Water Quality Prediction and Classification based on Linear Discriminant Analysis and Light Gradient Boosting Machine Classifier Approach

D. Sasikala[1], K. Venkatesh Sharma[2]

[1]Professor, Department of CSE, JB Institute of Engineering & Technology, Moinabad, Hyderabad, Telangana, India
[2]Professor, Department of CSE, CVR College of Engineering, Vastu Nagar, Mangalpally, Hyderabad, Telangana, India

E-mail: [1]godnnature@gmail.com, [2]venkateshsharma.cse@gmail.com

Abstract

Estimating water quality has existed as one of the vital factors embarked on the planet in the present eons. This paper illustrates a water quality estimate based on the Linear Discriminant Analysis (LDA) technique. Weighted arithmetic index technique is used in the computation of the Water Quality Index (WQI). At that moment, the LDA is linked to the dataset, and the ultimate principal WQI dynamics have been determined. Subsequently after predicting the WQI, Light Gradient Boosted Machine (LGBM) classification is performed in the LDA. Lastly, the LGBM classifier is activated to label the water quality. This proposed LGBM with LDA technique is demonstrated and evaluated on a Gulshan Lake-related dataset. The results show 96% forecast accuracy for the LDA and 100% categorization accuracy for the Light Gradient Boosted Machine classifier system that indicate consistent interpretation linked over the futuristic prototypes. This innovative model LDA-LGBM is aimed at enhancing the prediction of water quality and its classification through AI - ML approach.

Keywords: Water quality prediction, linear discriminant analysis, weighted arithmetic index, water quality index, light gradient boosted machine classifier, Gulshan Lake-related dataset

1. Introduction

Water quality is a significant factor in rural that has caused abundant good health in the state of human and wildlife ecosystem, however contrary in urban. Approval of continual water quality checking has been foreseen as the predicament situation, and the various

DOI: https://doi.org/10.36548/jiip.2022.3.002
Received: 03.07.2022, received in revised form: 08.08.2022, accepted: 22.08.2022, published: 01.09.2022

expensive atmosphere of such techniques consistently equalize its determination. Then, ample investigations have been accomplished to forecast water quality using miscellaneous Machine Learning (ML) processes. In this work, a proficient Water Quality Index (WQI) forecast model has been proposed to bring about additional genuine discerning dynamics. Three miscellaneous alternative collaborative prototypes are explored and evaluated, viz: Linear Discriminant Analysis (LDA), LightGBM Classifier, and LDA + LightGBM. These prototypes are combined with LDA - LightGBM to implement the technique in achieving good accuracy outcome and maintain GPU leaning by setting drop dispute. The outcomes indicate that LDA & LightGBM model reach the other prototypes with an R2 value up to 0.91. This blend of LDA - LGBM reliably tallies a superior enhanced R2 values even when skilled with a much lesser input feature and later with a swifter coaching timeline. This LDA - LGBM model will truly afford the robust water quality prediction in an effective and efficient mode.

2. Review of Literature

The representation of Artificial Intelligence (AI) techniques in [2] embracing of GMDH, SVM (Support Vector Machine) and ANN (Artificial Neural Networks) endured and appraised to forecast the water quality in parts of Tireh River (Iran) where the results of GMDH and SVM prototypes were reliable in tallying the valuation with ANN. The forecast by the SGR-WQI over ML practices depicted in [3] was a real technique to inspect and appraise water quality enactment. The prototypes determined the universal enhanced model by the logistic regression techniques, exposed the superior depiction in evaluating WQI and WQC, compatibly. Until now, for key-in decline, the MLR model via 12 dynamics is an enhanced selection.

Paper [4] denoted good deal amid the conclusions after the Piper and Sti-diagrams, and K-means clustering. A toning set of multivariate statistics pooled with geo-spatial exploration was discovered handy for unearthing of hydrogeochemical events in the headway of groundwater. The technique used was favorable for limited water reserve officials for embryonic stratagems to ease and obviate groundwater impurity. In [5], via state-of-the-art AI, water quality prediction and managing were maintained. The ANFIS model attained accuracy in the testing stage, by 96.17% regression constant for forecasting WQI, and the FFNN model achieved the supreme 100% accuracy destined for WQC. The practice put forward in the empirical work [6], utilized a WQI with amazing sensitivity to dynamic values that are exterior of legal confines, likewise runs by a profuse proficiency to categorize midst of specimen points and periods, plainly replicating temporal and spatial conflict, beside making evident yearly cycle and implications in water quality. As a stratagem to alert the state-owned evolution of the Santiago-Guadalajara River, the establishment of a receptive interface by a collaborating tariff was endorsed in contrary of the trends of the SGR-WQI to the universal public and to aid as a tool for the officials of this water reserve to build more firm resolves in its governing.

The work in [7] suggested that SVM is a grander system devising strange accuracy degree than additional prototypes - K-nearest neighbor (KNN), and Naive Bayes (NB). The research in [8] visibly exemplified that SVM classifier is the proficient model for water quality valuation in both Pre-Monsoon (PRM) and Post-Monsoon (PSM). The representation metrics of SVM refined accuracy, recall, precision, f1 score, and lesser error degrees commonly than additional classifiers. Paper [9] developed a water quality forecast model by the benefit of water quality facets via ANN and time-series analysis. The review work imposed the water quality of the year 2014 historic data, over a time break of six minutes. United States Geological Survey (USGS) online reserve entitled National Water Information System (NWIS) data was used, which grips the proportions of four dynamics that have an influence and control the water quality. For the insistence of gaging the representation of model, depiction estimates applied were Regression Analysis, MSE (Mean-Squared Error), and RMSE (Root Mean-Squared Error). Aforesaid mechanisms on Water Quality forecast devise too withstood, inspected likewise approaching intensifications have suggested surviving taking part in these investigation works.

The exploration [10] realized a chain of supervised ML practices to evaluate the WQI that is a precise indication to outline the universal quality of water, and the WQC (Water Quality Class) that is a unique set delineated on the core WQI. The endorsed system operated with four key-in dynamics, viz., total dissolved solids, temperature, pH, and turbidity. Between the intact dynamic practices, the WQI supreme proficiency was predicted attaining a MAE (Mean Absolute Error) of 1.9642 and 2.7273, gradient boosting with a learning rate of 0.1 and polynomial regression with a degree of 2, compatibly. But Multi-Layer Perceptron (MLP), labeled the WQC ultimate robustly, through an arrangement of (3, 7), as well as an accuracy of 0.8507 was attained. The recommended technique achieved true accuracy by a trivial number of dynamics to endorse the likelihood of the aforementioned practice in factual timeline water quality revealing techniques.

In the investigative work [11], cutting-edge AI practices were mechanized to forecast WQI and WQC. For the WQI forecast, ANN prototypes, viz. LSTM (Long Short-Term Memory) and NARNET (Nonlinear Autoregressive Neural Network) DL (Deep Learning) practice were used. Still, on behalf of the WQC venturing, the 3 ML practices, viz., K-NN, SVM, and NB, were applied. An abridged dataset had 7 vital dynamics, similarly the customary prototypes remained appraised put up continuously by definite arithmetic dynamics. These outcomes publicized that the foreseen prototypes will precisely forecast WQI, also label the quality version of water to its advanced robustness. Forecast fallouts publicized that the NARNET model attained fairly superior LSTM WQI values for its forecast, likewise the SVM WQC forecast practice achieved its supreme accuracy as (97.01%). Compatibly, the LSTM and NARNET prototypes put together reached identical accuracy in the regression coefficient for the appraising stage by a small variance (RLSTM = 94:21%, and RNARNET = 96:17%). Thereby the nature of proficient exploration will promote significantly to water handling.

3. Existing Systems

For the principal component regression practice 95% forecast accuracy and for the Gradient Boosting Classifier practice 100% labeling accuracy, exhibit consistent depiction linked by utilizing the ultra-modern prototypes for Gulshan Lake-related dataset.

Limitations:

1. The Gradient Boosting classifier labels were determined by the entire assessing data version to the quality level of water, where the extra prototypes miscategorized certain assessing data.
2. The mean WQI value was set up that exhibited the quality of water that is inapt for intake and agricultural watering in utmost regions.
3. At Gulshan Lake, the natural WQI value was strangely extraordinarily more owing towards the skewed values of TDS, COD, SS, EC, and alkalinity. So, this was fixed for ample precaution.
4. PCA + GB Regression learnt to stay a smaller quantity valued model.
5. More educational occasions built the model extra firm, and additional progression was credible on the forecast model. Those queries will be

conquered by apt PCR model fine-tuned also by DNN (Deep Neural Network).

4. Proposed System

Water quality forecast and categorization built on LDA by 95.68% accuracy and 100% labeling accuracy for the LGBM Classifier technique, reveals reliable interpretation linked by means of the modern prototypes for Gulshan Lake-related dataset.

Phases of the proposed LDA - LGBM model:

Phase 1: Data Preprocessing: Data quality appraisal, cleansing, amendment, and mitigation.

Phase 2: Data Pool and Dispensation: Among the available variables pH and DO are considered to calculate the WQI with their category stages 1 to 5: 1 being very poor, 2 in place of poor, 3 being fair, 4 being good and finally 5 as excellent.

Phase 3: Learning the existing and proposed techniques using Boosting-Centered Algorithms.

Phase 4: Building the ML model with the proposed LDA-LGBM simulation on or after performance appraisal of these Boosting-Centered – existing Gradient Boosting classifier model and proposed LDA-LGBM classifier model.

Phase 5: The proposed LDA-LGBM classifier model demonstrates the forecasting and categorization of water quality.

Benefits:

1. It is modest, rapid and a handy practice.
2. It imposes facts from the common features to create a unique axis that in turn cuts the variance and amplifies the class distance of the two variables.
3. By fine-tuning the LDA dynamics to fit disparate dataset outlines, the theme realization is determined and subsequently the document clusters.
4. Reduces the number of dynamics to an extra flexible extent in progress to labeling.

5. All new magnitudes begot a linear blend of pixel values that custom a prototype.

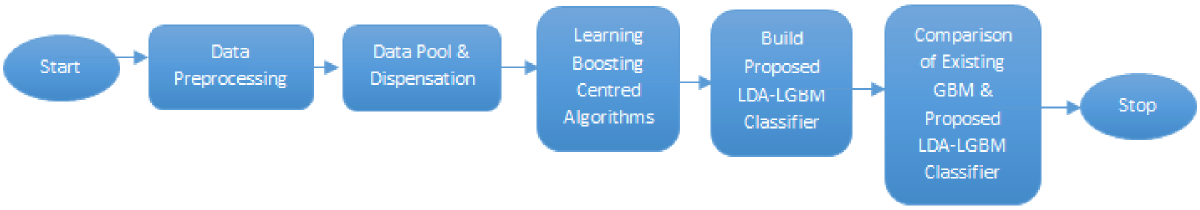
Start
Data Preprocessing
Data Pool & Dispensation
Learning Boosting Centred Algorithms
Build Proposed LDA-LGBM Classifier
Comparison of Existing GBM & Proposed LDA-LGBM Classifier
Stop

Figure 1. Phases of the proposed technique LDA - LGBM

5. Experimental Results

The brief specifics on the suggested architectures for forecasting and categorizing WQI in this LDA-LightGBM function, is indicated in Fig. 1. The LDA practice and LGBM model, are used in this operation.

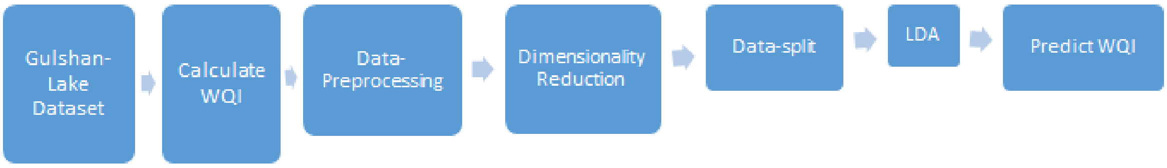

Figure 2a. Prediction of WQI by LDA Practice

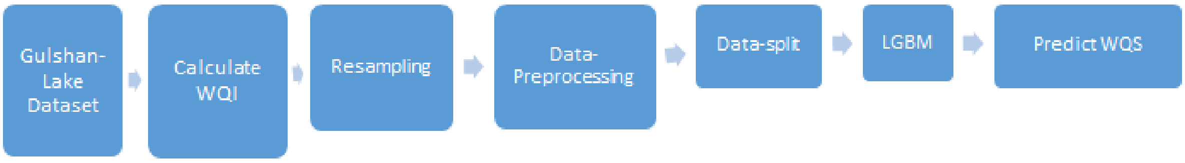

Figure 2b. Classification of WQS by LightGBM Practice

Calculation of WQI is shown in (1).

$$WQI = \sum w_j \, Q_j \Big/ \sum w_j \qquad (1)$$

where $w_j = 1/s_j$, w_j is the relative unit weight, s_j is customary value of the jth factors, similarly 1 is the persistent proportion.

$$Q_j = \left((M_j - l_j)/(S_j - l_j) \right) \times 100 \qquad 2)$$

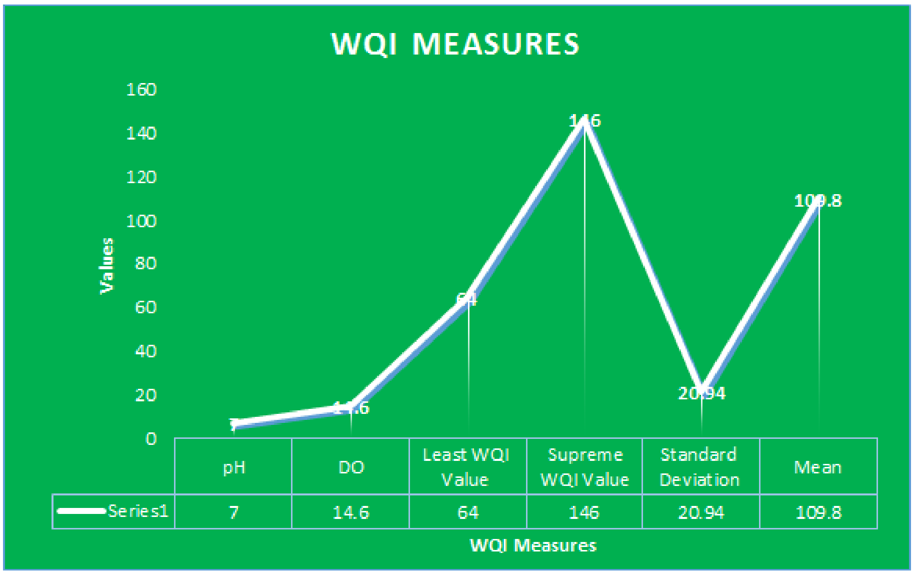

Figure 3. Line Chart of WQI Measures

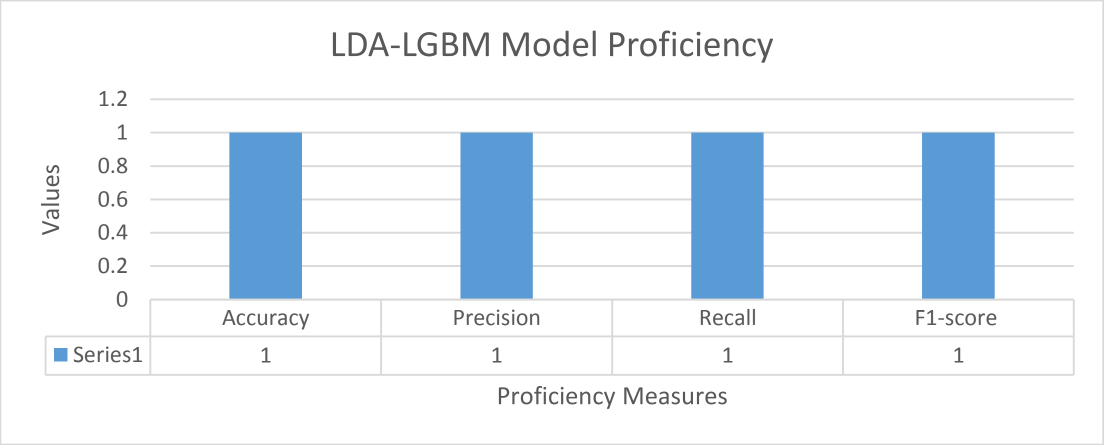

Figure 4. Bar Chart of LDA-LGBM Model Proficiency

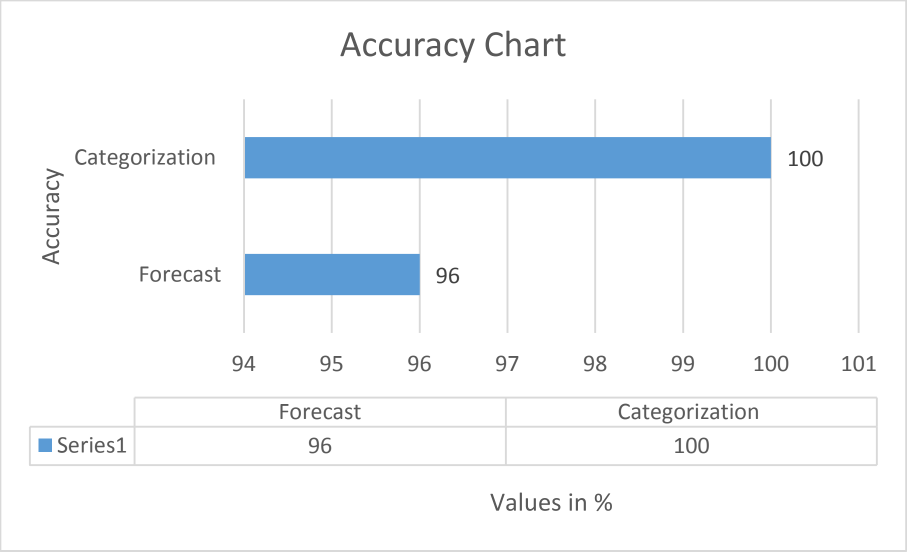

Figure 5. Bar Chart of Accuracy

Where Q_j is measured to be the value score of the jth measurement of water quality, $M_j$is denoted from the Gulshan lake as the appraised value, S_j is stated as the WHO advised accustomed rate of the water coefficient and l_j is conferred as the water quality dynamics perfect value.

The graph signifies the measures including the best value for DO and pH, are 14.6 mg/l and 7 respectively, and it is = 0 for other water quality. DO, pH, TDS, COD, SS, Turbidity Chloride, Alkalinity, EC and WQI are considered and with them the statistical measures such as, Min, Max, Mean, Count, Weight, Relative weight, and Standard Deviation are determined. WQI is determined by the above equations. After exhaustive learning, by a standard deviation of 20.94, the least WQI value is made real as 64, and with the supreme value as 146. The mean WQI value is obtained as 109.8 that exhibit that the water quality is unsuitable for intake and agricultural purpose in utmost regions.

Accuracy, Precision, Recall, F1-score are 1.0 R^2 value which indicates the first-rate LDA model proficiency. The LGB categorizer has attained the peak accuracy and similarly is a proficient model to forecast the quality of water. From [1], it is distinctive that the composed LDA and LightGBM techniques have outpaced the earlier reputable prototypes by achieving 96% forecast accuracy and 100% categorization accuracy.

6. Conclusion

The results indicate that while the accuracy of the integral dimensionality reduction practice is slightly progressive than that of the peripheral one, its training time will be doubled up. The dimensionality reduction techniques are diverse with various categories of data. The impact of dimensionality reduction techniques with LDA dimensionality reduction upon the ML prototypes entail to be furthermore learnt. So, this proposed LDA-LGBM analysis overcome these investigation breaks, and advances a new model LDA-LGBM: a pioneering LDA-LGBM model that reveals a boosting technique to amplify the representation for forecasting water quality, further stable and proficient. As LDA-LGBM is implemented alike PCR-GBM with a slight progress, estimating water quality signifies its forecast model employing the accessible practices. Subsequently, predicting the WQI practice used in the prior step obtains the output. Moreover, optimization to LDA-LGBM practice is performed for refining their performances, thus enhancing the water quality prediction and its classification by AI - ML practices.

References

[1] Md. Saikat Islam Khan, Nazrul Islam, Jia Uddin, Sifatul Islam, and Mostofa Kamal Nasir, "Water quality prediction and classification based on principal component regression and gradient boosting classifier approach", Journal of King Saud University – Computer and Information Sciences, Elsevier B.V., 2021.

[2] Amir Hamzeh Haghiabi, Ali Heidar Nasrolahi, and Abbas Parsaie, "Water quality prediction using machine learning methods", Water Quality Research Journal, vol. 53 no. 1,pp. 3–13, 2018.

[3] Alberto Fernández del Castillo, Carlos Yebra-Montes, Marycarmen Verduzco Garibay, José de Anda, Alejandro Garcia-Gonzalez, and Misael Sebastián Gradilla-Hernández, "Simple Prediction of an Ecosystem-Specific Water Quality Index and the Water Quality Classification of a Highly Polluted River through Supervised Machine Learning", Water 2022, vol. 14, no. 8, pp.1235, 2022.

[4] Ana Elizabeth Marín Celestino, José Alfredo Ramos Leal, Diego Armando Martínez Cruz, José Tuxpan Vargas, Josue De Lara Bashulto and Janete Morán Ramírez, "Identification of the Hydrogeochemical Processes and Assessment of Groundwater Quality, Using Multivariate Statistical Approaches and Water Quality Index in a Wastewater Irrigated Region", Water 2019, vol.11, pp.1702-1726, 2019.

[5] Mosleh Hmoud Al-Adhaileh and Fawaz Waselallah Alsaade, "Modelling and Prediction of Water Quality by Using Artificial Intelligence", Sustainability 2021, vol. 13, pp.4259, 2021.

[6] Luis Fernando Casillas-García, José de Anda, Carlos Yebra-Montes Harvey Shear, Diego Díaz-Vázquez, and Misael Sebastián Gradilla-Hernández, "Development of a specific water quality index for the protection of aquatic life of a highly polluted urban river", Ecological Indicators, vol. 129, October 2021, 107899, Elsevier, 2021.

[7] Aiswarya Vijayakumar, and A S Mahesh, "Quality Assessment of Ground Water on Small Dataset", International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, no. 5, pp. 475-478, March 2019.

[8] Aiswarya Vijayakumar, and A S Mahesh, "Quality Assessment of Ground Water in Pre and Post-Monsoon Using Various Classification Technique", International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 2, pp. 5996-6003, July 2019.

[9] Yafra Khan and Chai Soo See, "Predicting and analyzing water quality using Machine Learning: A comprehensive model", 2016 IEEE Long Island Systems, Applications and Technology Conference (LISAT), Farmingdale, NY, USA, 29-29 April 2016.

[10] Umair Ahmed, Rafia Mumtaz, Hirra Anwar, Asad A. Shah, Rabia Irfan and José García-Nieto, "Efficient Water Quality Prediction Using Supervised Machine Learning", Water 2019, vol.11, 2210, 2019.

[11] Theyazn H.H Aldhyani, Mohammed Al-Yaari, Hasan Alkahtani, and Mashael Maashi, "Water Quality Prediction Using Artificial Intelligence Algorithms", Applied Bionics and Biomechanics 2020, pp. 1-12, December 2020.