Predicting Water Quality Index for farm irrigation

Rajesh Kumar Yadav
Department of Computer Science and Engineering
Delhi Technological University
New Delhi-110042, India
rkyadav@dtu.ac.in

Adarsh Jha
Department of Computer Science and Engineering
Delhi Technological University
New Delhi-110042, India
adarshjha_2k17co22@dtu.ac.in

Aditya Choudhary
Department of Computer Science and Engineering
Delhi Technological University
New Delhi-110042, India
adityachoudhary_2k17co24@dtu.ac.in

***Abstract*—Agriculture sector of Indian economy in which more than half of the population is involved, contributes to less than quarter of the GDP. With advancement in ICT, tools and techniques can be developed that can help in analyzing and automating various phases of farming for improving productivity. This work focuses on analysing the quality of irrigation water and developing a model for prediction of Irrigation Water Quality Index(IWQI). Development of IWQI can save time and cost of lab tests for irrigation water. Five parameters of water mainly representing salinity and sodicity are measured using which the IWQI is calculated. These five parameters are further reduced to three parameters using correlation analysis and a classification model for prediction of water quality class is developed using various classification techniques. Best result is obtained by Random Forest Classifier followed by Gradient Boosting and Neural Network Classifier. The model developed in the paper can be used in agriculture to save time and cost of expensive lab tests for irrigation water.**

***Index Terms*—Irrigation Water Quality, Precision Agriculture, Water salinity, Sensor Data**

I. Introduction

In a survey conducted in 2018, it was found that 50% of the workforce in India is involved in the agriculture sector, however its contribution is only 16% to the GDP [1]. This trend can be seen all over the world, the major reason for which being wastage of resources used for farming. There exists a need to increase the efficiency of each stage in farming and at a cheap cost so that it is affordable. With improvement in technology, these needs can be addressed using innovative solutions like the Internet of Things(IoT). This technology allows us to use sensors which in turn can be used to obtain large datasets for physical and chemical characteristics of soil, water and weather. This data can be analysed and machine learning tools can be developed that can help in proper utilization of resources and increase the overall efficiency of crop production.

In this paper, we develop a model for prediction of Irrigation Water Quality Index(IWQI) that mainly indicates the degree of sodicity and salinity of water sample. We used seven classification techniques which are Support Vector Classifier, Neural Networks, Gradient Boosting , Random Forest, Decision Tree, Bagging and Naive Bayes classifier.

The rest of the paper is organized as follows: Section II describes the related works, Section III discusses the preliminary topics, Section IV describes the database and algorithm of the model, Section V gives the results and finally Section VI concludes the paper.

II. Related Work

With advancement in IoT, many new tools and techniques have been developed for agriculture sector. The term Precision agriculture [2] was developed in which various types of data such as time series, geolocation, sensor etc. are gathered and analysed to improve the efficiency of agricultural practices. For example, crop color and size can be measured using optical sensors and models can be developed for predicting optimal time of harvesting based on images.

Irrigation is one of the important phases of agriculture which involves adding water to soil after plantation which needs proper analysis of chemicals in water and quality of water for proper plant growth. Water quality Indices(WQIs) are one way to aggregate various parameters of water samples into a single value. Orozco et al [3] aggregated chemical properties of water and developed a water quality index(WQI) for drinking water based on WHO guidelines. Lerios et al [4] developed a prediction model for prediction of WQI for drinking water using parameters like acidity and temperature of water. The WQI calculated for drinking water was not suitable for use in irrigation water. Mark et al. [5] studied the effect of saline water on plants and proposed methods for using them for irrigation while Shrivastav et al. [6] studied the effect of salinity on soil nutrients which decreases with increase in salinity due to its ability to reduce soil permeability. Meireles et al [3] developed a WQI for irrigation water (which mainly focused on salinity and sodicity of water) using factor and principal component analysis to reduce thirteen parameters of water samples to just five parameters. Vij et al [7] proposes the use of wireless sensor network to measure soil parameters such as moisture, temperature, weather etc. to analyse and predict the amount of water required for irrigation.

III. Preliminaries

A. Development of Irrigation Water Quality

Calculation of water quality for irrigation is necessary to study its effect on plant and soil health. Water Quality Index (WQI) is used for representing the overall quality of water. It is the aggregate of various chemical and physical parameters of water and is calculated by multiplying relative weights of parameters with parameter value. The parameters chosen should be able to be measured in all the water sources used for irrigation. In this work, we use the Irrigation Water Quality Index (IWQI) based on salinity and sodicity of water. The following subsections describe the water quality parameters, their weights and formula for calculating IWQI.

1) Parameters for IWQI: The IWQI used in this work was formulated by Meireles et al who reduced 13 parameters of water using Principal Component Analysis(PCA) and Factor Analysis(FA) to 5 parameters which are: Sodium, Chloride and Bicarbonate ion concentration, Electrical Conductivity(EC) and Sodium Absorption Ratio (SAR) [8]. Electrical conductivity is generally measured by passing current through solutions and the value of SAR is given by the combination of Calcium, Magnesium and Sodium ion concentration whose formula is given by Equation 1. It is the primary indicator of sodicity of water which affects the permeability of soil and water infiltration rate [9]. Both cases of water infiltration rate being too high or too low affects the growth of plants. The weights of the 5 parameters, which represents their contribution to water quality are converted to relative weights and are mentioned in Table I.

$$SAR = \frac{Na^{+}}{\sqrt{\frac{Ca^{2+}+Mg^{2+}}{2}}} \tag{1}$$

Here,
Na: Concentration of Sodium ion
Ca: Concentration of Calcium ion
Mg: Concentration of Magnesium ion

TABLE I
Weights for IWQI parameters

Parameters	Weights
Electrical Conductivity(EC)	0.211
Sodium(Na^{+})	0.204
Bicarbonate(HCO^{3-})	0.202
Chloride(Cl^{-})	0.194
Sodium Absorption Ratio (SAR)	0.189

2) Quality Measurement Values and IWQI: The value of parameters measured for water sample should neither be too high nor too low for optimal growth of plants due to which the parameters are normalized(between 0 to 100) to quality measurement values (q_i) according to limits shown in Table III [8] and Equation 2. The q_i values obtained are multiplied with relative weights of their respective parameters and finally added to obtain IWQI according to Equation 3 whose value is between 0 and 100. IWQI range is divided into various classes and each class represents the salinity characteristic where water with IWQI in class range 85-100 is suitable for all types of soil and plants, range 70-85 is suitable for soil with moderate permeability and should be avoided in plants with low salt tolerance, range 55-70 is suitable for soil with moderate to high permeability and plants with medium salt tolerance, range 40-55 is suitable for soil with high permeability and plants with high salt tolerance, while water with IWQI below 40 is not suitable for irrigation and they should be properly treated before use. The details for the water quality classes are given in Table II [8].

$$q_i = q_{imax} - \frac{(x_{ij} - x_{inf}) * q_{iamp}}{x_{amp}} \tag{2}$$

Here,
q_i: quality measurement values,
q_{imax}: maximum value in particular quality measurement class,
q_{iamp}: amplitude of quality measurement class,
x_{ij}: value of parameter x,
x_{inf}: minimum value of parameter x in quality measurement class,
x_{amp}: It is the class amplitude to which the parameter belong

$$IWQI = \sum_{i=1}^{n} q_i w_i \tag{3}$$

Here,
$IWQI$: Irrigation Water Quality Index,
q_i: quality measurement value ,
w_i: weight of parameter from Table I

TABLE II
Classification of water sample based on IWQI

IWQI	Soil	Plant
85–100	Can be used for any kind of soil	Most plants won't be affected
70–85	Can be used on soil with moderate permeability	Avoid use in plants with very low salt tolerance
55–70	Can be used on soils with moderate to high permeability	Avoid in plants with low salt tolerance
40–55	Can be used on soils with high permeability without dense layers	Used mainly in plants with high salt tolerance. Plants with moderate salt tolerance can be used with some control practices
0–40	Use for irrigation should be avoided	Avoid for all plants

B. Classification Algorithms

1) ANN: In ANN, we have an input layer, hidden layers and output layer each consisting of neurons in it [10]. Neurons receive some input and give an output after applying an activation function. When we give an input to this network of neurons, it produces some output which is tested with the

TABLE III
LIMITING VALUES FOR QUALITY MEASUREMENT

q_i	EC	SAR	Na^+	Cl^-	HCO^{3-}
85-100	0.2$\leq$ $EC < 0.75$	2$\leq$ $SAR < 3$	2$\leq$ Na$^+$ $<$ 3	1$\leq$ Cl$^-$ < 4	1$\leq$ HCO^{3-} $<$ 1.5
60-85	0.75$\leq$ $EC < 1.5$	3$\leq$ $SAR < 6$	3$\leq$ Na$^+$ $<$ 6	4$\leq$ Cl$^-$ < 7	1.5$\leq$ HCO^{3-} $<$ 4.5
35-60	1.50$\leq$ $EC < 3$	6$\leq$ $SAR < 12$	6$\leq$ Na$^+$ $<$ 9	7$\leq$ Cl$^-$ $<$ 10	4.5$\leq$ HCO^{3-} $<$ 8.5
0-35	$EC <$ 0.20 or EC $\geq$ 3.00	$SAR <$ 2 or SAR$\geq$ 21	$Na^+ <$ 2 or $Na^+ \geq$ 9	$Cl^- <$ 1 or $Cl^- \geq$ 10	$HCO^{3-} <$ 1 or $HCO^{3-} \geq$ 8.5

known output for that sample. The error is back propagated to the network so that it can adjust the weights for each neuron to improve the network for upcoming samples. This classifier is also known as MLP(Multilayer Perceptron Classifier).

2) SVM: This classification algorithm is very useful as it tries to maximize the margin between any two classes which makes the decision hyperplane more accurate [11]. This algorithm is also able to classify datasets with higher dimensions using the kernel trick which doesn't make use of vector calculations so that it can save computation time.

3) Gradient Boost Classifier: It is a classifier which uses loss function, weak learners and an additive component to reduce the loss [12]. It has a necessary condition that loss function should be differentiable. This algorithm uses decision trees as its weak learners. It is quite powerful but is prone to overfitting.

4) Random Forest Classifier: It is an ensemble learning method which uses multiple trees [13]. This is beneficial because one classifier may be prone to overfitting but when using a group of uncorrelated classifiers, the result is a refined classifier which doesn't overfit the dataset. Here each classifier gives its own output and the output which is chosen by more number of classifiers becomes the final output.

5) Decision Tree: It is a classifier which uses a tree-like structure where at each node a decision is made according to the feature in the dataset and final output is given by the leaf node [14].We can obtain the rules for any output by traversing from root to leaf. This algorithm is simple and easy to implement and also functions as weak learners in Gradient boosting classifier, Bagging classifier and Random forest.

6) Naive Bayes: It is a simple classifier which uses Bayes theorem with an exception that features do not depend on each other [15]. This allows us to easily implement and use the algorithm for classification but its results are not reliable.

7) Bagging Classifier: This algorithm uses weak learners like decision trees on a random portion of its dataset, using the results from those trees to produce a final result [16]. This algorithm takes advantage of the fact that the construction process is randomized which helps to improve results of unstable weak learners such as neural networks and decision trees.

IV. MATERIALS AND METHODS

A. Dataset - Major ions dataset

The dataset used in our work was obtained from US Geological Survey of Brackish Groundwater [17]. It consists of 66 thousand rows and contains values for concentration of dissolve solids, metals, nutrients, ions and physical properties such as conductivity out of which only EC, Cl$^-$, Na$^+$, HCO^{3-}, Ca^{2+} and Mg^{2+} are used for IWQI. EC stands for electrical conductivity which is measured in milliSiemens/cm. Cl$^-$, Na$^+$, HCO^{3-}, Ca^{2+} and Mg^{2+} represents Chloride, Sodium, Bicarbonate, Calcium and Magnesium ion concentration respectively. These concentrations were measured in mg/L which we converted to millimoles/L. EC needs to be in dS/cm so it was divided by 100. We use Ca^{2+}, Mg^{2+} and Na$^+$ to calculate SAR using Equation 1.

B. Irrigation Water Quality Index based classification

In this work, we develop a classification model and apply it on major ions dataset. This dataset consists of all the necessary features required to obtain the IWQI described in Section III-A1 and the model is built using these three steps: converting given values to quality measurement values, calculating irrigation water quality index(IWQI) using calculated values and relative weights of each parameter, assigning to a class according to ranges given in Table II and finally using different classification techniques and choosing the best of them.

1) Quality Measurement Values: The dataset consists of parameters with different units. Also, the value of any parameter being abnormally high or low decrease the quality of water. Hence, we need to normalize the values according to predefined limits of parameters. This is done by converting values of parameters to quality measurement values (q_i) that can be obtained using Equation 2 and Table 3. The quality measurement value (q_i) is a dimensionless quantity which can then be combined with other parameters without considering units. The Table IV indicates the number of values present in specific ranges of q_i for the dataset used. Here, we can see that majority samples have q_i values of EC in the range 85-100, bicarbonate in the range 35-85 and Cl$^-$, Na$^+$ and SAR in the range 0-35. This indicates that most of the samples have optimal value of EC and either excess or deficit value of Cl$^-$, Na$^+$ and SAR.

TABLE IV
NUMBER OF QUALITY MEASUREMENT VALUES IN EACH RANGE

q_i	Q_HCO^{3-}	Q_EC	Q_Cl$^-$	Q_Na$^+$	Q_SAR
0-35	5540	6580	32700	31400	32600
35-60	9950	0	0	0	794
60-85	20000	3500	0	2780	1950
85-100	2470	27900	5280	3810	2680

2) Calculating IWQI and developing a classification model: Using the quality measurement values calculated in section III-A2, we multiply them by their relative weights mentioned in Table I and add them to obtain IWQI according to Equation 3. The samples are further classified into quality classes according to quality ranges provided in Table II. Correlation matrix of quality parameters which was obtained using Pearson's Correlation is given in Table V. The following observations are made:

1) Cl^- is highly correlated with Na^+ and EC.
2) Na^+ is highly correlated with all the parameters.
3) EC is highly correlated with all the parameters except SAR.
4) HCO^{3-} is highly correlated with EC.
5) IWQI is highly correlated with Cl^-, Na^+ and EC.

$$r = \frac{\sum(U_i - \bar{U})(V_i - \bar{V})}{\sqrt{\sum(U_i - \bar{U})^2 (V_i - \bar{V})^2}} \tag{4}$$

Here,
r: correlation coefficient
U_i: value of U-variable in the sample
$\bar{U}$: mean value of U-variable
V_i: value of V-variable in the sample
$\bar{V}$: mean value of V-variable

TABLE V
CORRELATION MATRIX FOR VARIOUS PARAMETERS FOR WATER SAMPLES

	Cl^-	Na^+	EC	HCO^{3-}	SAR	IWQI
Cl^-	1	0.467	0.506	0.159	0.083	0.44
Na^+	0.467	1	0.482	0.297	0.56	0.607
EC	0.506	0.482	1	0.835	0.018	0.423
HCO^{3-}	0.159	0.297	0.835	1	0.018	0.293
SAR	0.083	0.56	0.018	0.018	1	0.302
IWQI	0.44	0.607	0.423	0.293	0.302	1

The three parameters: Na^+, Cl^- and EC are chosen to develop classification algorithms to predict water quality. Choosing three parameters saves us the cost of measuring HCO^{3-} and SAR. The classification algorithms mentioned in Section III-B are used to predict quality classes and the method having best accuracy and F2 score is selected for further use. The overall algorithm is mentioned in algorithm IV-B2.

Classification of water samples based on IWQI

```
Input: Water quality dataset with 5 parameters $Cl^-$, $Na^+$, EC, $HCO^{3-}$ and SAR
Output: algorithm with best score
Calculate quality measurement values($q_{param}$) for each parameter.
Calculate IWQI using the Equation 3 and assign classes from Table II
Choose features of dataset as: features $\leftarrow Cl^-$,$Na^+$,EC
Choose labels of dataset as:IWQI
Split dataset into training_set and testing_set
Apply the 7 classifiers mentioned in Section III-B on the training set and evaluate on the testing set using accuracy and F1 score.
return algorithm with best score
```

V. Experiments and Results

The algorithm was implemented in python 3 with pandas [18] and numpy [19] for loading and preprocessing datasets. All of the classification methods were applied using Scikit-learn [20]. Parameter settings are explained in V-A, evaluation criteria is explained in V-B and results in V-C.

A. Parameter Settings

The parameter settings for all the classification models are given in Table VI and the parameters are described below:

1) Decision Tree: criterion: function to evaluate split quality, splitter: method used to split at each node, min_sample_split: least amount of samples required to split a non-leaf node

2) Naive Bayes: alpha: the additive smoothing parameter, class_prior: estimated probability for new class prediction, class_count: number of samples under each class

3) Gradient Boosting: loss: function to optimize, learning_rate: for changing contribution of trees to the result, N_estimators: levels of boosting

4) Random Forest: n_estimators: Amount of tree before voting, criterion: function to evaluate split quality, base_estimator: base classifier to create sub estimator

5) Support vector classifier: C: Regularization, decision_shape: decision function's shape, tol: stopping tolerance

6) Bagging Classifier: base_estimator: base classifier used to fit on random subdomains, n_estimator: count of the base_estimator

7) MLP Classifier: hidden_layer_sizes: array representing sizes of hidden layers, activation: output function of hidden layer, learning_rate: step size of gradient descent

TABLE VI
PARAMETER SETTINGS FOR CLASSIFIERS

Classifier	**Parameter Settings**
Bagging Classifier	base_estimator:SVC, n_estimators:10
DecisionTree	criterion: gini, splitter: best, min_samples_split: 2
Naive Bayes	alpha: 1.0, class_prior: None, class_count: [7059. 15826. 5057. 2055. 402.]
Gradient Boosting	loss: deviance, learning rate: 0.1, n_estimators: 100
Random Forest	n_estimators: 100, criterion: gini, base_estimator_: DecisionTreeClassifier
SVM	C: 1.0, decision_function_shape: 'ovr', tol:0.001
MLP	hidden_layer_sizes:128, activation='relu', learning_rate=0.001

B. Evaluation Metrics For Classification Model

For testing our classifier, we use accuracy, precision, recall and F1 score. These metrics are calculated to test the reliability and robustness of classifiers used . These metrics are defined in Section V-B1, V-B2, V-B3 and V-B4 [21]. The equations for the same are given at 5, 6, 7 and 8.

1) Accuracy: It tell us how much data can be predicted accurately from a given classifier. This gives us an idea about the bulk of data which is correctly classified.

2) Precision: It tells us how many predictions are correct out of all observations saying output belongs to a particular class.

3) Recall: It tells us how many correct predictions were made for a class out of all the actual observations for a given class.

4) F1 Score: It can be defined as harmonic mean of precision and recall, can be used independently for evaluation of classifiers.

$$Accuracy = \frac{TPos + TNeg}{TPos + TNeg + FPos + FNeg} \quad (5)$$

$$Precision(P) = \frac{TPos}{TPos + FPos} \quad (6)$$

$$Recall(R) = \frac{TPos}{TPos + FNeg} \quad (7)$$

$$F1Score = 2 * \frac{P * R}{P + R} \quad (8)$$

Here,
TPos: Number of observations which are correctly classified for positive class in binary classification
TNeg: Number of observations which are correctly classified for Negative class in binary classification
FPos: Number of observations which are incorrectly classified for positive class in binary classification
FNeg: Number of observations which are incorrectly classified for negative class in binary classification

C. Results of Classification based on IWQI

Algorithm IV-B2 is applied to major ions dataset with Electrical Conductivity, Cl$^-$ and Na$^+$ as features and IWQI class as labels and results are evaluated using the metrics mentioned in Section V-B, details of which are mentioned in Section IV-B2. For the major ions dataset, the number of water samples in IWQI range 85-100 is around 1.3% of the total samples which affected the results of correctly classifying samples belonging to that range. This however does not adversely affects the overall result. Random Forest performed the best with an accuracy of 86.9% followed by Gradient Boosting and Neural Networks with accuracy of 85.8% and 84.6% respectively. Naive Bayes classifier performed the worst with accuracy of 52.6%. The results of all the algorithms are given in Table VII. We also performed five-fold cross validation and computed the accuracies whose results are in Table VIII. From both the results, performances of all the algorithms are good except that of Naive Bayes.

VI. Conclusion and Future Work

Analysis of irrigation water quality can help in increasing agricultural productivity and prevent damage to plants. In this work, we developed a classification model for prediction of Irrigation Water Quality Index (IWQI) class. The IWQI used here is an aggregate of five parameters that are SAR, EC, Chloride, Bicarbonate and Sodium ion concentration respectively. For classification, we used seven classification algorithms and selected three out of five parameters: EC, Cl$^-$ and Na$^+$ for predicting IWQI class. Random Forest Classifier performed the best followed by Gradient Boosting and Neural Networks. The water quality index used here is mainly representation of salinity of water and the dataset used here is of groundwater. Other parameters such as acidity, oxygen demand etc. can be used in addition to the parameters used here to develop an index covering more factors. Also, dataset covering different types of water bodies can be used to improve the model. This work can be incorporated in precision agricultural systems which can help in saving cost of lab tests for water quality.

TABLE VII
Classification results for algorithm IV-B2

Methods	Accuracy	Precision	Recall	F1
DecisionTree	0.832	0.728	0.740	0.733
Naive Bayes	0.526	0.105	0.200	0.138
Gradient Boosting	0.858	0.762	0.757	0.760
Random Forest	**0.869**	**0.790**	**0.765**	**0.776**
SVM	0.845	0.762	0.711	0.726
Bagging	0.813	0.750	0.664	0.690
MLP	0.846	0.734	0.743	0.738

TABLE VIII
Five fold cross validation scores

Methods	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Bagging Classifier	0.829	0.823	0.814	0.830	0.813
DecisionTree	0.833	0.836	0.828	0.823	0.839
Naive Bayes	0.518	0.518	0.518	0.518	0.519
Gradient Boosting	0.862	0.866	0.869	0.859	0.862
Random Forest	**0.866**	**0.870**	**0.870**	**0.862**	**0.871**
SVM	0.853	0.849	0.846	0.841	0.848
MLP	0.847	0.853	0.857	0.847	0.849

References

[1] S. Sunder, "India economic survey 2018: Farmers gain as agriculture mechanisation speeds up, but more r&d needed," *The Financial Express*, 2018.

[2] N. Zhang, M. Wang, and N. Wang, "Precision agriculture—a worldwide overview," *Computers and electronics in agriculture*, vol. 36, no. 2-3, pp. 113–132, 2002.

[3] D. La Mora-Orozco, H. Flores-Lopez, H. Rubio-Arias, A. Chavez-Duran, J. Ochoa-Rivero *et al.*, "Developing a water quality index (wqi) for an irrigation dam," *International journal of environmental research and public health*, vol. 14, no. 5, p. 439, 2017.

[4] J. L. Lerios and M. V. Villarica, "Pattern extraction of water quality prediction using machine learning algorithms of water reservoir," *International Journal of Mechanical Engineering and Robotics Research*, vol. 8, no. 6, 2019.

[5] M. E. Grismer and K. M. Bali, "Drought tip: Use of saline drain water for crop production," 2015.

[6] P. Shrivastava and R. Kumar, "Soil salinity: a serious environmental issue and plant growth promoting bacteria as one of the tools for its alleviation," *Saudi journal of biological sciences*, vol. 22, no. 2, pp. 123–131, 2015.

[7] A. Vij, S. Vijendra, A. Jain, S. Bajaj, A. Bassi, and A. Sharma, "Iot and machine learning approaches for automation of farm irrigation system," *Procedia Computer Science*, vol. 167, pp. 1250–1257, 2020.

[8] A. C. M. Meireles, E. M. d. Andrade, L. C. G. Chaves, H. Frischkorn, and L. A. Crisostomo, “A new proposal of the classification of irrigation water,” *Revista Ciencia Agronomica*, vol. 41, no. 3, pp. 349–357, 2010.

[9] R. S. Ayers, D. W. Westcot *et al.*, *Water quality for agriculture*. Food and Agriculture Organization of the United Nations Rome, 1985, vol. 29.

[10] M. H. Hassoun *et al.*, *Fundamentals of artificial neural networks*. MIT press, 1995.

[11] W. S. Noble, “What is a support vector machine?” *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[12] J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” *Annals of statistics*, pp. 1189–1232, 2001.

[13] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[14] Y.-Y. Song and L. Ying, “Decision tree methods: applications for classification and prediction,” *Shanghai archives of psychiatry*, vol. 27, no. 2, p. 130, 2015.

[15] K. P. Murphy *et al.*, “Naive bayes classifiers,” *University of British Columbia*, vol. 18, p. 60, 2006.

[16] L. Breiman, “Bagging predictors,” *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[17] S. Qi and A. Harris, “Geochemical database for the brackish groundwater assessment of the united states,” *US Geological Survey data release*, 2017.

[18] W. McKinney *et al.*, “pandas: a foundational python library for data analysis and statistics,” *Python for High Performance and Scientific Computing*, vol. 14, no. 9, pp. 1–9, 2011.

[19] S. Van Der Walt, S. C. Colbert, and G. Varoquaux, “The numpy array: a structure for efficient numerical computation,” *Computing in science & engineering*, vol. 13, no. 2, pp. 22–30, 2011.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[21] M. Hossin and M. Sulaiman, “A review on evaluation metrics for data classification evaluations,” *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, p. 1, 2015.