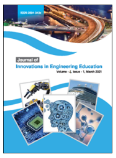

Journal of Innovations in Engineering Education
JIEE 2022, Vol. 5, Issue 1, pp 38-46.

https://doi.org/10.3126/jiee.v5i1. 42265

THAPATHALI CAMPUS
Institute of Engineering
Tribhuvan University

Comparison of machine learning algorithms in statistically imputed water potability dataset

Diwash Poudel[a,*], Dhadkan Shrestha[a], Sulove Bhattarai[a] and Abhishek Ghimire[a]

[a]*Department of Electronics and Computer Engineering, Thapathali Campus, Institute of Engineering, Tribhuvan University, Thapathali, Kathmandu, Nepal*

ARTICLE INFO

Article history:

Received 11 Jan 2022
Received in revised form
19 March 2022
Accepted 18 April 2022

Keywords:

ANN
K Nearest neighbour
LR
Missing data
RF

Abstract

Lack of safe drinking water is a growing concern in the present day and age. Since missing data is commonly found among most of the available datasets, the main purpose of this study is to find the best algorithm that works in the dataset that is statistically imputed and find the algorithm that gives the best prediction on whether water is potable or not. Water potability is predicted using its datasets with the help of the four algorithms evaluating nine features. Some values of the three features, specifically pH, chloramine, and trihalomethane, are found to be missing in the dataset. Missing values are filled in by the median of that particular feature. The performance of machine learning algorithms called LR, K-NN, RF, and ANN is compared in these given conditions. As per our research, RF, with 700 decision trees at a maximum depth of 30, is found to be the best-performing algorithm for the statically imputed water potability dataset. The study most certainly answers the question concerning the best algorithm, but still, further study is needed to optimize the algorithm in order to provide the best prediction.

1. Introduction

Machine learning is the way of programming computers so that they can learn and adapt from the data available to them without any instructions. Machine Learning algorithms learn from their experiences, measure their performance, and improve with experience. These algorithms help to identify patterns and behaviors of data that are not apparent. Machine Learning is the ability to automatically adapt to the changes in the data to get a desired range of output. Machine learning involves supervised, unsupervised, semi-supervised, and reinforcement learning procedures [1].

Missing data and information are frequently encountered in datasets used in machine learning. Such conditions can drastically impact the quality of the machine learning model. It is better to impute those missing values using inference from the available data in the datasets prior to the learning process. There are different imputation techniques such as imputation using stochastic regression imputation, imputation using mean or median, imputation using k-NN. Using these techniques, the missing values can be filled up as if they were actually observed values.

Water is one of the most important elements needed for the continuity of existence of life on this planet. Global species are highly contingent on the bodies of water for their existence. However, issues like increased urbanization, economic growth, rapid industrial expansion, etc. are responsible for polluting different bodies of water on the earth. Each year, 485 000 diarrhoeal fatalities are estimated to be caused by contaminated drinking water. 80% of the health problems in developing countries are associated with contaminated water [2]. Hence, the quality or potability of the water must be tested using an effective procedure.

Potable water simply means water that is safe for human consumption. Potability Test of water is done to know whether it is drinkable or not. U.S. Environmental Protection Agency (EPA) has introduced maximum contaminants levels for various contaminants. Limit has been set on over 90 contaminants in drinking water by EPA [3]. Nepal Government has also set the limits for 27 contaminants under the Water Resources Act, 2005,

[*] Corresponding author:
poudeldiwash13@gmail.com (D. Poudel)

Clause 18, and Sub Clause [4].

Safe drinking water is essential to health, so effective policies should be made by all nations regarding the water potability section. As water is one of the crucial elements for our survival, it is obvious to have healthy drinking water to be safe from water-borne diseases. The potability of drinking water can be determined using machine learning algorithms as well. Parameters used in this project to analyze the potability of water are:

A. Ph value

The pH value of water is a measure of its acidity or alkalinity. World Health Organization (WHO) has recommended 6.5- 8.5 as the range of pH value which is suitable for drinking [5].

B. Hardness

Water hardness is linked with the amount of dissolved calcium and magnesium in water. Traditionally, it is defined as the capacity of water to react with soap. The limit of hardness generally lies between 120 to 170 mg/L.

C. Total dissolved solid (TDS)

Total Dissolved Solid deals with the mineral content in the water and include dissolved organic materials. TDS is the total amount of substances remaining after the evaporation of water. Many minerals are water-soluble, so they can accumulate with water. The limit for TDS which is acceptable for drinking water is 1000 mg/L [6].

D. Chloramine

Chloramine is the combination of chlorine with ammonia. The mixture of ammonia helps chlorine to sustain longer in the solution. To kill microorganisms, chlorine is added to the water. EPA in the USA has established a maximum level for chloramine to be 4 mg/L [7].

E. Sulfate

Sulfates are found in nature in various minerals and are typically used in the industrial sector. Sulfates are available in air, water, food, and plants. According to the Environmental Protection Agency (EPA), the limit for sulfate is set to be 500 mg/L which is safe for drinking [8].

F. Conductivity

Pure water is a bad conductor of electric current. Dissolved impurities like salts and chemicals increase the conductivity of water. Every water body has its level of conductivity. It is necessary to measure the conductivity of water to infer if impurities have been present in water sources. According to EPA, the electrical conductivity of water should not cross the limit of 500 $\mu S/cm$ [9].

G. Total organic carbon

Total Organic Carbon (TOC) is the measure of the quantity of carbon in the organic compounds present in water. Total Organic Carbon test does not extract out the exact carbon-containing compounds present in water but provides information on the amount of carbon present in it [10].

H. Trihalomethanes

Trihalomethanes are present in water that is treated with chlorine. It is obtained as a result of the water treatment process. Trihalomethanes are produced with the reaction of natural organic material found in water with chlorine used to treat water. Trihalomethanes are toxic to the liver, kidneys, reproductive system and are a potential risk to cancer if consumed for a long time. The acceptable concentration of trihalomethanes for drinking water is 0.1 mg/L [11].

I. Turbidity

Turbidity is the measure of light getting scattered and absorbed by suspended sediment in the water. It measures the cloudiness of water. High turbidity of water reduces the beauty of water sources. High turbidity harms the fish and other aquatic life in water. World Health Organization has established the limit for turbidity of water that should not cross more than 5 Nephelometric Turbidity Units (NTU) [12].

After testing and analysis of all the aforementioned parameters, one can conclude whether the given quality sample is potable or not.

2. Related Works

Machine learning was branched as a subdivision of AI in the late 1970. Machine learning has been used to analyze sales data, fraud detection, natural language processing, and product recommendation along with the prediction of different events [13].

Similarly, machine learning is being used to determine the potability of water. As a consequence of increased urbanization and industrialization, water potability evaluation is one of the crucial tasks in guaranteeing the purity of drinking water. To check groundwater quality, 8 artificial intelligence algorithms like random forest (RF), artificial neural network (ANN), M5P tree (M5P), random subspace (RSS), additive regression (AR), support vector regression (SVR), Multilinear regression (MLR), and locally weighted linear regression (LWLR) were used to determine water quality index (WQI) in Illizi region [14]. Performance of machine learning techniques like the artificial neural network, group method of data handling (GMDH), and support vector machine (SVM) were studied to predict water quality components of Tireh River located in the southwest of Iran where it was found that GMDH accuracy was less compared to ANN and SVM [15]. Nonlinear autoregressive neural network (NARNET) and long short-term memory (LSTM) deep learning methods have been developed to determine the water quality index [16]. Sustainable use of water resources was evaluated using Random Forest for Lishui River Basin in China where results were more accurate and stable than ANN and SVM [17]. The group of researchers from Penn State implemented artificial intelligence to determine the quality of the river with respect to human disturbances and climate change using dissolved oxygen as the main indicator to support aquatic life [18]. Researchers from the University of Stirling used a new algorithm known as the meta-learning method that uses data from satellite sensors to monitor the quality of water bodies, determining shifts of quality due to pollution or climate change [19]. Hence, we compared 4 different machine learning techniques to predict the potability of water.

3. Methodology

This paper is aimed at comparing the different highly used machine learning techniques on the basis of their abilities to perform efficiently using statistically imputed datasets. We used Logistic Regression, K-Nearest Neighbors, Artificial Neural Networks, and Random Forest techniques to obtain the prediction. A technique was established to achieve the research's goal, which involved the collection of data sets, data preprocessing, building the model using LR, KNN, ANN, and RF, training the model using the training dataset, and evaluating the built model with the help of a testing dataset.

3.1. Data generating and preprocessing

After the finalization of the objectives of this undertaking, data suitable for our project were collected. The quest to find a suitable dataset for this research began with the reading of similar research of the past, browsing the internet, and going through several other resources. Through the initial research, we collected 5 datasets in total and started examining them. Finally, we chose the dataset called "Water Quality" from the Kaggle website for water quality prediction, which was then used to train and test the model. There is a total of 3276 data in the dataset. In which there are 10 columns, 9 are for features of the water and 1 for "potability". The "potability" value is either 0 or 1, 0 for water is not drinkable or 1 for water is drinkable. The dataset contains 1998 features for not drinkable and 1278 for drinkable. The values missing for the 3 features were about 14% for "pH", 24% for "Sulfate", and 5% for the "Trihalomethanes" feature.

We tried to run the algorithms with NaN values. However, algorithms won't work with NaN values because NaN is an undefined or unpresentable result. Mathematical operations involving a NaN will either return a NaN or raise an exception. Hence, mathematical operations cannot be formed with NaN values [20].

Therefore, to run a mathematical operation, we either need to remove the row containing NaN, which is not ideal because it reduces valuable information, or we can use the statistical imputation technique.

Statistical imputation is a technique to fill missing values by replacing NaN with mean, median, mode, or other central tendency measuring methods.

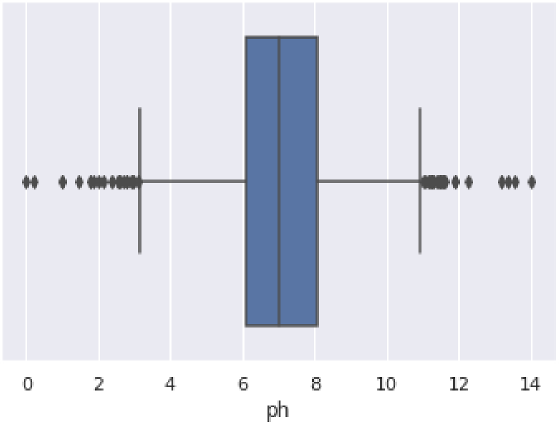

Figure 1: Box plot of 'pH' feature

From the box plot, as shown in Figure 1, we can observe that there are lots of outliers in the 'ph' feature of the dataset. The same goes for the "Sulfate" and "Trihalomethanes" features. In such cases, where there are lots of outliers in the dataset, the median is the best imputation method [21]. As a result, the median depends primarily on the order of data. As a result, it is least affected by outliers [22].

We then filled missing values using the median value of that particular feature. For the "ph" feature the median was calculated as 7.036, for the "sulfate" feature median was calculated as 333.0735457 and for the "Trihalomethanes" feature median was calculated as 66.6224851 and these values were filled in the missing values. Furthermore, we separated the 80% dataset as a training dataset to build the models and the 20% testing dataset to check the accuracy of the models.

3.2. Software Requirements

To conduct this project, different kinds of python libraries were used. Pandas (1.3.5) is used to analyze and

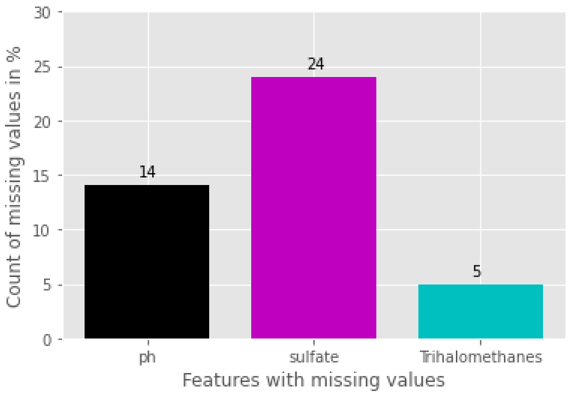

Figure 2: Comparison of % of missing values among different features

manipulate data and NumPy (1.21.5) is used to perform various mathematical operations on our array of data. Matplotlib (3.2.2) library is used to plot graphs and create interactive visualization whereas seaborn (0.11.2) is used to assist matplotlib in statistical graphs.

For machine learning purposes, TensorFlow (2.8.0), as well as the popular sci-kit learn (1.0.2) package allows different ML tasks like classification, prediction, etc. Keras uses TensorFlow to use different kinds of models like sequential and dense models. We chose sci-kit learn to preprocess the data and helped in clustering, classification, regression, and splitting our dataset into train and test sets.

3.3. Machine Learning Algorithms

After the preprocessing phase, we created two Data Frames, X that contains all features except 'Potability', and Y that contains only the 'Potability' feature. After that, we started to scale X and gave it the name X_scaled, and we searched for its maximum and minimum values in one set. We reshaped the Y with (-1, 1), which will reshape the data in such a way that it contains only one column, and the number of rows counts to multiple original rows and columns. We used sklearn train test split to split the x_scaled and y_scaled data into X_train, X_test, y_train, and y_test with a test size of 0.2. We used Tensorflow and Keras to build the model for ANN and we used Sklearn to build a model of K-nearest neighbor, Logistic Regression, and Random Forest.

3.3.1. Artificial Neural Network

Artificial Neural Network (ANN) is useful for establishing the relationship between input and output based on the provided data. ANN is capable of learning and finding out the results from a complex data set. ANN is used to create a model which will assist to predict the events based on learning. It has been solving a real-life problem that cannot be solved by subjective and traditional methods [23].

Artificial Neural Network is the resemblance of the biological neural network which consists of millions of neurons. ANN attempts to bring life to the machine with help of machine learning algorithms [24]. The brain consists of the number of neurons that form a network through synaptic connections. Neurons are used to receive and transmit a piece of information and signal to various parts for proper functioning. The brain consists of dendrites that act as a receiver, neurons as a biological information processor, and soma sums up the input signal. Synapse is located at the end of soma. If the total signal exceeds the threshold limits, it will fire the signal to the axon. Axon transmits the neural signal to the other cells of the body. [25] Similarly, a basic neural network consists of the following:

Inputs: Inputs are considered as an attribute and each attribute has values. These values come from an environment, data sets.

Weights: Weights are the values that signify the importance of the inputs, and features for analysis and prediction.

Bias: Bias helps to shift the results along with the weighted sum of inputs of the neuron. Without bias, neural network model has a limited movement; it introduces flexibility in the neural network.

Activation function: The output of the neural network is determined by the activation function. It is a function used to obtain the output of any node. The resulting outputs are mapped from -1 to 1. The activation function determines whether the receiving information is relevant or irrelevant. Different types of activation functions are step function, ramp function, sigmoid function. For this research we have used ReLU and SIGMOID activation functions while choosing the ANN method. The reason to choose the sigmoid function is that it gives out output in the range of 0 to 1. And the reason to choose ReLU is that this function is fast and simple as well as it does not trigger all the neurons at the same time. ReLU activation function is also known as piecewise activation function because it will give input as output directly if it is positive or else it will give zero output. With the help of ReLU, we can overcome the vanishing gradient problem which helps the model to perform and learn at a quick speed.

In our ANN model, we selected sequential models with a dense layer. For preprocessing, we chose MinMaxS-

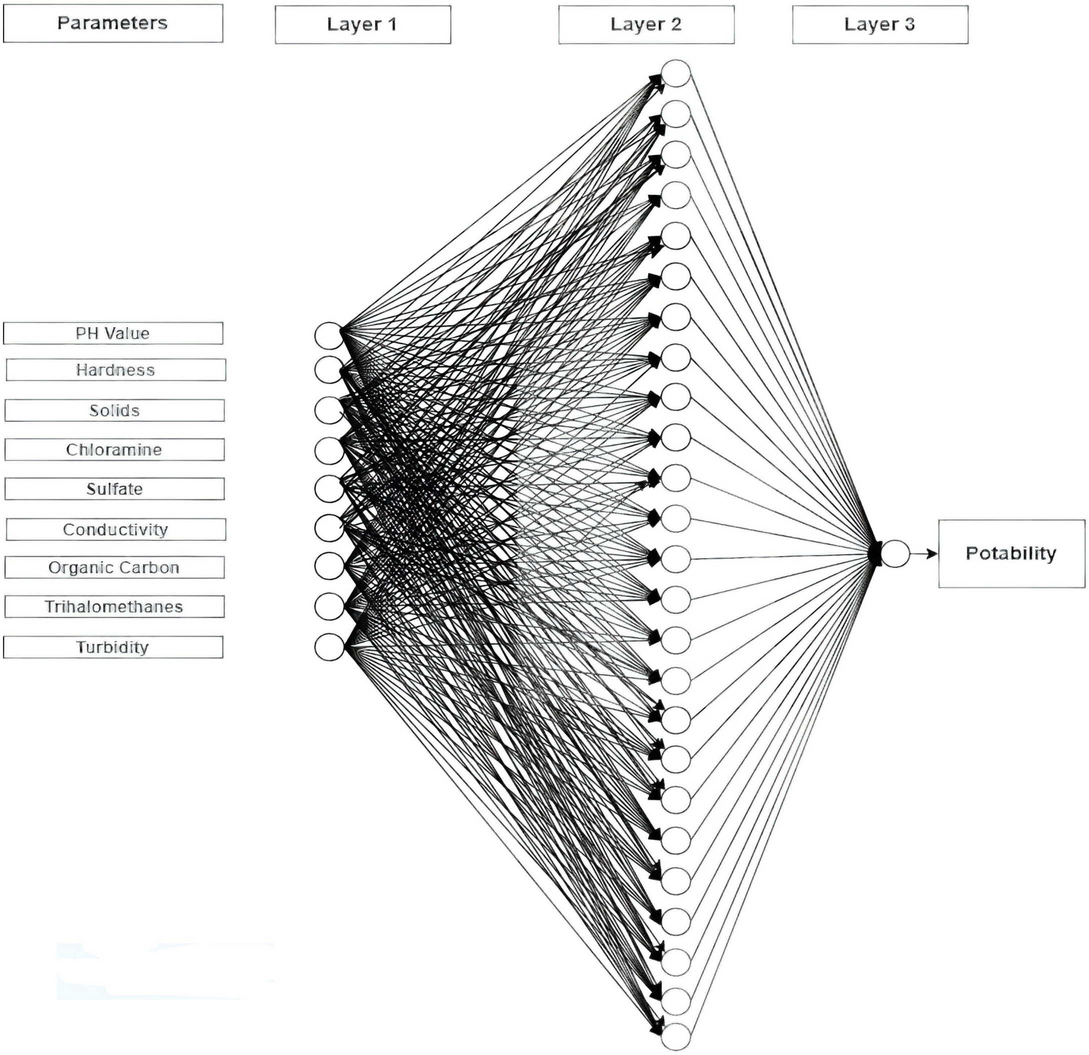

Figure 3: Structure of ANN

caler. There is a total of 3 layers in the model. 1st layer was the input layer consisting of 9 units because of the input dimension of 9 and a Relu activation unit. The hidden layer consisted of 25 units with Relu activation. The final or Output layer consisted of 1 unit with sigmoid activation. We compiled the model with adam optimizer, mean squared error as loss, and accuracy as a metric. Finally, the model was fitted with X_train, Y_train. A total of 150 epochs were used and the size of the batch was 30.

Figure 3 represents a model of the ANN. This neural network has 9 neurons in its input layer, and they are pH value, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic carbon, Trihalomethanes, and Turbidity. There is a single hidden layer with 25 neurons and all the inputs are connected individually to these neurons, as shown in the figures. At the end of each layer, we have used different types of activation functions that determine whether the neurons are supposed to be activated or not based on weight and bias. Another function of this activation function was to produce a nonlinear output. We used the Relu activation function at the first and second layers and the sigmoid activation function at the end layer. The reason for choosing the Relu function was that it produces output for positive input and zero output for negative inputs. We used the sigmoid activation function at the output layer so that it would produce output in the range of 0 to 1. For this purpose, we used different functions like tanh, leaky relu, etc. but none of them were as good as these two.

3.3.2. *Random Forest*

Random Forest is a supervised machine learning algorithm that is applied using an arrangement of decision trees that have been produced using random sets of data. Ensemble methods are generally used to reduce variance and avoid overfitting. It is a widely used machine learning algorithm for feature selection, regression, and classification. Random Forest does not require pre-processing the data which is its major advantage [26].

While the decision tree makes a conclusion based on a single parameter, the random forest takes the average of all the decision trees to conclude a result. Random Forest was developed by Leo Breiman in 2001 and was modified from the concept of bagging that includes random selection without replacement from training datasets [27]. Each tree is constructed in different bootstrap samples. The process of the growing tree is described below: [28]

- Sample the N cases at random if there is N number of cases in the training datasets using with replacement.
- If there are K input features, a number k«K is allocated at each node.
- K features are selected at random from K, and the best feature among the k is used to split the node.
- Each tree is grown to the largest size possible without pruning.

Random Forest is based on the majority of voting among the group of the decision tree, hence named as forest.

Out-of-bag (OOB) error is an error estimation parameter that calculates the error on samples that were not selected during bootstrap sampling. To understand this, let us consider a training datasets D = [a1, a2, a3, a4]. Then, during bootstrap sampling random sample is taken from the training datasets as D1= [a1, a2, a1, a1]. About one-third of the data are left during bootstrap sampling for the construction of trees. Approximation of an error during training is performed by out-of-bag error. While choosing a random bootstrap sample, it leaves some of the observations like a3, a4 as shown in D1, and left observation is known as OOB sample. Parameter tuning is based on the finding of parameters that would produce low OOB error [29]. Random Forest provides remarkable improvement on the performance over decision trees and solves the problem of overfitting.

In the model of Random Forest, we tried different combinations of the number of trees and max depth and concluded that the number of trees should be 700 and the max depth be 30. In the Random Forest algorithm, 700 trees gave their predictions of what the outcome should be, and the majority vote was taken from those 700 trees whose outcomes got the most votes, thus becoming the outcome of the Random Forest.

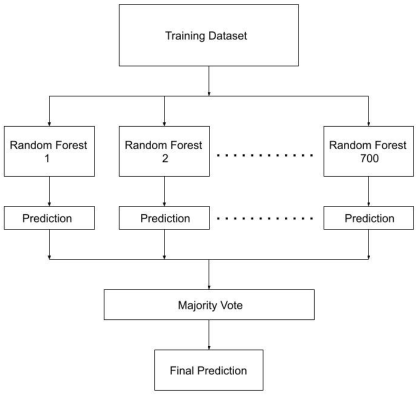

Figure 4: Block diagram of Random Forest Algorithm

3.3.3. *K-Nearest Neighbor (KNN)*

K Nearest Neighbor algorithm is one of the supervised learning machine algorithms that can be used for classification and regression. Resampling datasets as well as imputing missing values can be done with this algorithm. Since it stores the dataset and performs operations on the dataset during classification instead of learning straight from the training set, it is widely known as the lazy learning algorithm. This algorithm supposes that the similar things appear in close proximity [30]. KNN calculates the distance between points to find the closeness between data. Euclidean distance is used to find the distance between data points which is simply given by the following equation, Equation 1.

$$D(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + ... + (a_n - b_n)^2} \tag{1}$$

Where, subjects to be compared are ‘a’ and ‘b’.

The selection of the k factor plays important role in the performance of the model. Larger the k less the impact of variance caused by random error but it may ignore the small yet important patterns [31]. The value of k should be chosen in such a way as to maintain the balance between overfitting and underfitting.

In K-NN, we tried different values of Number of Neighbors (K) and got the best result in K as 10.

3.3.4. *Logistic Regression (LR)*

Logistic regression is one of the supervised machine learning algorithms which is used in the field of classification to predict the probability of target variable. It is used to explain the relationship between the dependent variable and one or more independent variables. Since it is used to solve classification problems, the output of this algorithm ranges from 0 to 1 [32]. It can be used in different tasks like fraud detection, disease diagnosis, and emergency detection. Logistic regression uses a mathematical function known as the sigmoid function that maps the predicted values to probabilities. The sigmoid function is given as:

Its output becomes 0 or 1 for any range of real value in its input. Logistic regression uses the concept of threshold in such a way that any value greater than threshold results in 1 and the value below it results in zero [33]. In our model of LR, we had tried to tune the hyperparameter of LR called Inverse Regularization Parameter(C) by trying out different values and got the best result when C was 0.1.

Table 1: Hyperparameter values of all algorithm

Algorithms	Hyperparameters	Values
LR	Inverse Regularization Parameter (C)	0.1
K-NN	Number of Neighbors (K)	10
RF	Number of Trees	700
	Max depth	30
ANN	Number of features in hidden (second) layer	25
	Activation functions	Relu (In first & second layer), Sigmoid (in the final layer)
	Ephochs	150
	Batch size	30 per epoch

4. Result and Analysis

As stated earlier, the four algorithms or models called LR, K-NN, RF, and ANN were tested. Among the four algorithms compared, the accuracy of LR is 60.51%, and K-NN is 60.98%. Similarly, the accuracy of ANN is 69.5% and that of RF is 70.42%. From the results of accuracy testing, it is clear that K-NN and LR have significantly lower accuracy than ANN and RF, whereas RF has the highest accuracy. Since ANN and RF accuracy margins are very small, we need other performance metrics to reach a clear conclusion.

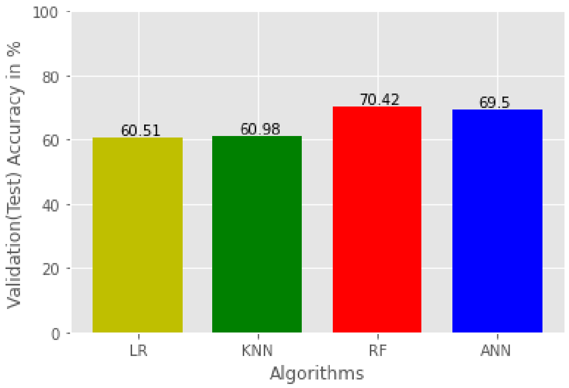

Figure 5: Accuracy comparison of various algorithms

Comparison of algorithms was done with confusion matrices, precision, and recall values. We also compared AUC by plotting the ROC curve of all the algorithms.

Table 2: Confusion matrix of RF algorithm

		Actual	
		Positive	Negative
Predicted	Positive	366	51
	Negative	416	96

Here, Table 2 shows the confusion matrix of the RF algorithm. From the confusion matrix, we can tell that it predicted 366 true positives and 96 true negatives. It made 51 type-1 errors, also called false positives, as well as 143 type-2 errors, also called false negatives.

Table 3: Confusion matrix of ANN algorithm

		Actual	
		Positive	Negative
Predicted	Positive	350	67
	Negative	136	103

Here, Table 3 depicts the confusion matrix of the ANN algorithm. From the confusion matrix, we can confirm that it predicted 350 true positives and 103 true negatives. It made 67 type-1 errors, also called false positives, and 136 type-2 errors, also called false negatives.

The confusion matrix of the K-NN algorithm is presented in Table 4. It informs us that it predicted 368 true positives and 53 true negatives. It made 49 type-1 errors, also called false positives, as well as 186 type-2 errors, also called false negatives.

The confusion matrix of the LR algorithm is depicted in

Table 4: Confusion matrix of K-NN algorithm

		Actual	
		Positive	Negative
Predicted	Positive	368	49
	Negative	186	53

Table 5: Confusion matrix of K-NN algorithm

		Actual	
		Positive	Negative
Predicted	Positive	417	0
	Negative	239	0

Table 5. From the confusion matrix, we can posit that it predicted 417 true positives and 0 true negatives. It made 0 type-1 errors, also called false positives, as well as 239 type-2 errors, also called false negatives.

From the confusion matrix of LR, it can be observed that the LR only predicted "potability" irrespective of any data given to it. Hence, it has performed worse than the other 3 algorithms. From our observation, LR is practically useless in the case of this research.

Among all the algorithms, K-NN predicted the most true positives but predicted the least true negatives. It has also made most type 2 errors. By comparing RF and ANN, we can see that RF has predicted more true positives than ANN, but ANN has predicted more true negatives than RF. RF has made the least type 1 errors, but more type 2 errors compared to ANN.

From the confusion matrix, it can be observed that the most correct prediction was made by RF, which made 462 correct predictions out of 656 total test data, followed by ANN, which made 453 correct predictions out of 656 total test data, and the least was made by K-NN, with 421 correct predictions out of 656 total test data.

Here, Figure 6 depicts the ROC curve of LR, K-NN, RF, and ANN. It can be discerned that the AUC score of LR was 0.5079, K-NN was 0.6273, ANN had 0.6884, and RF had 0.6971. From the results of the AUC score, we can see that K-NN and LR had significantly lower scores than ANN and RF. Here, RF has the highest AUC.

From the above table, Table 6, it can be observed that the LR has a Precision of 1 because it didn't predict any values as negative which is 0 true negative and 0 types 1 error. ANN has the highest precision and F1 score, whereas RF has the highest recall score.

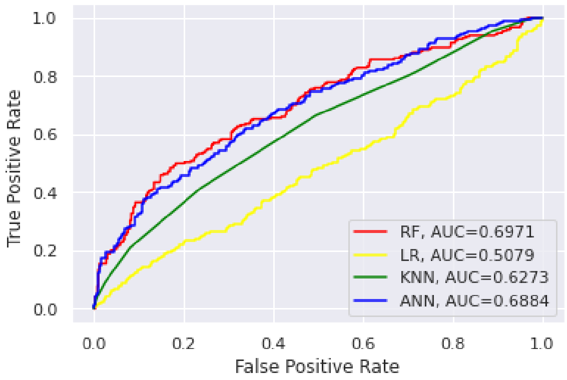

Figure 6: Comparison of ROC Curve between different algorithms

Table 6: Performance metrics comparison of various algorithms

Algorithm	Accuracy	Precision	Recall	F1	AUC
LR	0.6051	1	0.6357	0.7773	0.5079
K-NN	0.6098	0.8825	0.6643	0.7580	0.6273
RF	0.7042	0.8393	0.7202	0.7752	0.6971
ANN	0.6950	0.8777	0.7191	0.7905	0.6884

5. Conclusion

The selection of an appropriate machine learning model or algorithm considering the nature of the work can significantly impact the efficiency of the machine learning procedure. Hence, the use of any machine learning technique or algorithm should be carefully considered based on the nature of the task that needs to be performed. The use of different ML algorithms in the statistically imputed dataset was successfully demonstrated in this research work. We discussed a variety of machine learning algorithms along with their methodology. We calculated the accuracy, confusion matrix for a better understanding of the True Positive, False Positive, True Negative, and False Negative occurrences, precision, and recall, as well as plotted the ROC curve.

According to the results obtained, the use of RF for this particular task was found to be more efficient compared to the ANN, LR, and K-NN because RF had the highest accuracy, recall, and AUC scores, which makes it slightly better than other algorithms. The other theoretical benefit RF provides is that it is made up of a different number of decision trees, and decision trees isolate outliers in a limited region of the feature map. That's why decision trees are resistant to outliers. Outliers won't affect the remainder of the predictions since each leaf's prediction is the majority class. In the end, in RF, you don't worry about outliers. This is also an important reason that RF doesn't need much prepossessing.

The main purpose of machine learning is to learn itself and to lessen the burden of programmers' work by learning itself over a course of time. As long as the volume of the data set is small, there is a clear limit on the data set where the machine could learn itself. The algorithms to predict the potability of the water are simple instead of using complex algorithms and deep learning methods.

Future Works

Our future work includes the use of voluminous contemporary datasets, which are conducive to a plethora of information and better implementation of the algorithms. More studies on water quality prediction can also be carried out by adopting hybrid methodologies employing GIS, IoT, and remote sensing algorithms.

Acknowledgemant

We would like to acknowledge the Department of Electronics and Computer Engineering, Thapathali Campus, Tribhuvan University for providing the necessary resources and platform to complete this project. We would like to appreciate the suggestions and insight from Er. Umesh Kanta Ghimire regarding the issues encountered in the course of this research. We thank Er. Amod Khatiwada for his constant guidance and encouragement throughout this undertaking.

Conflict of interest

No conflict of interest

References

[1] Burns E. Techtarget[J/OL]. 2020, 6. https://www.dataversity.net/a-brief-history-of-machine-learning/}.

[2] A snapshot of the world's water quality: Towards a global assessment[M]. United Nations Environment Programme, 2016.

[3] The united states environment protection agency[EB/OL]. 2021. https://www.epa.gov/dwreginfo/drinking-water-regulations.

[4] National drinking water quality standards[M]. Government of Nepal, 2005.

[5] ph in drinking-water[M]. World Health Organization, 2007.

[6] Kent health care products[M/OL]. Kent RO Systems, 2020. https://www.kent.co.in/blog/what-are-total-dissolved-solids-tds-how-to-reduce-them/.

[7] Chloramine in drinking water[M]. World Health Organization, 1998.

[8] United states environmental protection agency[EB/OL]. 2003. https://www.epa.gov/sites/default/files/2014-09/documents/support_cc1_sulfate_healtheffects.pdf.

[9] United states environmental protection agency[EB/OL]. https://archive.epa.gov/water/archive/web/html/vms59.html.

[10] Whitehead P. Elga veolia[J/OL]. 2021. https://www.elgalabwater.com/blog/total-organic-carbon-toc.

[11] Guidelines for canadian drinking water quality: Trihalomethanes[M]. Federal-Provincial-Territorial Committee on Drinking Water of the Federal-Provincial-Territorial Committee on Health and the Environment, Ottawa, 2006.

[12] Water quality and health - review of turbidity: Information for regulators and water suppliers[M]. World Health Organization, 2017.

[13] Foote K D. Dataversity[J/OL]. 2021. https://www.dataversity.net/a-brief-history-of-machine-learning/}.

[14] S. Kouadri A R M T I, A. Elbeltagi, Kateb S. Performance of machine learning methods in predicting water quality index based on irregular data set: application on illizi region (algerian southeast),"[M]. SpringerLink, 2021.

[15] Haghiabi A H, Nasrolahi A H, Parsaie A. Water quality prediction using machine learning methods[J]. IWA Publishing, 2018, 53(1).

[16] Aldhyani T H H, M. Al-Yaari H A, Maashi M. Water quality prediction using artificial intelligence algorithms[J]. Hindawi, 2020, 2020.

[17] Xie C, Chao L, Shi D, et al. Evaluation of sustainable use of water resources based on random forest: A case study in the lishui river basin, central china[J]. Journal of Coastal Research, 2020, 105.

[18] Schley T. Penn state[J/OL]. 2021. https://www.psu.edu/news/research/story/artificial-intelligence-predicts-river-water-quality-weather-data/.

[19] Science daily[EB/OL]. 2021. https://www.sciencedaily.com/releases/2021/05/210504112514.htm.

[20] dataset[EB/OL]. https://datatest.readthedocs.io/en/stable/how-to/nan-values.html}:.

[21] Australian bureau of statistics[EB/OL]. https://www.abs.gov.au/websitedbs/D3310114.nsf/Home/Statistical+Language+-+measures+of+central+tendency}:.

[22] Concepts in statistics[EB/OL]. https://courses.lumenlearning.com/wmopen-concepts-statistics/chapter/mean-and-median-2-of-2/.

[23] J. Ćetković M L M S V J C, S. Lakić, Gogić M. Assessment of the real estate market value in the european market by artificial neural networks application[J]. Hindawi, 2018, 2018.

[24] Hsu H. Computer history museum[J/OL]. 2020. https://computerhistory.org/blog/how-do-neural-network-systems-work/.

[25] Cherry K. very well mind[J/OL]. 2020. https://www.verywellmind.com/structure-of-a-neuron-2794896.

[26] R S E. Analytics vidhya[J/OL]. 2021. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/.

[27] Breiman L. [J/OL]. 2001. https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf.

[28] Breiman L, Cutler A. [J/OL]. https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm}prox.

[29] Schonlau M, Zou R Y. The random forest algorithm for statistical learning for statistical learning[Z]. 2020.

[30] Harrison O. towards data science[J/OL]. 2018. https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761.

[31] Zhang Z. Introduction to machine learning: k-nearest neighbors[J]. Annals of Translational Medicine, 2016, 4(11).

[32] Bonthu H. Analytics vidhya[J/OL]. 2021. https://www.analyticsvidhya.com/blog/2021/07/an-introduction-to-logistic-regression/.

[33] Pant A. towards data science[J/OL]. 2019. https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148.