BioGecko Vol 12 Issue 02 2023

ISSN NO: 2230-5807

SOCIAL SPIDER OPTIMIZATION FOR VISUALIZING AND PREDICTING AQUA POND STATUS

Dr.Ch. Suresh Babu, M.Tech,Ph.d[1],D.Venkata Naga Rupasri[2],G.Sravana Lakshmi[3],E.Sai Kishore[4],P.Dinesh Kiran[5]

1Professor & Mentor, Department of Information Technology , SRGEC, Gudlavalleru,
2Undergraduate Student, Department of Information Technology , SRGEC, Gudlavalleru,
3Undergraduate Student, Department of Information Technology , SRGEC, Gudlavalleru,
4Undergraduate Student, Department of Information Technology , SRGEC, Gudlavalleru,
5Undergraduate Student, Department of Information Technology , SRGEC, Gudlavalleru,

ABSTRACT

Aquaculture is becoming very popular economically. The aquatic Condition Index (WQI), a numerical expression, is used to evaluate the state of an aquatic body. This study's main objective is to assess water quality using machine learning algorithms. The WQI and the following water quality parameters were used to evaluate the project's total water quality. These included the presence of chloramines, trihalomethanes, pH, hardness, particles, conductivity, organic carbon, BOD, sulphates, turbidity, and potability. These variables are used as feature vectors to characterise the water quality.In this project, we used five distinct classification algorithms: Logistic Regression, Decision Tree (DT), K-Nearest Neighbor (KNN), Random Forest (RF), and Social Spider Optimization to forecast the water quality class. Both the real dataset, which includes data from different sources, and a synthetic dataset produced using randomly selected parameters were used in the experiments. The performance of five various classifier types was compared, and it was found that Social Spider Optimization outperforms the others with a 94.1 accuracy rate.
.

1. INTRODUCTION

The quality of the water directly affects both the environment and the welfare of the general population. Drinking, gardening, and manufacturing are just a few uses for water. Due to their accessibility, rivers have been used more frequently for the development of human civilizations than other water sources. The use of other water sources, such as groundwater and seawater, can occasionally be beneficial. For example, using saltwater is usually associated with the spread of pollution, and using groundwater without sufficient recharge will result in land sinking. As a consequence, interest has grown in the use of waterways. Numerous studies into waterways around the world have led to the suggestion of a new branch of engineering known as "river engineering." It is pertinent to conduct research on morphological changes, sediment movement, water quality, and pollution transmission mechanisms.Research on river water quality is a common subject in earth sciences. When assessing the quality of rivers, two methods are taken into account: defining the mechanism of pollution transmission and measuring the components of water quality. Among other water quality indicators, it has been proposed to measure dissolved oxygen (DO), chemical oxygen demand (COD), biological oxygen demand (BOD), electrical conductivity (EC), pH, temperature, K, Na, and Mg. In order to accomplish this, governments have constructed hydrometry sites along urban rivers, agro-industrial projects, industrial estates, and rivers that link dam reservoirs. The evaluation of water quality is a basic stage in developing agricultural initiatives, choosing the type of irrigation system to use, the cropping pattern, and industrial water purification systems.

2. LITERATURE SURVEY

Cao et al. (2019) developed a dissolved oxygen prediction model using grey relational analysis, Ensemble Empirical Mode Decomposition (EEMD), sample entropy and a regularized extreme learning machine (RELM). Initially, the correlation among the dissolved oxygen and various factors were analysed by the grey correlation approach from which the pH and water temperature having a maximal correlation with the dissolved oxygen. Hence, water temperature, dissolved oxygen and the highest correlation are the three factors considered for the model output. Consequently, the three factors were decomposed based on EEMD, and then the sample entropy was employed for analysing the complexity of the subsequence. Here, the higher prediction precision was found better but failed to consider the knowledge of the environmental system with optimal aquaculture environment. Xu and Boyd (2016) designed a regression model to mitigate the monitoring parameters of water quality. The method use minimum variables for evaluating agriculture water quality, but it consumed more memory and time. Peng et al. (2017) developed Primary Component Analysis (PCA) and Fuzzy Neural Network (FNN) of dissolved oxygen in the aquaculture water quality for prediction. This approach employed PCA for extracting PC of the aquaculture ecologicalindexes, and then, input vector dimension was reduced. Also the differential evolutionary algorithm was introduced for optimizing the weight parameter ofFNN for achieving the optimal parameters automatically. The method was not considered other methods for managing and predicting water quality. Barzegar et al. (2018)employed an Extreme Learning Machine (ELM) and hybrid wavelet-extreme learning machine (WA-ELM) for combining the advantages of the WA-ELM based on boosting ensembles. Here, the performance was found better, but the method provides less accurate forecasts.

Antanasijevic et al. (2019) developed a self-organizing network-enabled location similarity index (LSI) with Ward NNs (WNNs) to achieve less complex, and various sites' model was employed to predict dissolved oxygen content. Also, the multilevel splitting method was introduced for monitoring the locations using similarity, and the virtual splitting of the processed data was done in terms of features. The method was effective with only limited inputs. Xu et al. (2017) developed statistical and mechanistic methods for water temperature prediction using Water Temperature Mechanism Model (WTMM). Here, the best parameters were chosen using Improved Artificial Bee Colony (IABC). This approach was utilized for searching the required optimal combinational parameters in WTMM model that solves the blindness and limits the selection of parameters. The method failed to enhance the information management level in aquaculture. Shiet al. (2019).presented Clustering based on Soft plus Extreme Learning Machine method (CSELM) to efficiently and accurately determine the dissolved oxygen change from the time series data. This approach uses k-medoids clustering for grouping datasets into several clusters using Dynamic Time Warping (DTW) distance, and then, the Soft plus ELM was employed for discovering time series pieces. The unnecessary losses were not reduced and avoided, which was caused by the hypoxia. Huan et al. (2018) developed a hybrid decomposition–prediction–reconstruction model that integrates EEMD and Least Squares Support Vector Machine (LSSVM) optimized by the Bayesian evidence method. Initially, the EEMD was pre-processed with dissolved oxygen to reduce the noise with features extraction. After that, the LSSVM was optimized by the Bayesian evidence to obtain a faster convergence rate. The method failed to improve the forecasting accuracy. Gautam and Pagay (2020) provide a review to determine the water status of horticultural crops with the use of remote sensing. The plant's instantaneous response to water stress can be captured using thermal cameras and potentially narrow-band hyper spectral sensors. However, further developments are required to establish crop-specific thresholds of remotely sensed indices. Liu et al. (2019) proposed a water quality forecasting method with the help of LSTM deep neural networks, and this offers a feasible approach for water quality prediction. However, this model only considers single-dimensional input, while there are more complex datasets with many different dimensions for water quality monitoring. Aldhyani et al. (2020) proposed a model by using advanced artificial intelligence algorithms to measure the future water quality. The SVM algorithm has achieved the highest accuracy; however, the analysis should be extended to different types of water.

3. PROBLEM STATEMENT

Therearemanyways topredict theaquastatussomeofthem areusing LSTM, Navie Bayes, The above methods gives less accuracy and the error rate for this method is high. Mainly in existing system we did not use any visualization models.

Limitations

Thepredictedvaluecannotbecorrectalways

4. PROPOSED SYSTEM

The recommended system uses a variety of visualisation models, such as bar graphs, box plots, scatter plots, etc. To assess precision, five different models—Logistic Regression, Decision Tree, K-Nearest Neighbor, Random Forest, and Social Spider Optimization—were used.

FEATURES

With the use of machine learning technique algorithms, a model is built to provide accurate result. The output of this work would help to predict the quality of water.

5. IMPLEMENTATION

5.1 LogisticRegression:

To predict the probability of a specific class or occurrence, logistic regression's "Supervised machine learning" algorithm can be used. When the outcome is binary or dichotomous and the data can be linearly divided, it is used.That means logistic regression is usually used to address problems involving binary classification.

Binary classification is the process of predicting an output variable that is discrete and divided into two categories. Yes/No, Pass/Fail, Win/Lose, Cancerous/Non-cancerous, and many other binary categories are examples. What is the Logistic Regression Process?

Consider a model where "x" is the only predictor, "" is the only Bernoulli response, and "" has a probability ("p") of being equal to 1.

The formula p = b0+b1x >eq 1

represents the linear equation.

The right-hand side of the equation, b0+b1x, can have values outside the range because it is a linear equation. (0,1). But we understand that chance will always fluctuate between (0,1).

In order to avoid that, we predict odds rather than likelihood.

Odds: The ratio of a situation's probability of occurring to its probability of not occurring. Odds = p/(1-p)

Equation 1 being rewritten as p/(1-p) = b0+b1x >eq 2

Only positive numbers can be assigned to odds; to handle negative values, we forecast the odds logarithm. Log of odds equals ln(p/(1-p)) Equation 2 being rewritten as ln(p/(1-p)) = b0+b1x >eq 3

In order to update the parameter p in equation 3, we use exponential on both sides.exp(ln(p/(1-p)) equals exp(b0+b1x)

p/(1-p) equals eln(b0+b1x)

P/(1-p) = e(b0+b1x) by the logarithm inverse method.

P is a simple mathematical expression: (p-1) * e(b0+b1x).

P is equivalent to p / e(b0+b1x)(b0+b1x-p).

In the event that p dominates the right side, then p is equal to p * (e(b0+b1x)/p - e(b0+b1x)).

P equals (1 + e(b0+b1x))/e(b0+b1x).

p = 1 / (1 + e-(b0+b1x)), where e(b0+b1x) is used to split the numerator and denominator on the right side.

The following equation describes a logistic model with n factors:

P = 1/(1 + e-(b0 + b1 x b2 x b3 x + --- +bn x n)

5.2 KNN:

K-Nearest Neighbour is one of the simplest supervised learning-based machine learning methods. The K-NN algorithm puts the new case in the category that resembles the existing categories the most, presuming that the new case and the existing cases are comparable. After storing all of the previous data, a new data point is classified using the K-NN technique based on similarity. This suggests that new data can be accurately and swiftly categorised into the appropriate category using the K-NN algorithm.The K-NN algorithm can be used for regression even though classification problems are where it is most commonly applied. K-NN is a non-parametric approach.which shows that it doesn't depend on any underlying presumptions. As a result of storing the information rather than immediately learning from the training set, it is also known as a lazy learner algorithm. Instead, when classifying data, it utilises the dataset to carry out an action..

5.3 DecisionTree:

Classification and error issues can be resolved using the supervised learning technique known as a decision tree, but this approach is frequently preferred. It is a tree-structured classifier where each child node represents the classification outcome and internal nodes represent the characteristics of a dataset. The two components of a decision tree are the Decision Node and Leaf Node. Decision nodes are used to make decisions and have many branches, whereas Leaf nodes are the outcomes of decisions and do not have any extra branches.

Decisions or tests are made based on the characteristics of the dataset that is given. It is a graphical method of obtaining every possible solution to a problem or choice based on predetermined parameters. It is referred to as a decision tree because, similar to a tree, it starts with the base node and expands by adding more branches to produce a shape approximating a tree.

5.4 RandomForest:

favoured method for machine learning The technique of guided learning includes Random Forest. It can be used for ML problems incorporating both regression and classification. It is based on the concept of ensemble learning, a technique for combining different classifiers to deal with complicated problems and improve model performance. As its name suggests, Random Forest is a classifier that averages several decision trees applied to various sections of the supplied dataset to improve the predictive accuracy of the dataset. Instead of relying solely on one decision tree, the random forest takes the forecast from each tree and bases it on the majority votes of predictions. it predicts the final output. Thegreater number of trees in the forest leads to higher accuracy and prevents the problem ofoverfitting.

5.5 SocialSpiderOptimization:

The SSO thinks of each possible solution in the population as a spider, and it considers the entire search space to be a giant spider web. According to the health value of the treatment it stands in for, each spider is assigned a weight. The method uses two different search sets of evolutionary operators to simulate the various cooperative behaviours presumed in the colony.

The technique was developed to handle a nonlinear global optimisation problem with the following box constraint: minimization of f (x)x = (x1, x2,..., xd) Rd under x X

Where d is a reduced feasible space constrained by the lower (lh) and upper (uh) limits and X = x Rd | (lh x, uh) h = 1,..., and f: A nonlinear function is Rd R.

The SSO utilises a population S of N potential solutions to resolve the optimisation problem. The overall web represents the search space X, and each solution denotes a location of a spider. Male (Ms) and female (F) search agents are divided into the population S in the method. (Fs). The number Nf is composed of females and is randomly selected from between 65 and 95 percent of the overall population S in order to simulate an actual spider colony. Male people make up the remaining Nm (Nm = S - Nf). In these conditions, the Fs group is made up of a collection of female people (Fs = fs1, fs2,..., fs Nf), and the Ms group is made up of male people (Ms = ms1, ms2,..., ms nm), where S = Fs U Ms*(S={s1,s2,...,sN})*.

5.6 ACCURACY:

The accuracy algorithm makes measurement errors easier to comprehend. If the measured value matches the actual value, then the measurement is said to be highly accurate and error-free. Accuracy and error rate cannot coexist. When accuracy is high, the error rate is low, and when accuracy is low, the mistake rate is high. The accuracy algorithm calculates the accuracy as a proportion; the sum of accuracy and error rate is 100%. Accuracy is one metric for rating classification algorithms. Accuracy is the proportion of forecasts that our model accurately predicted. The formal definition of accuracy is as follows:

$$\text{Accuracy}= \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Forbinaryclassification, accuracycanalsobecalculatedintermsofpositivesandnegatives

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Where *TP* = True Positives, *TN* = True Negatives, *FP* = False Positives, and *FN* = FalseNegatives

6. RESULTS

For the data set water Quality we have 9 attributes in it, those nine attributes are appearing in different visualizations like box plot, bar graph, histography ,heatmap, scatter plot etc…

Boxplot

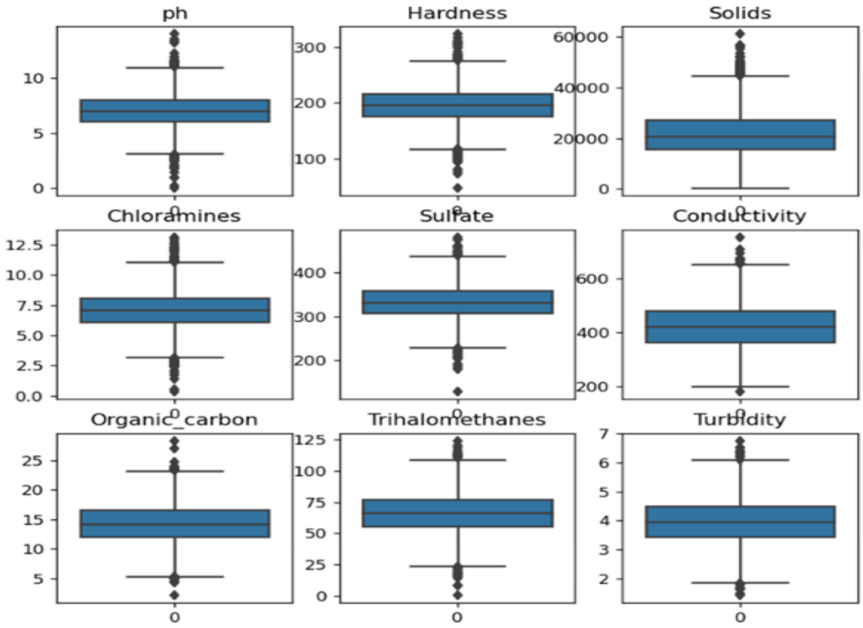
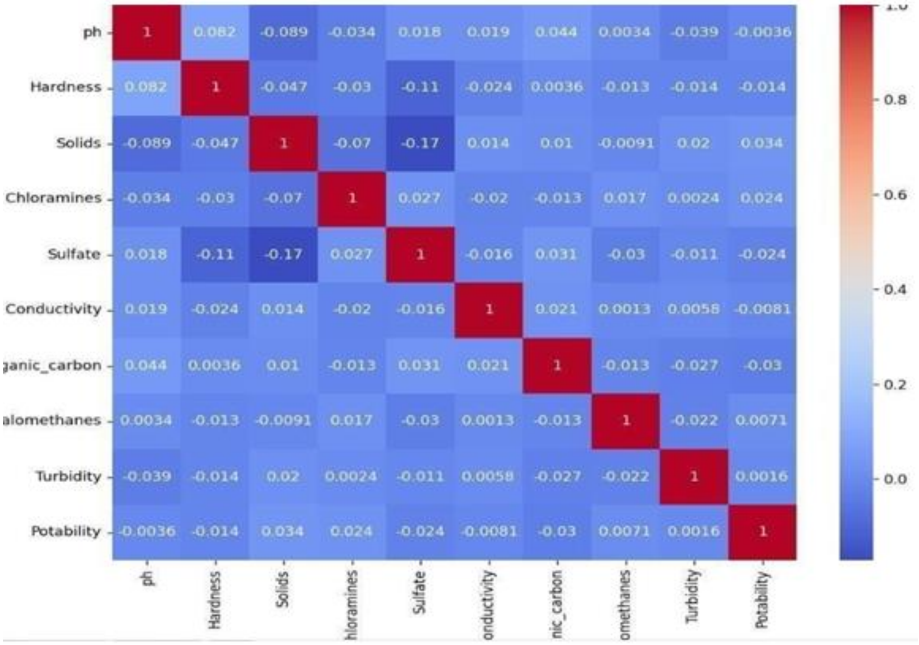


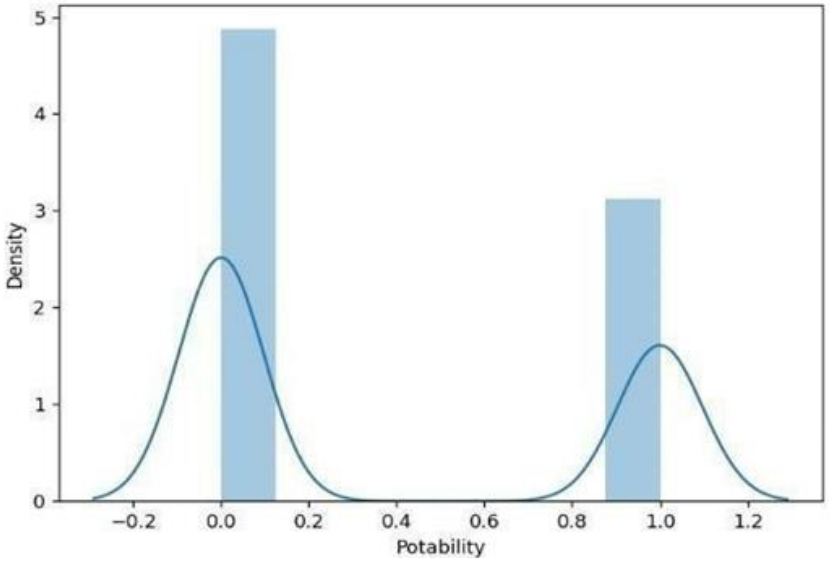
Fig 7.1: Box plot

HeatMap:

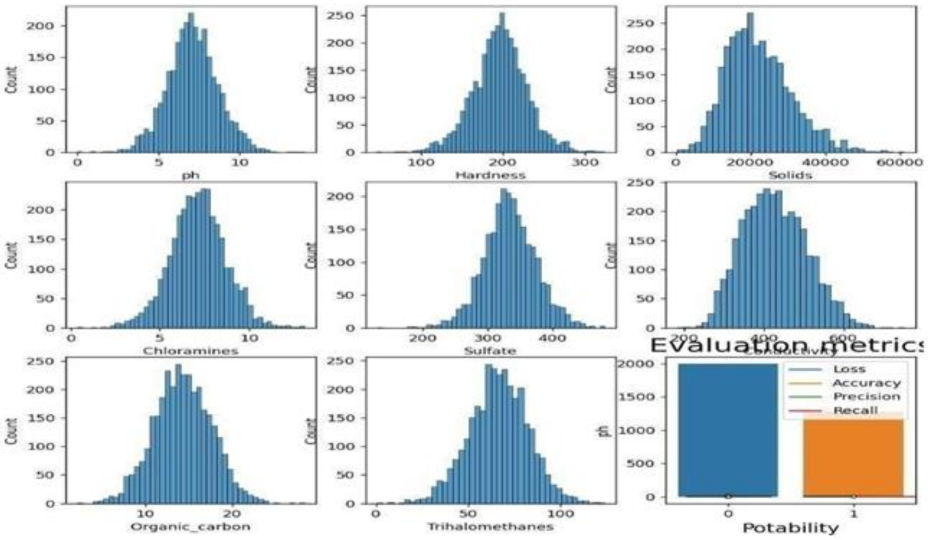


	ph	Hardness	Solids	hloramines	Sulfate	onductivity	nic_carbon	omethanes	Turbidity	Potability
ph	1	0.082	-0.089	-0.034	0.018	0.019	0.044	0.0034	-0.039	-0.0036
Hardness	0.082	1	-0.047	-0.03	-0.11	-0.024	0.0036	-0.013	-0.014	-0.014
Solids	-0.089	-0.047	1	-0.07	-0.17	0.014	0.01	-0.0091	0.02	0.034
Chloramines	-0.034	-0.03	-0.07	1	0.027	-0.02	-0.013	0.017	0.0024	0.024
Sulfate	0.018	-0.11	-0.17	0.027	1	-0.016	0.031	-0.03	-0.011	-0.024
Conductivity	0.019	-0.024	0.014	-0.02	-0.016	1	0.021	0.0013	0.0058	-0.0081
ganic_carbon	0.044	0.0036	0.01	-0.013	0.031	0.021	1	-0.013	-0.027	-0.03
alomethanes	0.0034	-0.013	-0.0091	0.017	-0.03	0.0013	-0.013	1	-0.022	0.0071
Turbidity	-0.039	-0.014	0.02	0.0024	-0.011	0.0058	-0.027	-0.022	1	0.0016
Potability	-0.0036	-0.014	0.034	0.024	-0.024	-0.0081	-0.03	0.0071	0.0016	1

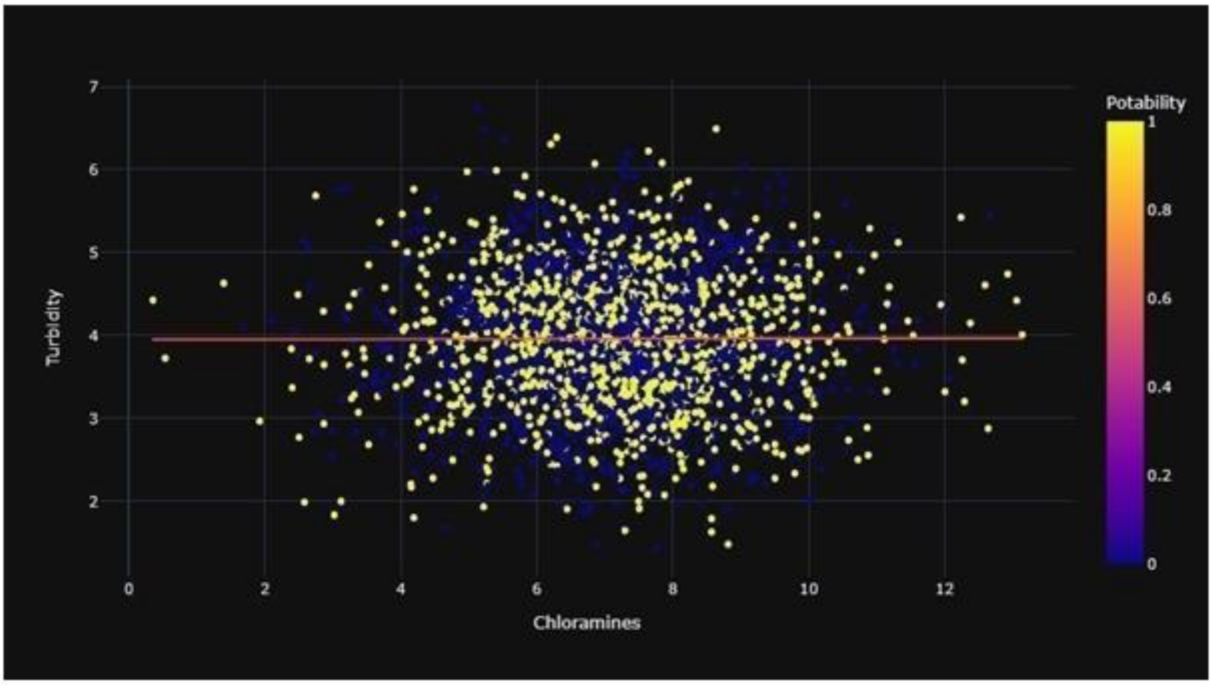
Distribution graph



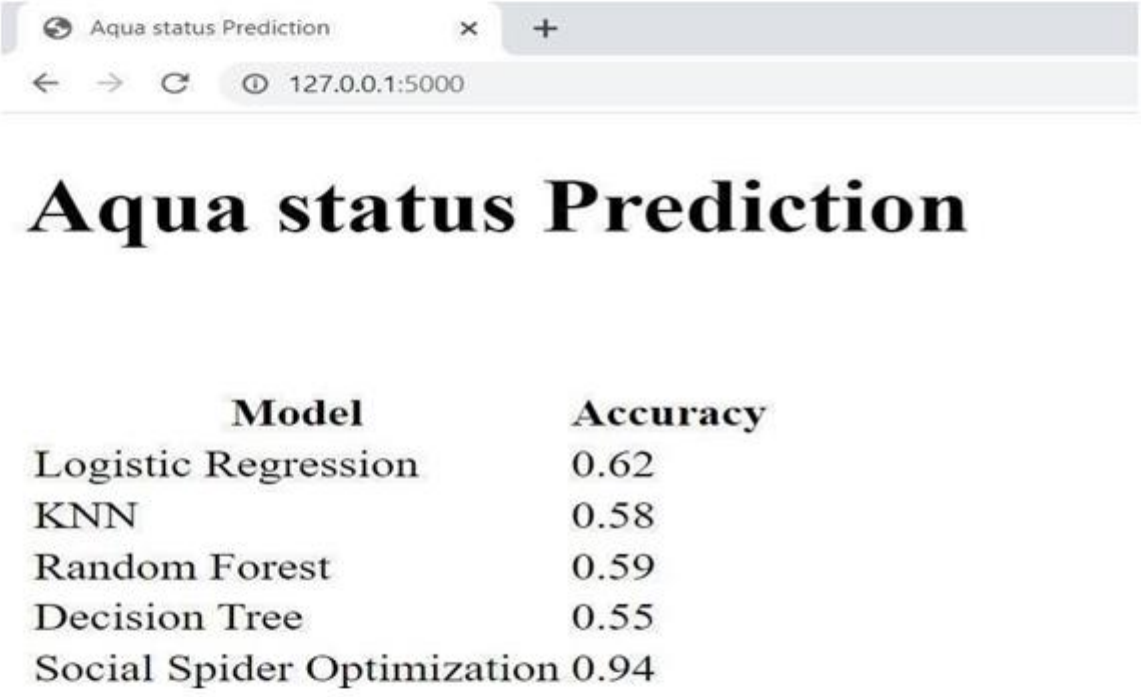
Histplot



Scatterplot



After visualization models we find the accuracy with five different models finally the model Social Spider Optimization gave higher accuracy



Aqua status Prediction

Model	Accuracy
Logistic Regression	0.62
KNN	0.58
Random Forest	0.59
Decision Tree	0.55
Social Spider Optimization	0.94

7. CONCLUSION

The social spider optimisation algorithm has the best results when predicting water quality, with an accuracy rate of 94%. The social spider optimisation algorithm can thus reasonably forecast water quality. This model was intended to serve as a guide for estimating water purity..

8. FUTURE SCOPE

In order to ensure that future works only use the required parameter instruments, we advise incorporating the research's results into a sizable internet of things system. On the basis of the real-time statistics provided by the IoT system, the developed algorithms would instantly project the water quality.

9. *REFERENCES*

1. A. N. Prasad, K. Al Mamun, F. R. Islam, and H. Haqva, “Smart water quality monitoring system,” in Proceedings of the 2nd IEEE Asia Pacific World Congress on Computer Science and Engineering, IEEE, Fiji Islands, December 2015.

2. P. Li and J. Wu, “Drinking water quality and public health,” Exposure and Health, vol. 11, no. 2, pp. 73–79, 2019.

3. Y. Khan and C. S. See, “Predicting and analyzing water quality using machine learning: a comprehensive model,” in Proceedings of the 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT), IEEE, Farmingdale, NY, USA, April 2016.

4. D. N. Khoi, N. T. Quan, D. Q. Linh, P. T. T. Nhi, and N. T. D. Thuy, “Using machine learning models for predicting the water quality index in the La buong river, Vietnam,” Water, vol. 14, no. 10, p. 1552, 2022.

5. U. Ahmed, R. Mumtaz, H. Anwar, A. A. Shah, R. Irfan, and J. García- Nieto, “Efficient water quality prediction using supervised machine learning,” Water, vol. 11, p. 2210, 2019.

6. S. Kouadri, A. Elbeltagi, A. R. M. T. Islam, and S Kateb, “Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast),” Applied Water Science, vol. 11, no. 12, p. 190, 2021.

J. P. Nair and M. S. Vijaya, “Predictive models for river water quality using machine learning and big data techniques - a Survey,” in Proceedings of the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), IEEE, Coimbatore, India, March 2021