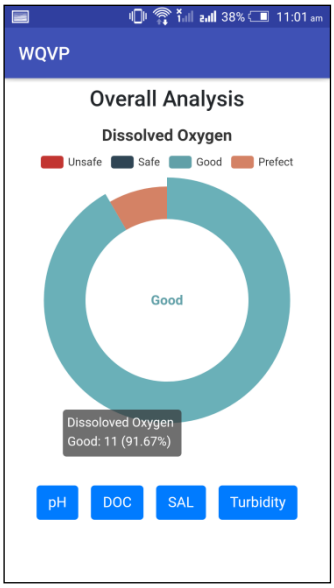


Figure 15: Dissolved Oxygen scores for Australian Rivers

Figure 15 shows overall 'Dissolved Oxygen' analysis. For given four parameters 11 Australian rivers with 91.67% 'good' level of dissolved oxygen were found. Remaining 8% rivers were found to have perfect dissolved oxygen.

Figure 16 shows overall analysis of water 'salinity' levels in various rivers. Salinity is important to a certain level for marine growth, but the presence of a large amount of salinity in water makes it unsuitable for drinking purposes. We divided salinity into three categories. Below 30 was considered good, between 30 to 40 was considered tolerant and beyond 40 was considered as the bad level of salinity. From the data analysis for salinity, we observed that more than 78% of the rivers have acceptable saline levels whereas remaining ones fall under tolerable levels of 30 to 40.

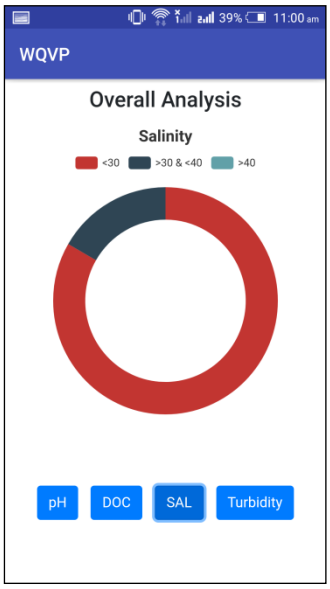


Figure 16: Salinity scores for Australian Rivers

Figure 17 shows overall analysis of all rivers on the basis of Turbidity. Good turbidity levels are required for fishery industry. High turbidity damages the fish and is not considered good for consumable fish for human beings.

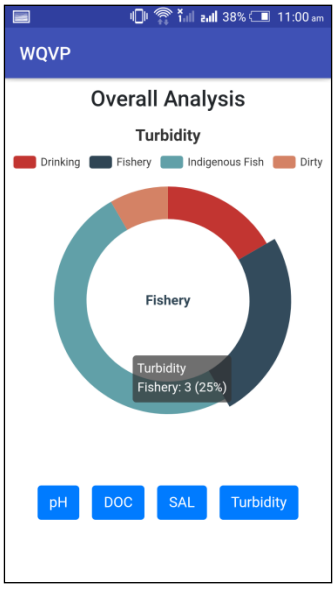


Figure 17: Turbidity scores for Australian Rivers

We divided turbidity levels into four classes. Good turbidity levels for 'drinking', second class for all types of 'fishery', the third level for 'indigenous' Australian fish and las one 'dirty' which is not fit for the fishery. It was observed that of all the Australian rivers we found 3 rivers with 25% good conditions for the fishery. We also found that rivers in the metropolitan areas were unfit for the fishery with the percentage of 4%.

Figure 18 & Figure 19 show comparisons of different rivers. The user was also given the option to choose and compare all the parameters or specific parameters for all the rivers with the help of 'wind rose' charts. Wind rose charts are two dimensional

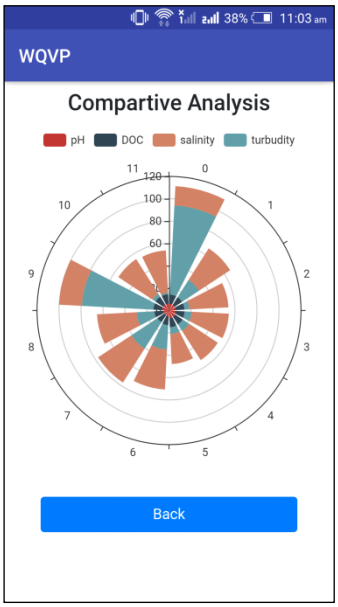


Figure 18: Wind Rose Chart for Salinity & Turbidity Comparison of all Rivers

charts used to display wind speeds and strengths in different directions at the same time. We used it to show comparisons within different water pollution parameters. Figure 18 shows a comparison of Salinity & Turbidity levels for all rivers individually. Figure 19 shows a comparison of individual Salinity levels of all the rivers. Therefore given on the user input the required chart could be generated and analyzed in real time.

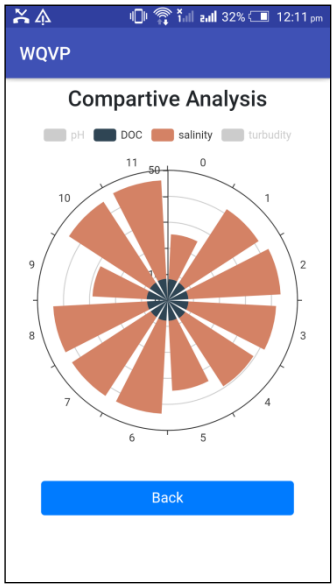


Figure 19: Wind Rose Chart for Individual Salinity Comparison of all Rivers

Visualization was performed with the help of a map of Australia as well. Figures 20 & 21 show where different sensor sites are located in Australia. We used APIs to fetch temperature data to distinguish various sites on the basis of current temperature present there. On the basis of temperature, we categorized the sites as cool, normal or hot. Temperature plays a crucial role in maintaining good aquatic life. Australia provides both very high temperatures and cold temperatures in its regions. Though there are websites that do provide temperature updates but we were looking for an app

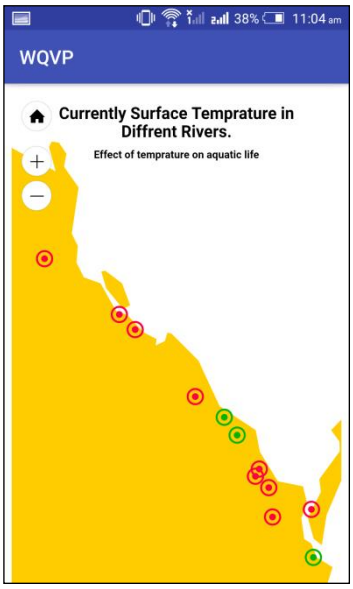


Figure 20: Temperature Map showing Surface Temperatures of rivers

that could provide updated temperature information to sailors etc. in the same app. This way they did not have to surf various websites for different information. All information could be searched in a single app. Figure 20 shows the temperature conditions present on the water surface in Baffle Creek.

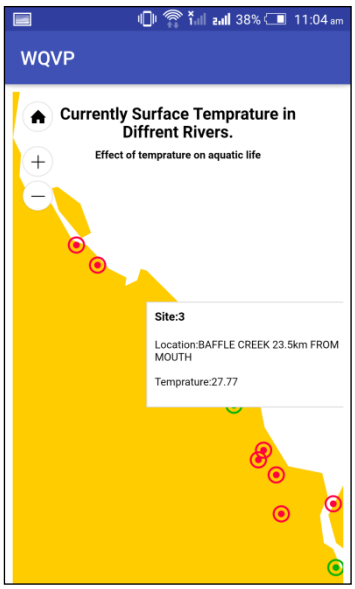


Figure 21: Temperature Map showing Surface Temperatures of rivers

3) Prediction

The prediction option predicts the category of water quality on the basis of values provided. In this user-defined values of different parameters are used to identify the water quality.

The set of figures in Figure 22, Figure 23 and Figure 24 show how the prediction process was performed. Figure 22 shows the way user can enter details to preform cluster prediction. Figure 23 shows the results in which it displays the category to which the quality of water belongs to. Further, it also contains an option to display the detailed results. Figure 24 shows the detailed results. This detailed result contained the upper and the lower bound for the safe region. This safe region is shown by dotted lines. The straight line is the current user-defined value. It shows

deviation of the value from safe point.

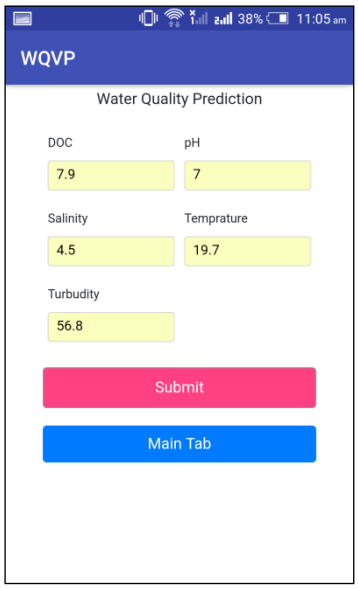


Figure 22: Water Quality Prediction for given input parameters

As shown in Figure 22 for given value inputs by the user, namely, DOC=7.9, pH=7, Salinity=4.5, Temperature=19.7 and Turbidity=56.8, the application provided results after cluster based prediction. The overall stats of the inputs could be displayed to the user showing him the quality of the water, as shown in Figure 23.

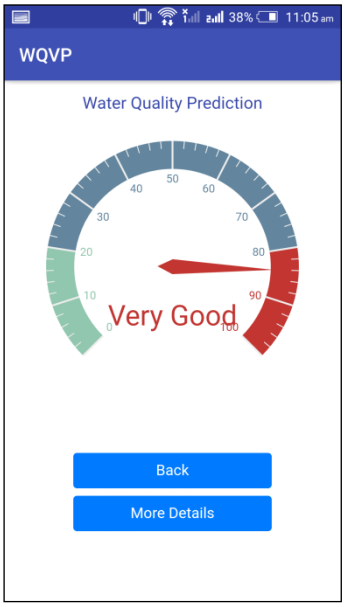


Figure 23: Water Quality Barometer to show predicted water quality

As per the water guidelines, we provided a barometer chart to display the water quality to the user. Figure 23 shows three scales of water divided from zero to hundred. Good quality water lied between 80 to 100 range and tolerable levels were from 20 to 80 range. Below 20, water quality was considered bad.

Figure 24 shows the Radar chart for displaying the safe level relations between the different parameters. We believe that a regular citizen finds it difficult to know how the various

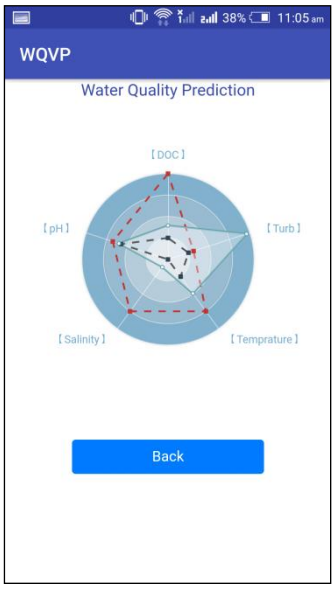


Figure 24: Radar Chart for Water Safe Limits for different Parameters

parameter ranges affect each other and water quality. A certain set of values may be good for one user but not for the other like for fishing fresh water fish, and salt water fish require different values of Salinity and Turbidity. Using Radar charts as shown in Figure 24 one could benefit from the same app.

V. CONCLUSION

Water exists in a different form on the earth. It is used by humans since its inception for various kinds of activities are it for washing, drinking, and agriculture purposes and for industrial work. The increasing consumption of water has led to water scarcity, and various efforts are being made to conserve water for future generation.

In this project, we have taken data of 12 Australian rivers for making their water quality prediction. The project consists of two phases. The first phase provides insights of the dataset with the help of API and graph libraries. This phase does the visualization of the insights we get from the data. The second phase is a prediction model. It predicts the category to which the water quality belongs. The different categories that are used are as follows: Excellent, Fair, Good, Very Good, Poor, Marginal and Worst. Initially, our data does not contain any categories. So, for categorization K-means is used and for prediction process Decision tree J48 Algorithm is used. The overall accuracy of the prediction model is 99%. We have also developed an Android application for displaying results of the project.

.

REFERENCES

[1] Drinking-water fact sheet, Online available at http://www.who.int/mediacentre/factsheets/fs391/en/. Accessed on July 2017

[2] Dataset available at : https://data.qld.gov.au/dataset/ambient-estuarine-water-quality-monitoring-data-1993-to-2013

[3] B. Srivastava, S. Sandha, V. R. Choudhury, S. Randhawa, V. Kapoor and A. Agrawal, "An Open, Multi-Sensor, Dataset of Water Pollution of Ganga Basin and its Application to Understand Impact of Large Religious Gathering", in arXiv Journal, 2016

[4] S. Emamgholizadeh, H. Kashi, I. Marofpoor, E. Zalaghi, "Prediction of Water Quality Parameters of Karoon River (Iran)by Artificial Intelligence-based models", in Int. J. Environ. Sci. Technol. (2014) vol. 11, pp. 645-656

[5] P. Varalakshmi, S. Vandhana, S. Vishali, "Prediction of Water Quality using Naïve Bayesian Algorithm", in International Conference on Advance Computing,2016

[6] S. Y. Muhammad, M. Makhtar, A. Rozaimee, A. Abdul, Aziz and A. A. Jamal, "Classification Model for Water Quality using Machine Learning Techniques", in International Journal of Software Engineering and Its Applications, Vol. 9, No. 6, pp. 45-52, 2015.

[7] M. A. Dota, C. E. Cugnasca, D. S. Barbosa, "Comparative Analysis of Decision Tree Algorithms on Quality of Water Contaminated with Soil", in Cienc. Rural, Vol.45, no.2, pp.267-273, Santa Maria, Feb. 2015.

[8] X. Fan, B. Cui, H. Zhao, Z. Zhang, H. Zhang, "Assessment of River Water Quality in Pearl River Delta using Multivariate Statistical Techniques", in International Society for Environmental Information Sciences 2010 Annual Conference (ISEIS), 2010.

[9] A. Barakat, M. El Baghdadi, J. Rais, B. Aghezzaf, M. Slassi, "Assessment of Spatial and Seasonal Water Quality Variation of OumErRbia River (Morocco) using Multivariate Statistical Techniques", International Soil and Water Conservation Research, pp. 284–292, 2016.

[10] S. C. Azhar, A. Z. Aris, Mohd K. Yusoff, M. F. Ramli, H. Juahir, "Classification of River Water Quality using Multivariate Analysis", in International Conference on Environmental Forensics,pp. 79-84, 2015.

[11] Y. Magara, "Classification of Water Quality Standards", Water Quality and Standards- Vol. I.

[12] H. Effendi, Romanto, Y. Wardiatno, "Water Quality Status of Ciambulawung River, Banten Province, based on Pollution Index and NSF-WQI", in 1st International Symposium on LAPAN-IPB Satellite for Food Security and Environmental Monitoring, pp. 228-237, 2015

[13] K. Gu, J. Qiao and W. Lin, "Recurrent Air Quality Predictor Based on Meteorology- and Pollution Related Factors", IEEE Transactions on Industrial Informatics, 2017

[14] S. Farhan, A. Sharif, G. M. Al-Saadi, "Caused by South Baghdad Power Plant South Baghdad Power Plant", International Conference on Environment Impacts of the Oil and Gas Industries (EIOGI), Koya , Kurdistan Region - Iraq, 2017