Turkish Journal of Engineering
https://dergipark.org.tr/en/pub/tuje
e-ISSN 2587-1366

Machine learning empowered prediction of geolocation using groundwater quality variables over YSR district of India

Jagadish Kumar Mogaraju *1

[1]*Lovely Professional University, Department of Geography, India, jagadishmogaraju@gmail.com*

Cite this study: Mogaraju, J. K. (2024). Machine learning empowered prediction of geolocation using groundwater quality variables over YSR district of India. Turkish Journal of Engineering, 8 (1), 31-45

Keywords
Machine learning
Haversine distance
Prediction
Extra trees regressor
Groundwater quality

Research Article
DOI: 10.31127/tuje.1223779

Received:24.12.2022
Revised: 01.04.2023
Accepted:12.04.2023
Published:15.09.2023

Abstract
Machine Learning (ML) has been used in the prediction of geolocation with improved accuracies in this work. The pre-processed data was subjected to prediction analytics using 22 machine learning algorithms over regression mode. It was observed that Extra Trees Regressor performed well with better accuracies in predicting latitude, longitude, and Haversine distance, respectively. Regression models like CatBoost, Extreme Gradient boosting, Light Gradient boosting machine, and Gradient boosting regressor were also tested. The R^2 values were computed for each case, and we obtained 0.96 (Longitude), 0.98 (Latitude), and 0.96 (Haversine), respectively. The evaluation of models was done using metrics like MAE, MASE, RMSE, R^2, RMSLE, and MAPE and R^2 is considered most important than others. The effect of data point was calculated using Cooks' distance, and the variable fluoride has a significant impact on the prediction accuracy of Longitude followed by RSC, Cl, SO4, SAR, NO3, NA, Ca, EC and pH variables. In the prediction of latitude, the SAR variable played a significant role, followed by Na and TH. According to the t-SNE manifold, three longitude values were quite different from the others. This work is supported by some of the manifests like Cooks' distance outlier detection, feature importance plot, t-SNE manifold, prediction error plot, residuals plot, RFECV plot, and validation curve. This work is done to report that the challenge of predicting both latitude and longitude on a common ground is solved partially, if not completely, and machine learning tools can be used for this purpose. Haversine distance can be obtained from latitude and longitude and can be used in the prediction of geolocation.

1. Introduction

Groundwater is an essential source that needs to be protected from external and internal pollutants and its overexploitation to achieve the goals set by the United Nations through SDGs [1-2]. Machine learning tools can be effectively used in the prediction of environmental factors that can disturb the purity and extent of groundwater across the world [3]. Most of the modeling techniques that are considered predictive ML algorithms are adaptable and can simulate nonlinear and complicated interactions within a small window of time [4]. We can observe different hydrogeological environments that can also alter the availability of groundwater [5]. The availability of data is limited, and this is the main problem that hinders experimentation and analysis [6]. The prediction of groundwater levels was possible using simulation methods, and this has helped in groundwater management effectively [7-9]. Several numerical models were used in the prediction of the quantity and quality of groundwater [10-11]. We can use Long Short-Term Memory (LSTM), Extreme Learning Machine (ELM), and Deep Learning (DL) methods for accurate and meaningful predictions [12-16]. The autoML frameworks can be used in handling the whole ML pipeline, starting from data input to the display of outputs in graphical modes and also solving other data-related problems [17]. AutoML tools were previously used in the investigation of drinking water quality [18]. Pycaret is one of the important libraries that can be used in the AutoML frameworks with appreciable results in the form of metrics and graphs [19]. Regression methods were used in the investigation of water quality to know correlation and other insights [20]. The latitude and longitude values can be merged in the form of Haversine distance, and it is being used for some of the location-based services [21]. Haversine distance is least affected by some of the features like the width of a valley, the height of a mountain, etc., and hence can be used in solving some of the navigation problems [22]. The interpolation techniques based on deterministic geostatistical techniques and artificial neural networks were used to create digital elevation models [23]. The extent of solar radiation was forecasted/predicted using machine learning approaches [24]. The water levels in the surface water body like lake was predicted using neurocomputing intelligence methods [25]. The empirical equations that are associated with climatic regions were calibrated using genetic algorithm, particle swarm optimization techniques along with multi-gene genetic programming method [26].

There is some gap in research that embraces the fact that variables of groundwater quality can be used in the prediction of geolocation with reasonable accuracy. Though the attempt to achieve the same is naïve at this point in time, considering data availability, this work might lead to higher enhancements in the future in solving navigation problems and improving location-based services.

2. Material and Method

2.1. Data

The datasets essential for this study were collected from the WRIS system of Government of India website [27] and Central Groundwater Control Board website [28].

The study area is shown in Figure 1. 1000 random points were selected for this study.

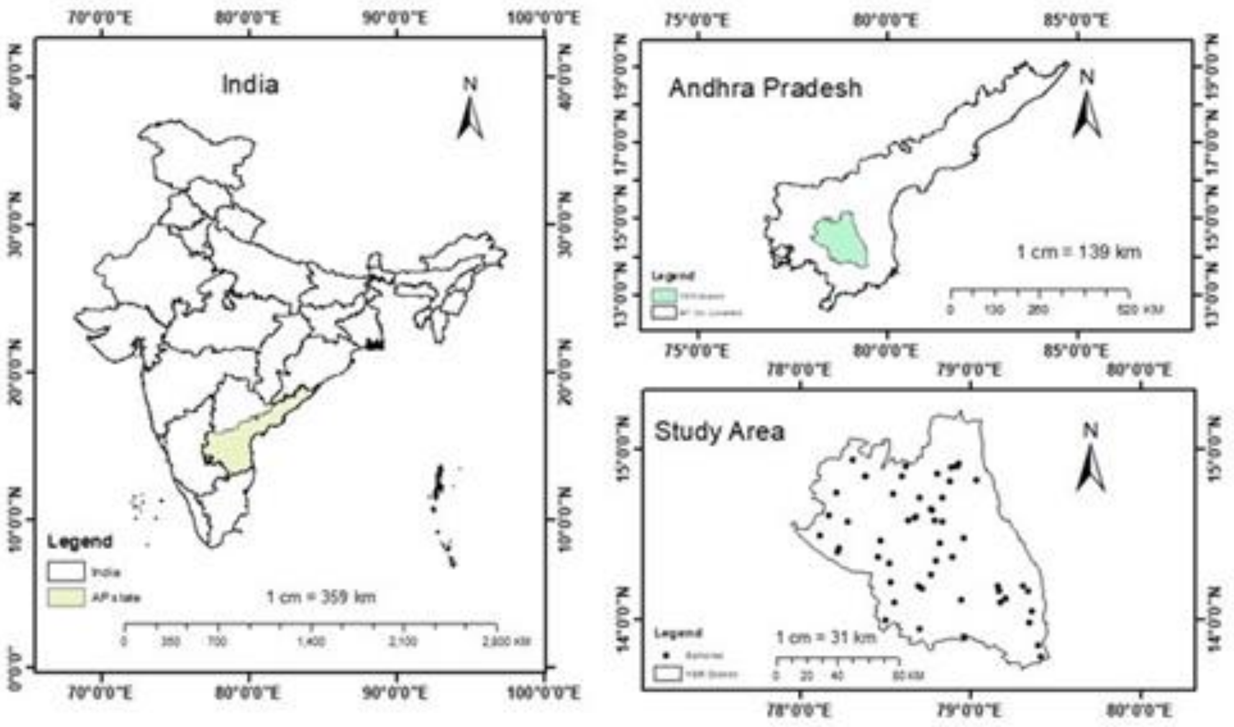

Figure 1. Study area.

2.2. Method

The models that were considered in this study are Extra Trees Regressor, CatBoost Regressor, Extreme Gradient Boosting, Light Gradient Boosting Machine, Random Forest, Gradient Boosting Regressor, Decision Tree, AdaBoost Regressor, k Neighbors Regressor, Ridge Regression, Linear Regression, Least Angle Regression, etc. Evaluation metrics like MAE, MSE, RMSE, R^2, RMSLE, and MAPE for each algorithm were considered. The datasets needed for this study were subjected to standard procedures that deal with data imbalance, missing values, and errors under the ML framework. The variables that were considered are HCO_3, Ca, Cl, F, K, Mg, Na, NO_3, pH, RSC, SAR, SO_4, TH (Total hardness), TA (Total alkalinity), and EC (Electrical conductivity). These variables were combined with latitude, longitude, and Haversine values separately, and they were considered as dependent or response variables, respectively. The combined datasets were passed onto the ML framework separately, and prediction accuracies with evaluation metrics were reported. The methodology employed in this study is given in Figure 2. The python packages built-in H_2O AutoML was used in this study. The information linked with the machine learning algorithms and H_2O AutoML package can be viewed through the online sources i.e., [29] (for models/algorithms) and [30] (for H_2O autoML).

The extra-trees regressor is a meta estimator that can fit randomized decision tress on sub-samples and averaging is done to enhance the prediction accuracy and regulates overfitting. CatBoost regressor relies on gradient boosted decision trees and a specific set of these trees will be built accordingly. Every tree that is built can be devoid of loss compared to the original ones. XGBoost or Extreme Gradient Boosting regressor can be scalable and it supplies parallel tree boosting. It can lower the error caused due to bias. Light Gradient Boosting Machine regressor uses traditional gradient boosting decision tree algorithms and also uses exclusive feature bundling (EFB) and Gradient-based One-Side sampling (GBDT) algorithms. More information on the algorithms/regressors can be obtained from the aforementioned links.

Central Haversine distance can be calculated between two points using Equation 1.

Where 'r' is the radius of earth, 'd' is the distance between two points, $ø_1$, $ø_2$ is the latitude of the two points and λ_1 and λ_2 are the longitudes of the two points.

$$\text{Haversine}\,(d/r) = \text{haversine}\,(ø_1 - ø_2) + \cos(ø_1)\cos(ø_2)\,\text{haversine}\,(\lambda_2 - \lambda_1) \quad (1)$$

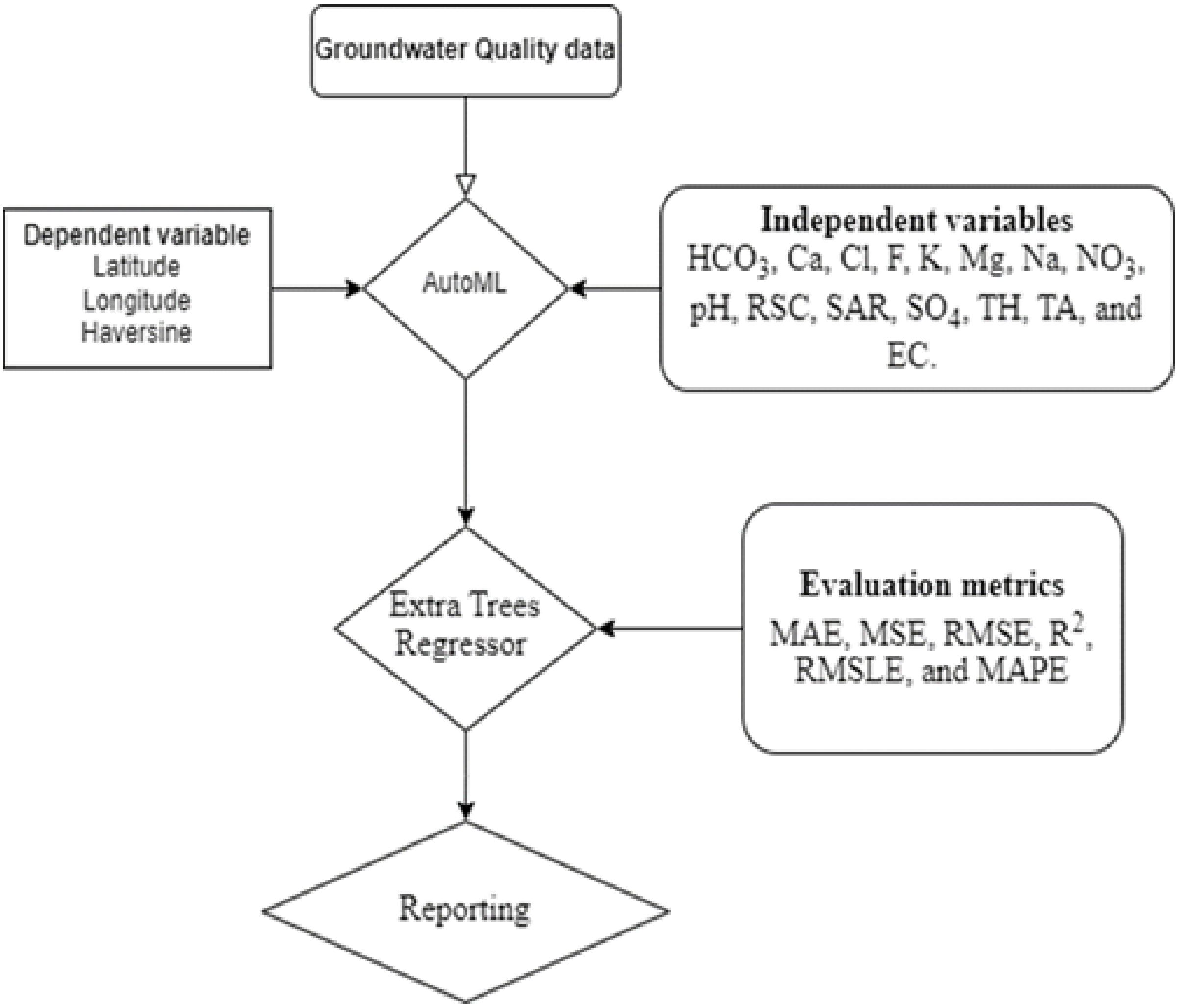

Figure 2. Methodology.

3. Results

3.1 Prediction of longitude

The Extra Tress Regressor (et) performed better than other regressors like CatBoost, Extreme Gradient boosting, Light Gradient boosting machine, Gradient boosting regressor, etc., predicting the Longitude variable. The 'et' regressor yielded an R^2 value of 0.96 with MAE (0.0391), MSE (0.004), RMSE (0.062), RMSLE (0.0008), and MAPE (0.0005), showing that the model fitted well and is shown in Figure 3 and model metrics for 'et' are given in Figure 4. The Cook's Distance Outlier Detection plot (6.98%) is given in Figure 5, and it represents the importance of an instance that might affect the prediction. The feature importance plot (Figure 6) shows that the variable 'F' has more influence than other variables like RSC, Cl, SO_4, SAR, NO_3, Na, Ca, EC and pH. According to the t-SNE Manifold plot, the longitude values 78.6, 78.4, and 78.2 are quite different from the others (Figure 7). The prediction error of the Extra Trees Regressor is shown in Figure 6, and we can observe an R^2 value of 0.965 with a best fit and identity (Figure 8). The residuals R^2 for training data is 1 and for test data, it is 0.965 (Figure 9). The Recursive feature elimination with cross-validation (RFECV) plot showed a score of 0.97 (Figure 10). The training score and the CV or cross-validation score are steadily increasing with each other, and it reflects that this model fitted well (Figure 11).

The plot given in Figure 3 shows the combined metrics of models used in this study. The dominance of R^2 metric is given in dark green.

3.2 Prediction of latitude

The Extra Tress Regressor (et) performed better than other regressors like CatBoost, Extreme Gradient boosting, Light Gradient boosting machine, Random

Forest, etc., predicting the Latitude variable. The 'et' regressor yielded an R^2 value of 0.98 with MAE (0.0261), MSE (0.0019), RMSE (0.0419), RMSLE (0.0027), and MAPE (0.0018), showing that the model fitted well and is shown in Figure 12 and model metrics are given in Figure 13. The Cook's Distance Outlier Detection (6.78%) is given in Figure 14. The feature importance plot shows that SAR highly influences the prediction, followed by Na, TH, Ca, Mg, SO_4, NO_3, EC, Cl, and RSC (Figure 15). The t-SNE Manifold plot shows that the latitude values 14.8, 114.6, and 14.4 differ from others (Figure 16). The prediction error plot shows that the R^2 value is 0.984 with best fit and identity (Figure 17). The predicted value versus Residuals is given in Figure 18 with Train R^2 of 1 and Test R^2 value of 0.984. RFECV plot scored 0.983 with 15 features (Figure 19). The validation curve showed that the training and cross-validation scores are growing gradually, reflecting that model performed well (Figure 20).

The plot given in Figure 12 shows the combined metrics of models used in this study. The dominance of R^2 metric is given in dark green.

3.3 Prediction of Haversine distance

The Extra Tress Regressor (et) performed better than other regressors like CatBoost, Extreme Gradient boosting, Light Gradient boosting machine, Random Forest, etc., predicting the Haversine distance variable. The 'et' regressor yielded an R^2 value of 0.96 with MAE (4.235), MSE (47.2008), RMSE (6.7658), RMSLE (0.0008), and MAPE (0.0005), showing that the model fitted well and is shown in Figure 21 and model metrics are given in Figure 22. The outlier detection plot (7.22%) is shown in Figure 23. The feature importance plot showed that the F variable strongly influences prediction, followed by RSC, SO_4, and Cl. SAR, Ca, NO_3, Na, pH, and EC (Figure 24). The t-SNE Manifold showed that the values 8780, 8760, and 8740 of Haversine distance are quite different from others (Figure 25). The prediction error plot shows that the R^2 value is 0.963 with the best fit (Figure 26). The residuals plot showed a Train R^2 and Test R^2 of 1 and 0.963, respectively (Figure 27). The RFECV plot showed a score of 0.969 (Figure 28). The validation curve showed that the cross-validation and training scores grew gradually, showing that the model fits well (Figure 29).

The plot given in Figure 21 shows the combined metrics of models used in this study. The dominance of R^2 metric along with other metrics is given.

The evaluation metrics of all models considered in this study is given in Table 1.

Table 1. Evaluation metrics.

Model	Latitude (R^2)	Longitude (R^2)	Haversine (R^2)
Extra Trees Regressor	0.9829	0.9697	0.9691
CatBoost Regressor	0.9814	0.9532	0.9537
Extreme Gradient Boosting	0.9691	0.9452	0.9427
Light Gradient Boosting Machine	0.9682	0.9322	0.9326
Gradient Boosting Regressor	0.9678	0.9072	0.9134
Random Forest	0.9646	0.9047	0.913
AdaBoost Regressor	0.9483	0.8322	0.8399
Decision Tree	0.9287	0.8279	0.8261
Linear Regression	0.9082	0.7606	0.7627
Ridge Regression	0.9059	0.7604	0.7624
Bayesian Ridge	0.9057	0.7579	0.76
TheilSen Regressor	0.9057	0.7306	0.7413
Random Sample Consensus	0.9049	0.7288	0.735
K Neighbors Regressor	0.9043	0.6856	0.7309
Elastic Net	0.8833	0.4237	0.6875
Lasso Regression	0.8725	0.3684	0.6766
Orthogonal Matching Pursuit	0.8467	0.3501	0.3658
Support Vector Machine	0.8112	0.3093	0.121
Lasso Least Angle Regression	0.6901	-0.0258	-0.001
Least Angle Regression	-0.0127	-4.799	-5.0102
Huber Regressor	-68.0315	-1740.73	-2006.42
Passive Aggressive Regressor	-554.398	-14405	-15444.1

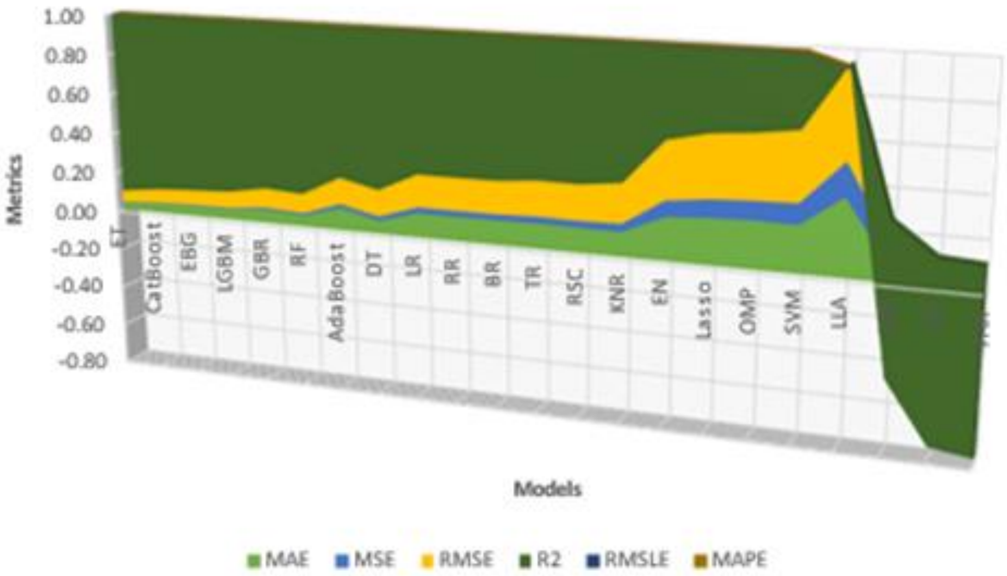

Figure 3. Model contrast (Longitude).

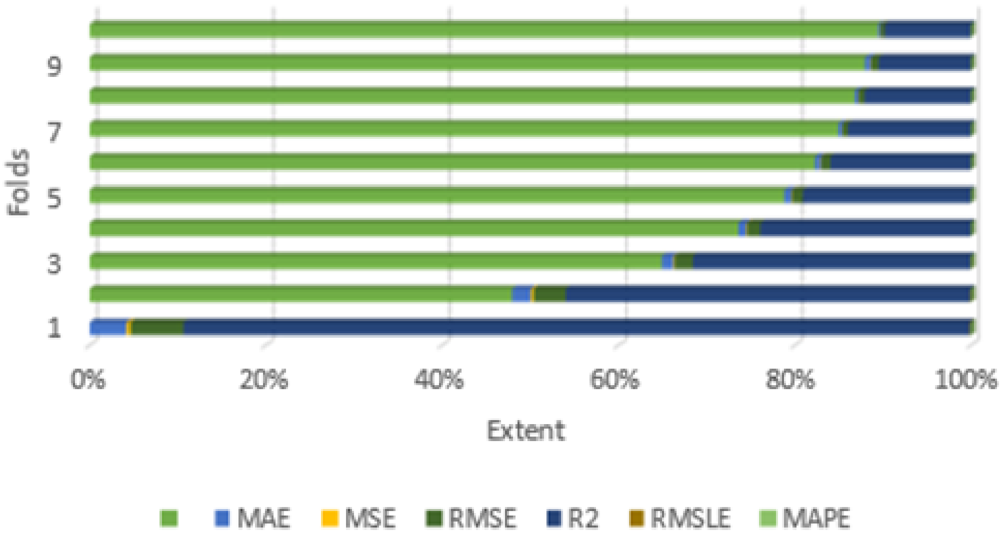

Figure 4. Model performance (Longitude).

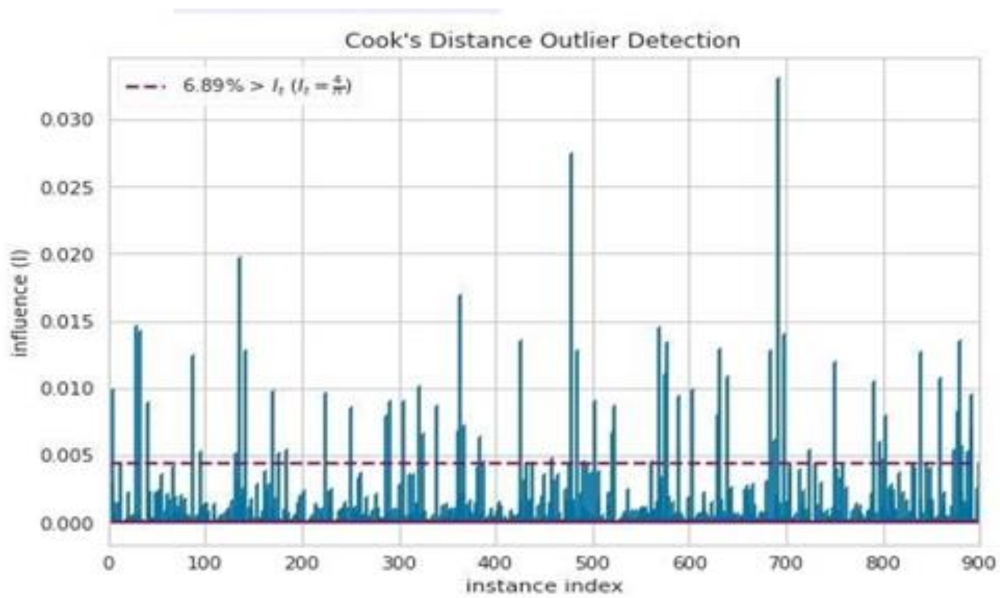

Figure 5. Outliers (Longitude).

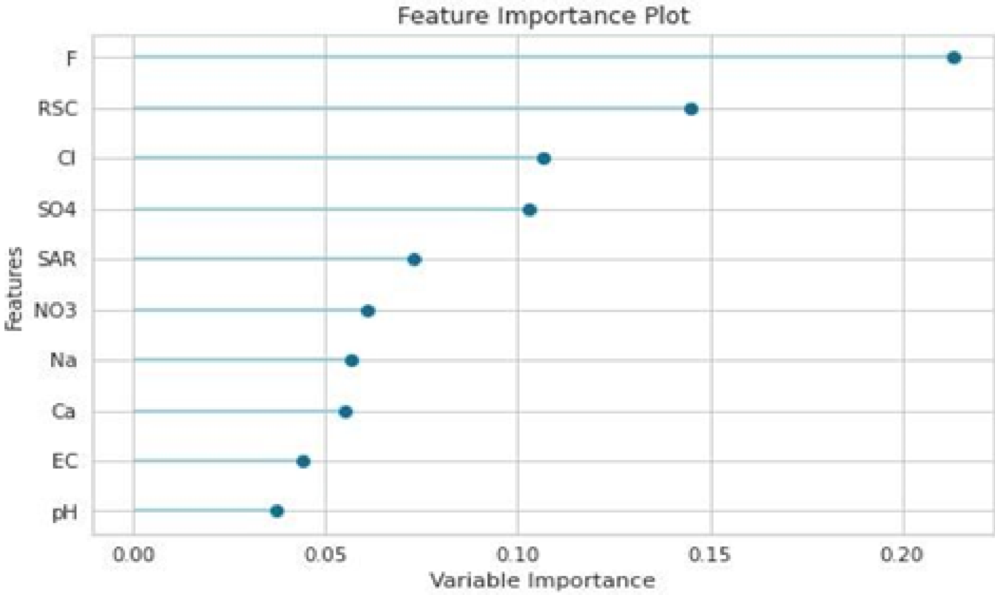

Figure 6. Feature importance (Longitude).

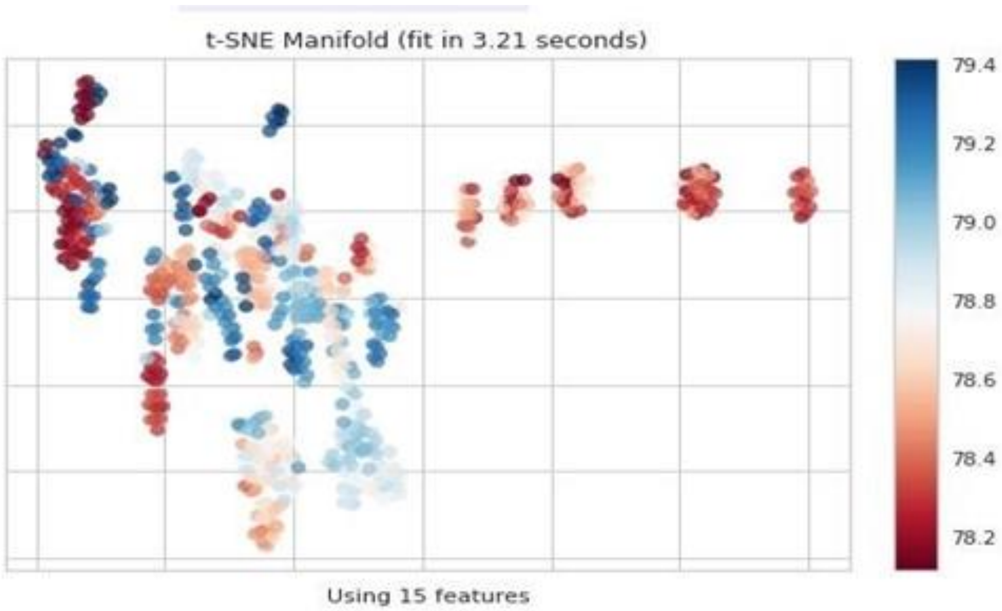

Figure 7. t-SNE Manifold (Latitude).

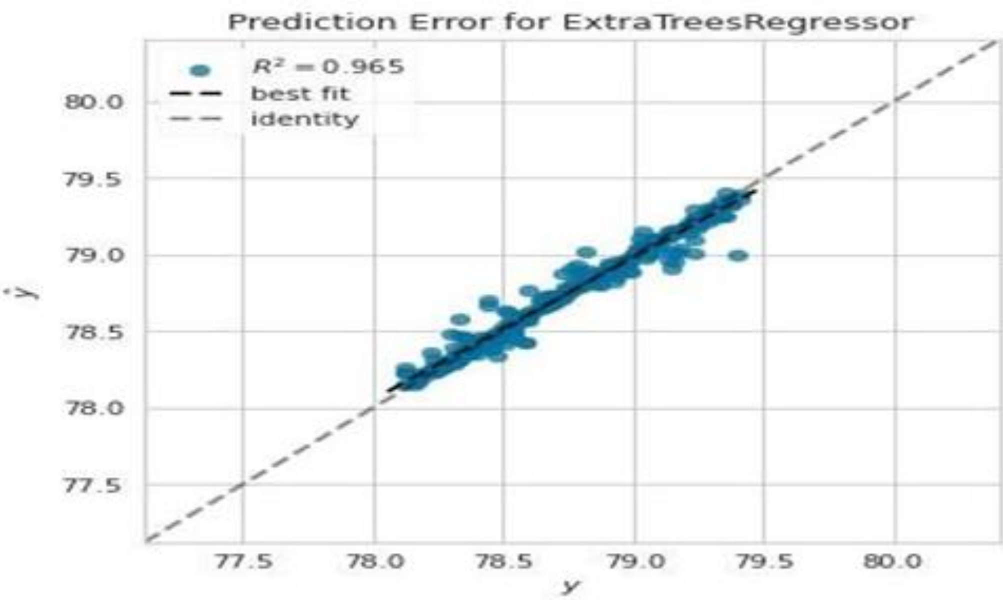

Figure 8. Prediction error (Longitude).

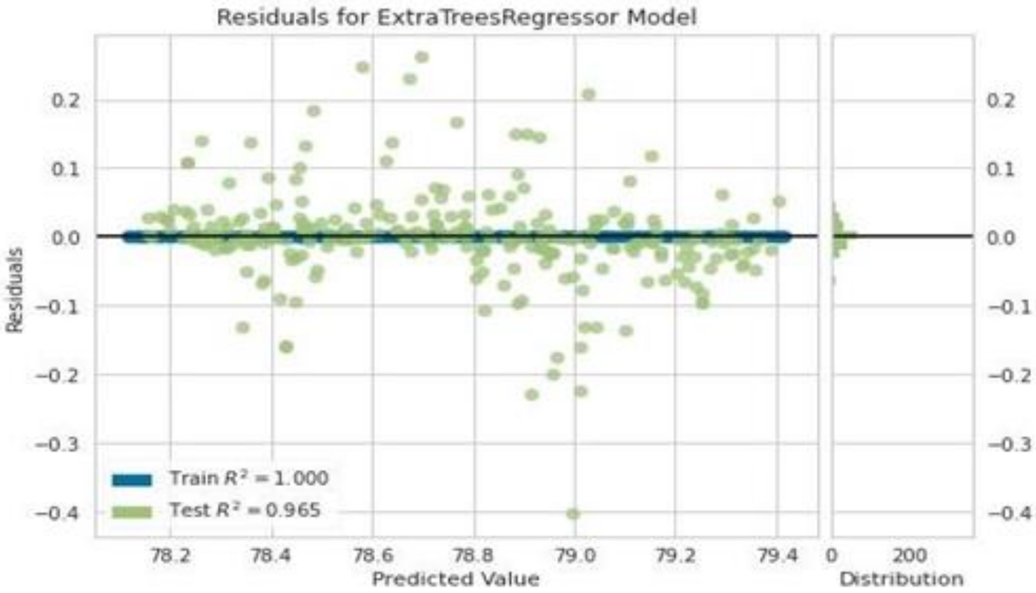

Figure 9. Residuals (Longitude).

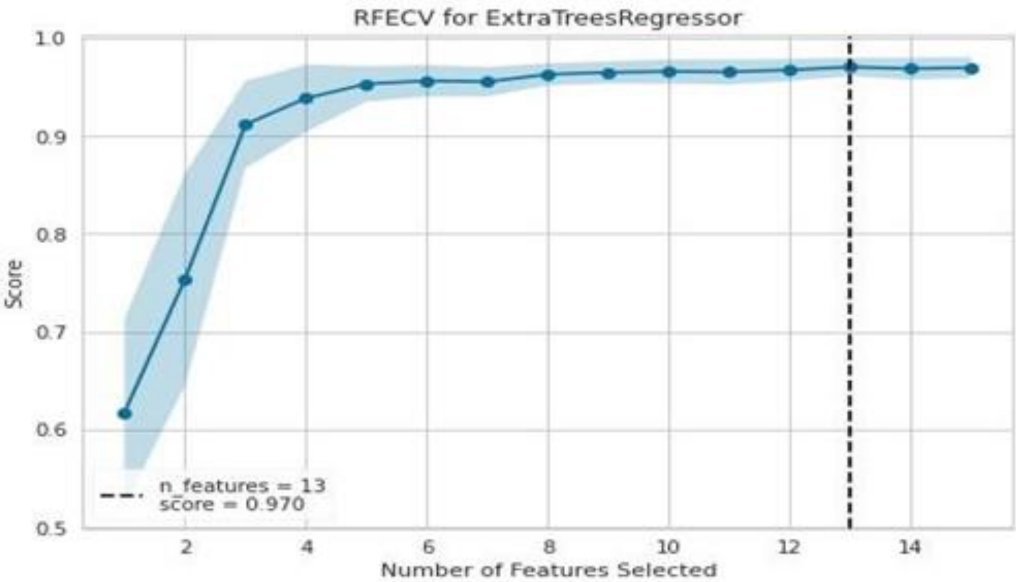

Figure 10. RFECV (Longitude).

Figure 11. Validation curve (Longitude).

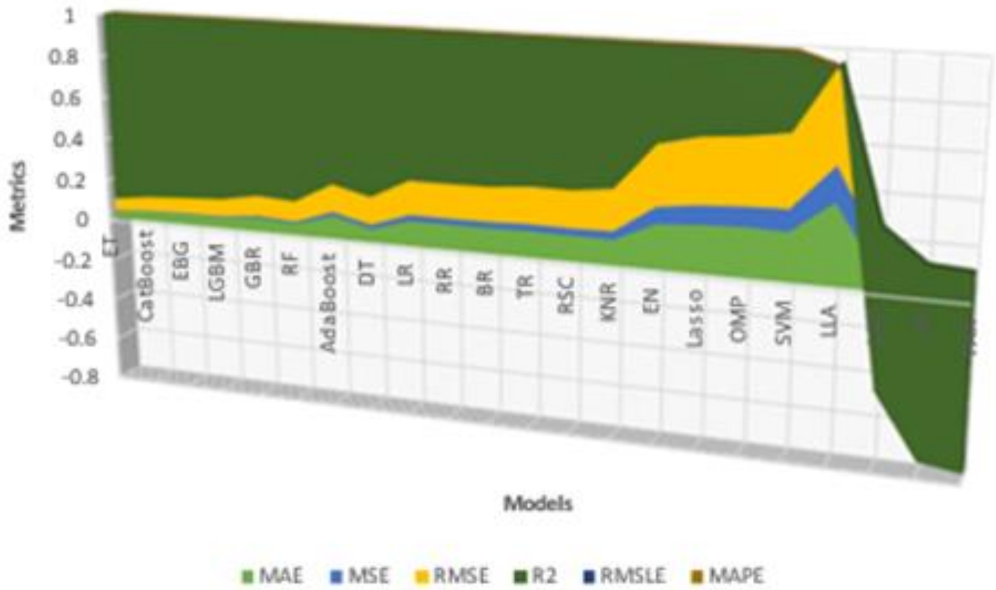

Figure 12. Model contrast (Latitude).

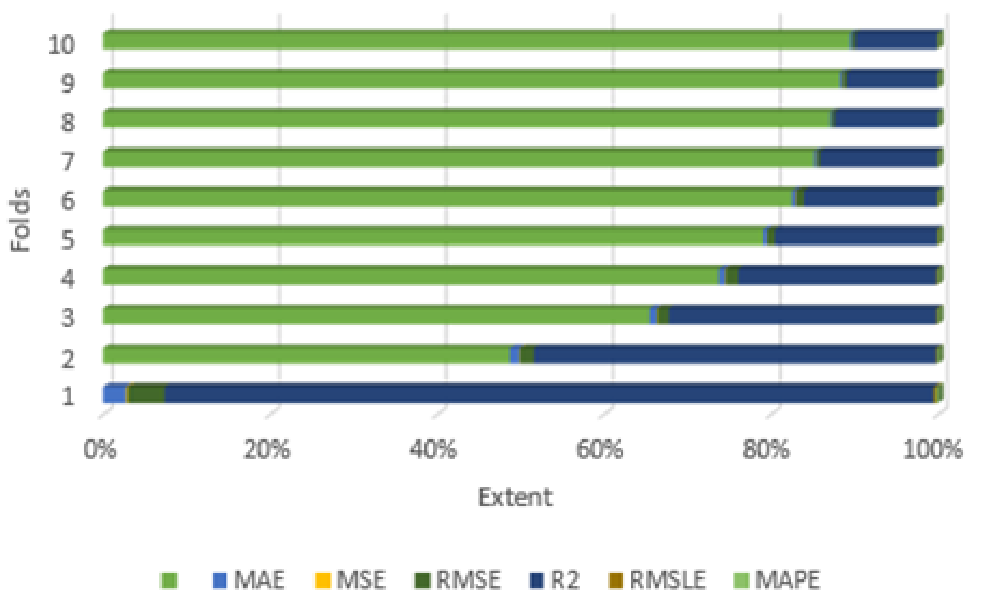

Figure 13. Model performance (Latitude).

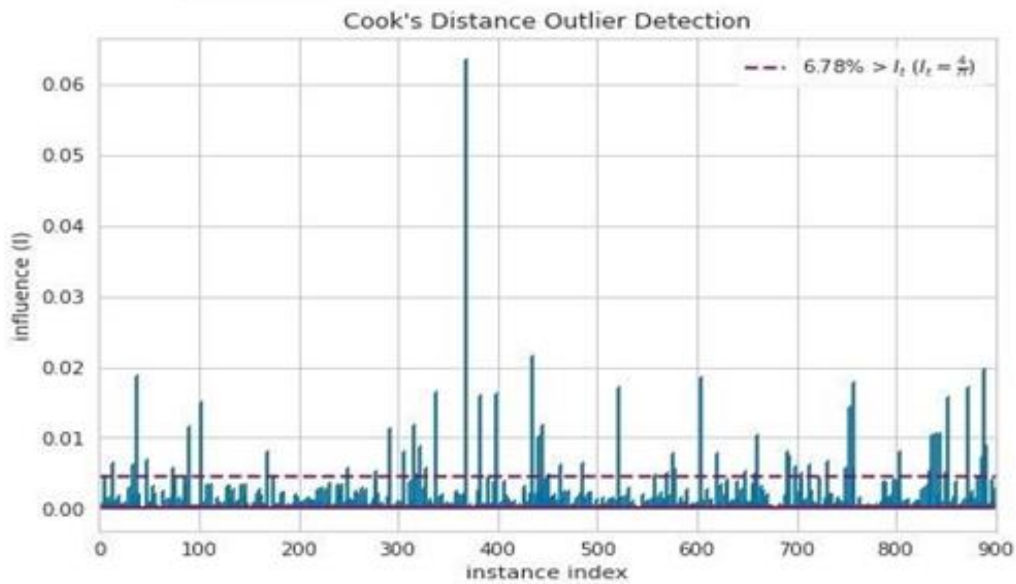

Figure 14. Outliers (Latitude).

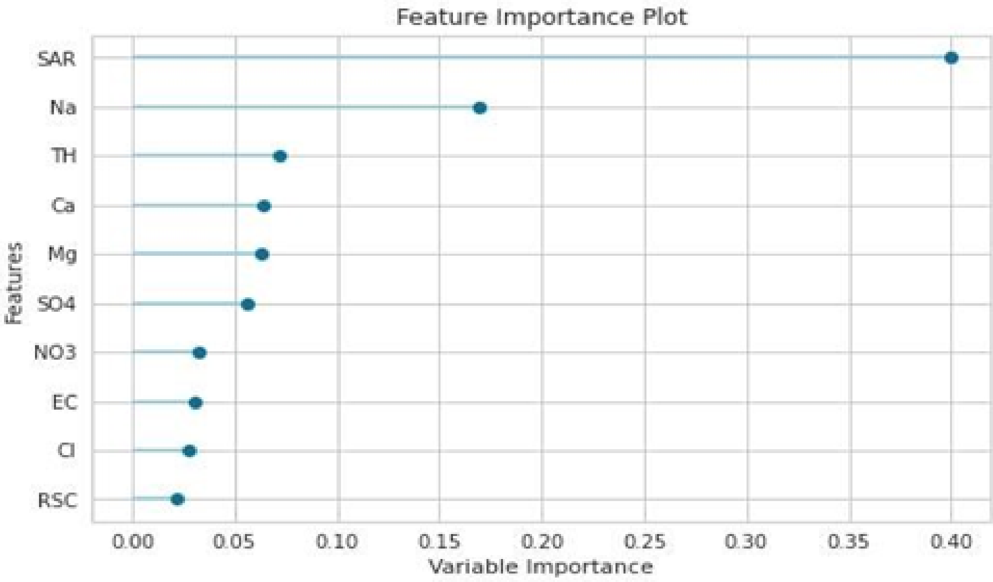

Figure 15. Feature importance (Latitude).

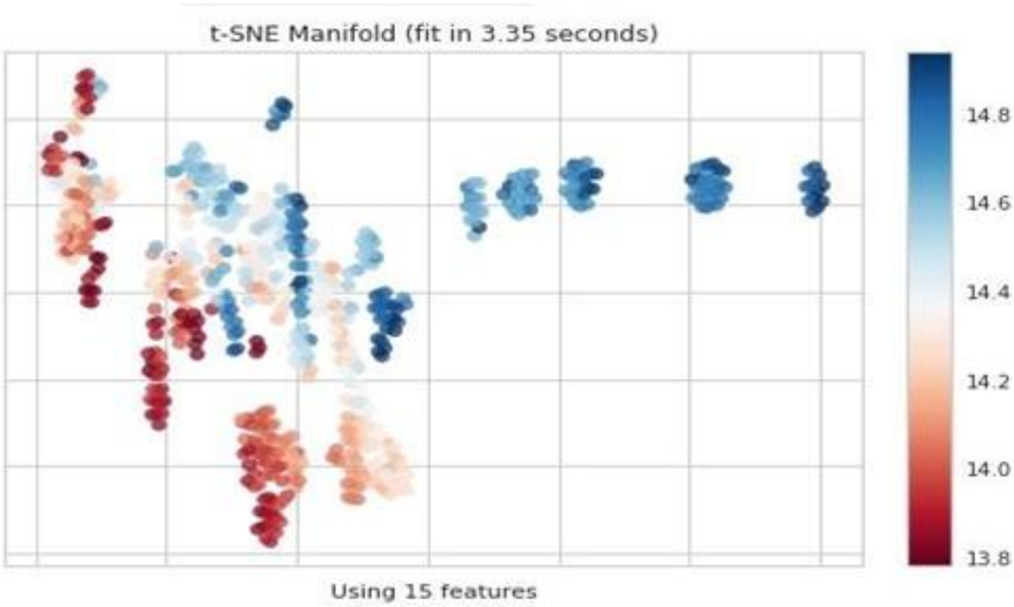

Figure 16. t-SNE Manifold (Latitude).

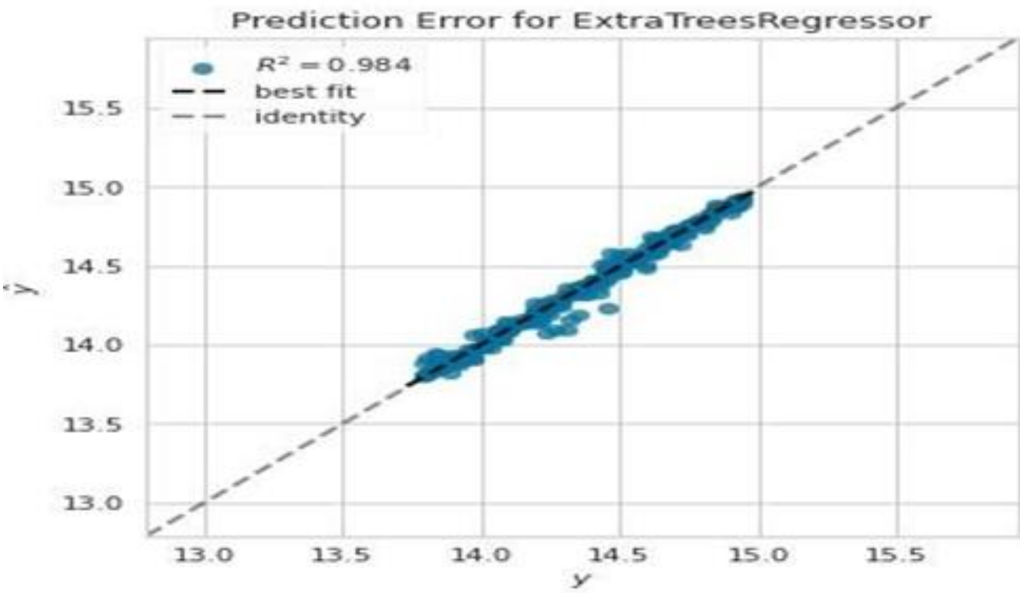

Figure 17. Prediction error (Latitude).

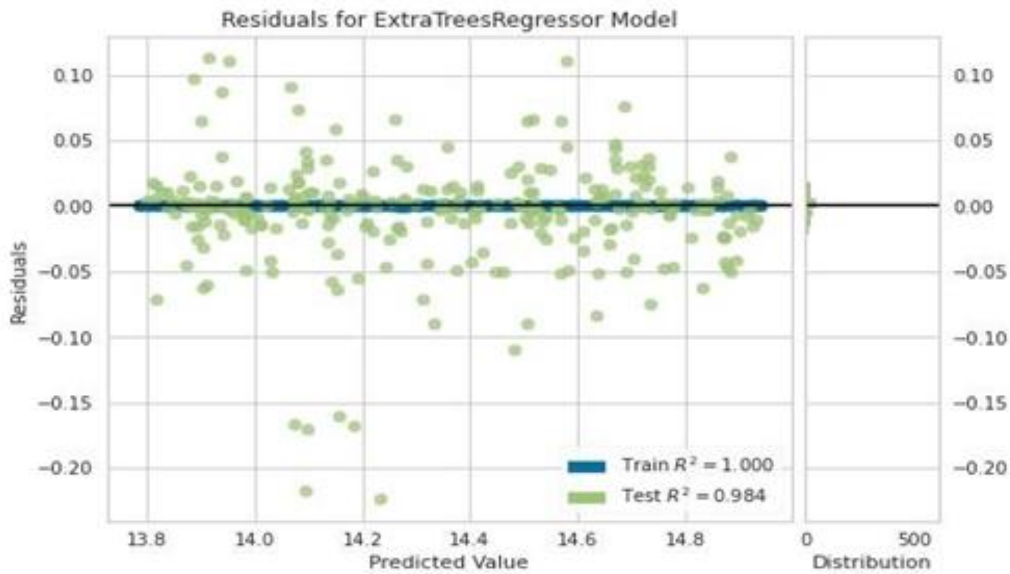

Figure 18. Residuals (Latitude).

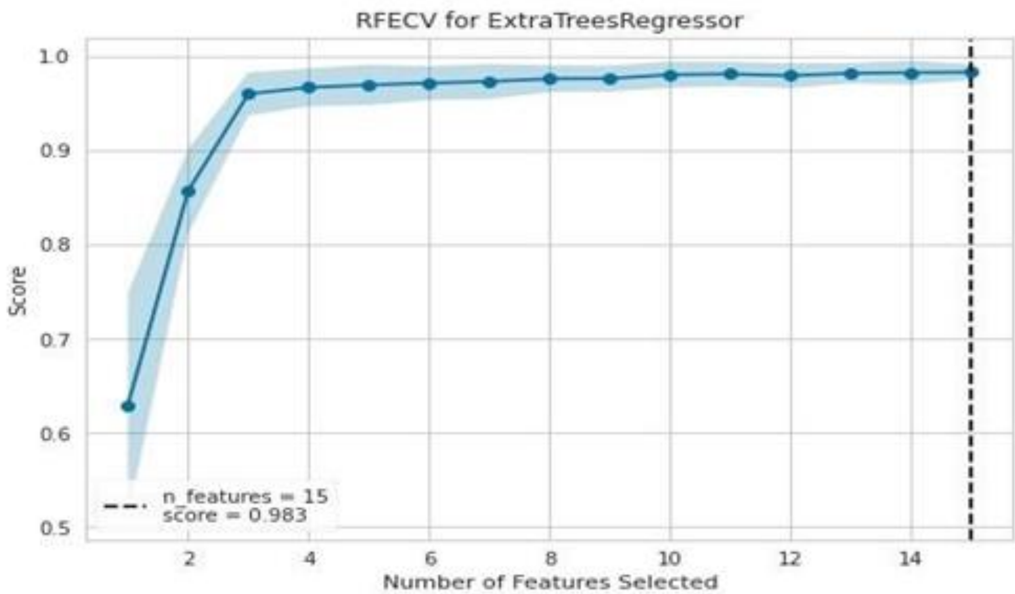

Figure 19. RFECV (Latitude).

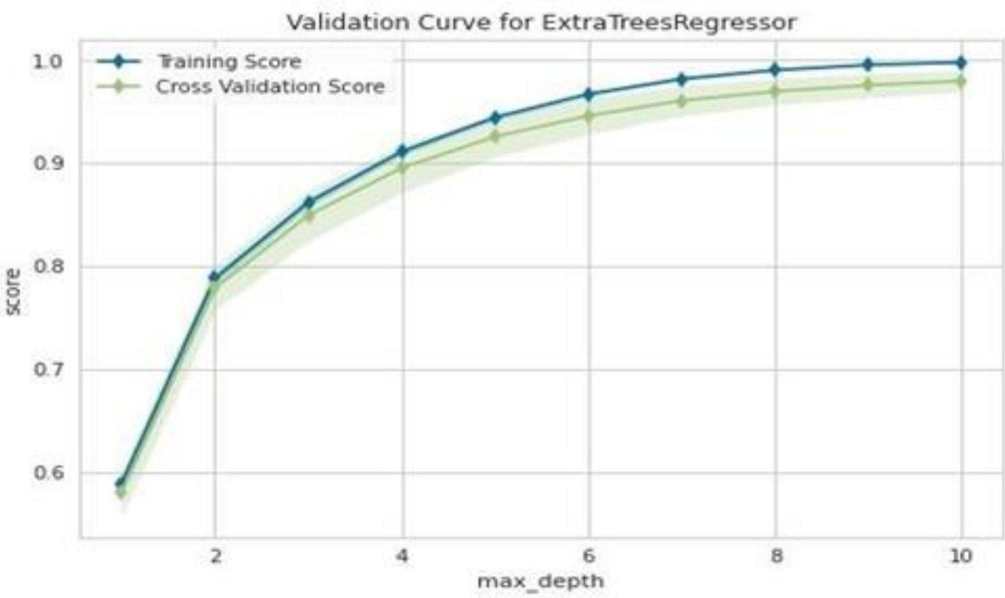

Figure 20. Validation curve (Latitude).

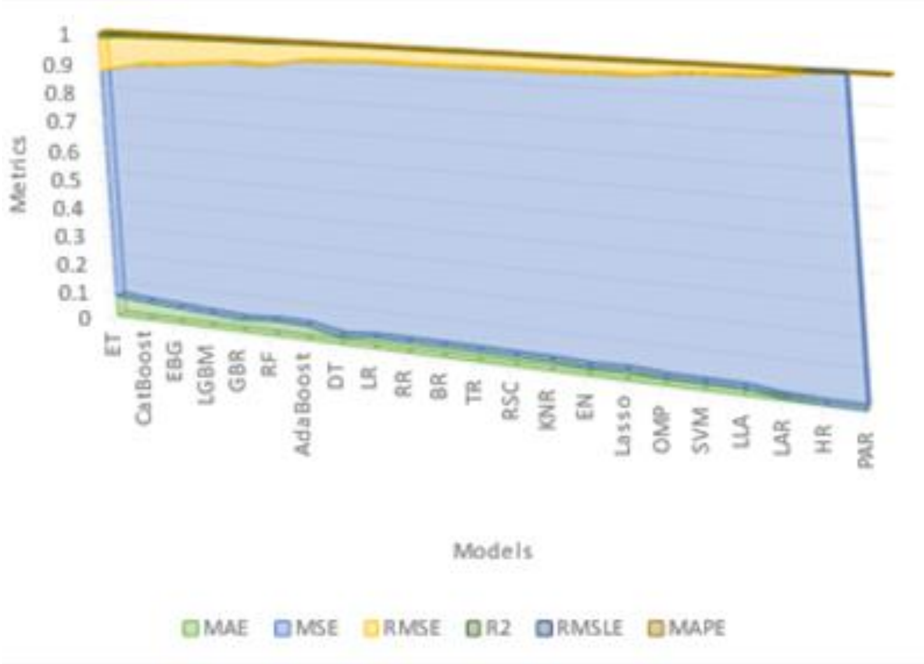

Figure 21. Model contrast (Haversine distance).

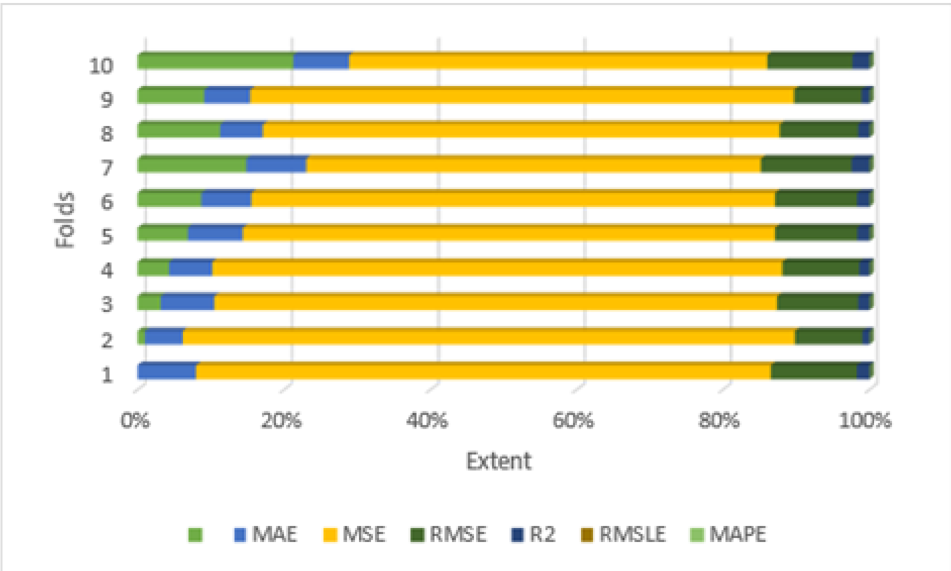

Figure 22. Model performance (Haversine distance).

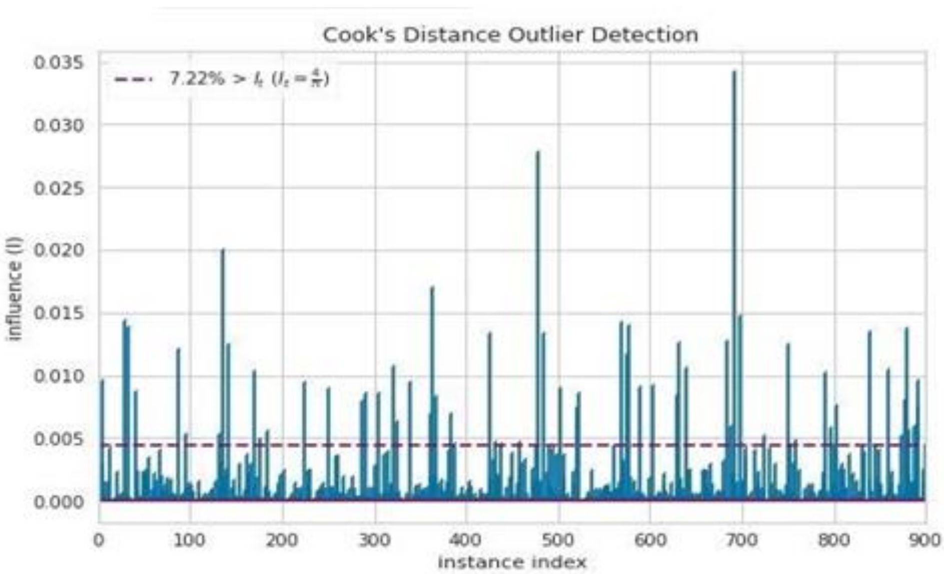

Figure 23. Outliers (Haversine distance).

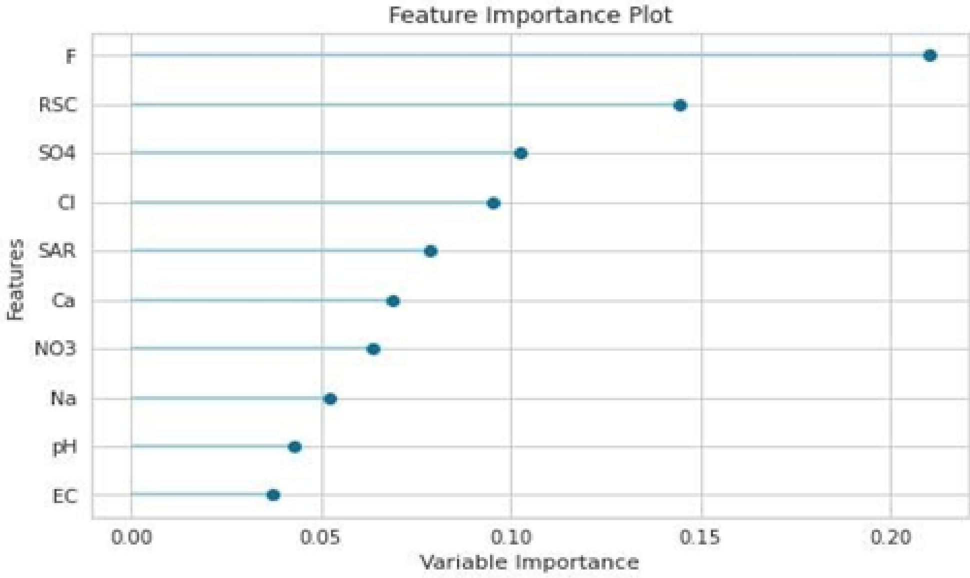

Figure 24. Feature importance (Haversine distance).

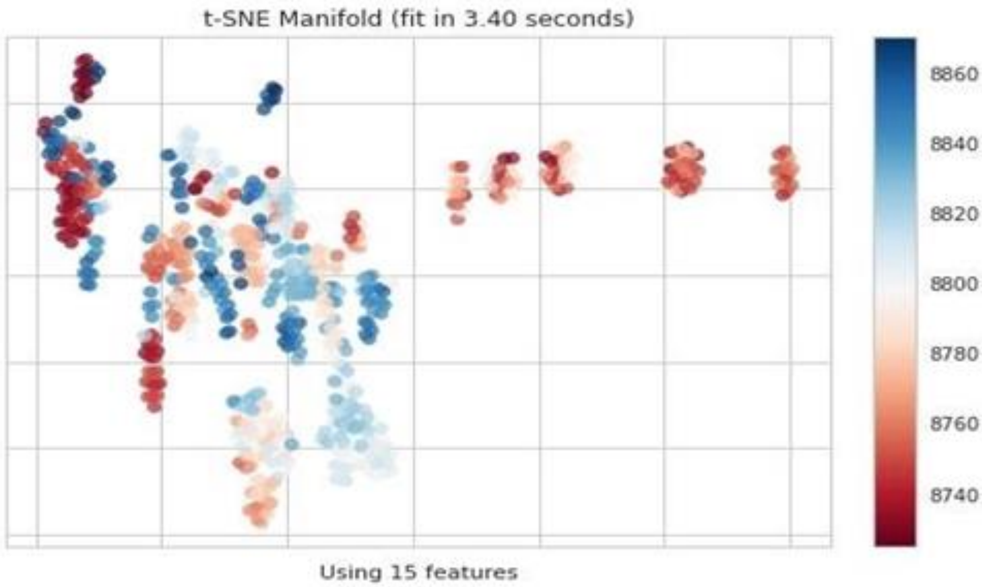

Figure 25. t-SNE Manifold (Haversine distance).

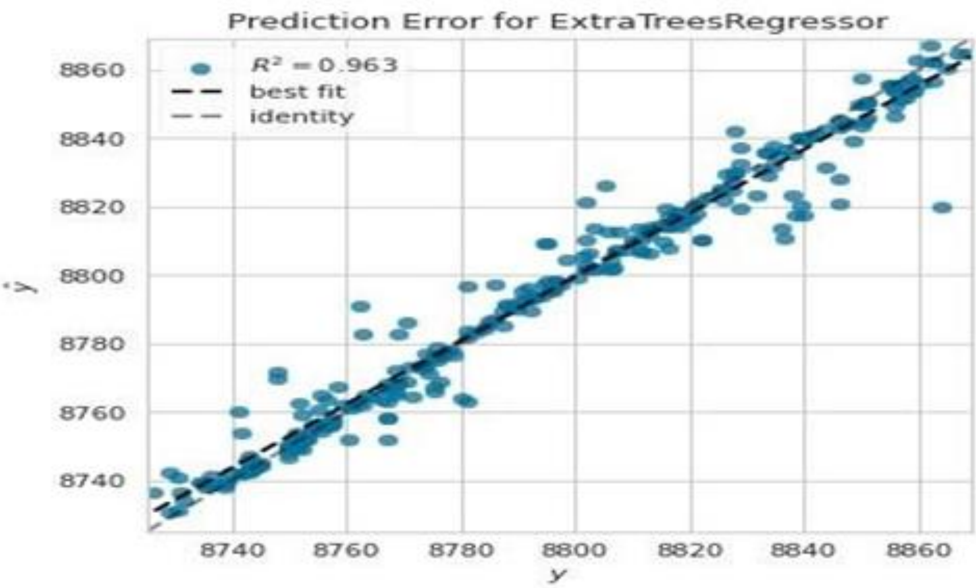

Figure 26. Prediction error (Haversine distance).

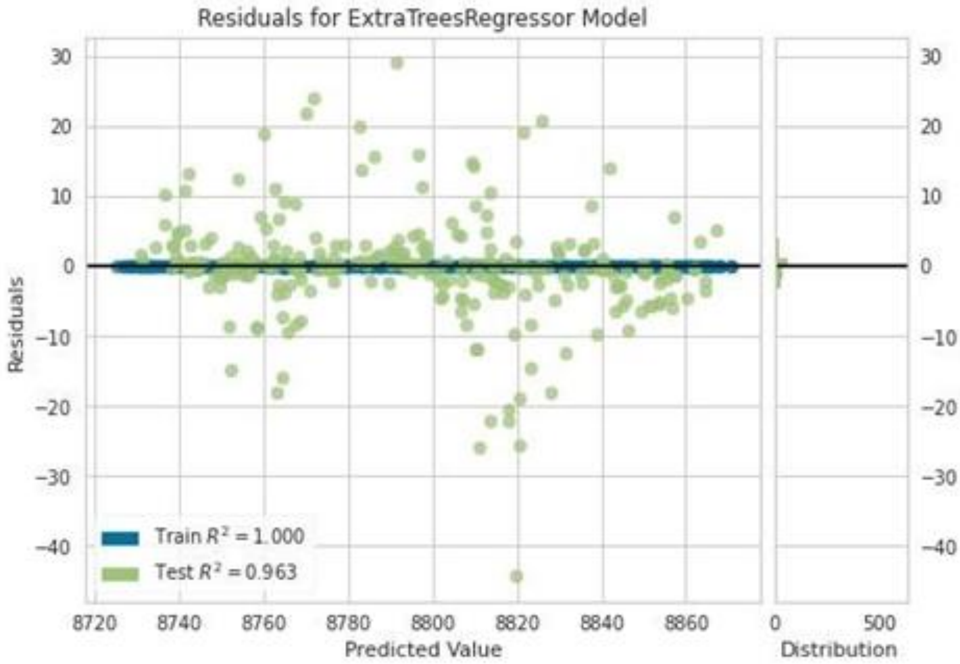

Figure 27. Residuals (Haversine distance).

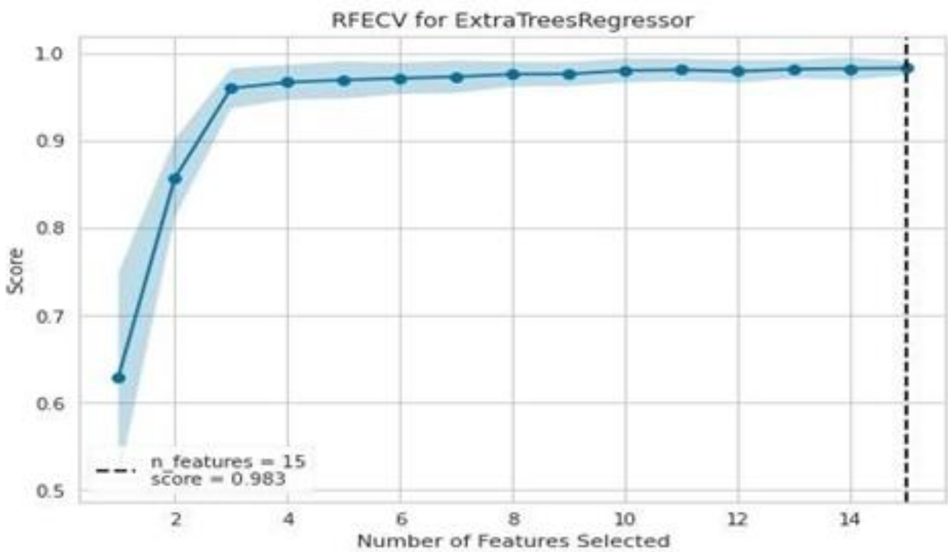

Figure 28. RFECV (Haversine distance).

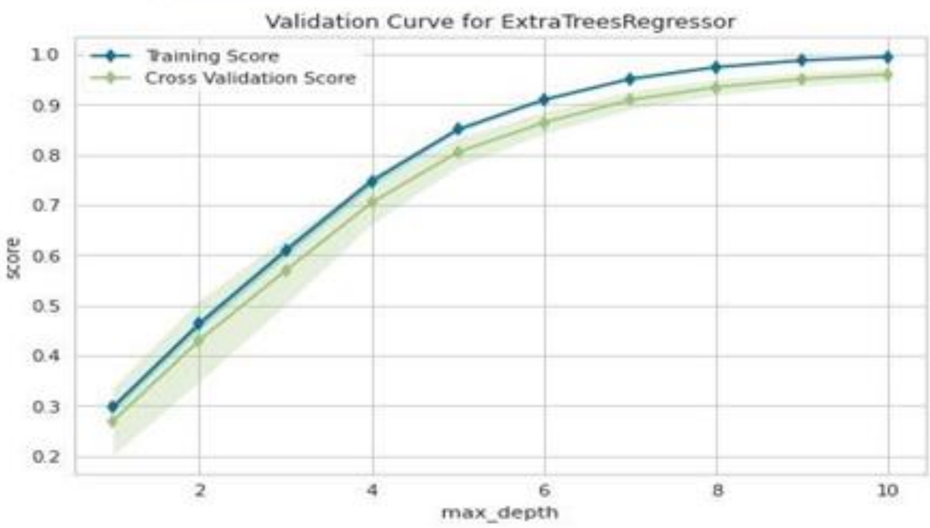

Figure 29. Validation curve (Haversine distance).

4. Discussion

The Extra Tress Regressor (et) performed better than other regressors i.e., CatBoost, Extreme Gradient boosting, Light Gradient boosting machine, Gradient boosting regressor, etc., predicting the Longitude variable. The 'et' regressor yielded an R^2 value of 0.96 reflecting that the model fitted well. Fluoride has more influence than other variables i.e., RSC, Cl, SO_4, SAR, NO_3, Na, Ca, EC and pH. The Extra Tress Regressor (et) responded well in predicting the Latitude variable. The 'et' regressor yielded an R^2 value of 0.98 showing that the model fitted well. The SAR highly influenced the prediction, followed by Na, TH, Ca, Mg, SO_4, NO_3, EC, Cl, and RSC. The Extra Tress Regressor (et) performed better than other regressors in predicting the Haversine distance. The 'et' regressor yielded an R^2 value of 0.96 reflecting that the model fitted well. Fluoride strongly influenced prediction, followed by RSC, SO_4, and Cl. SAR, Ca, NO_3, Na, pH, and EC. The present observations obtained through this study suggests that the groundwater quality variables if collected at large from several points will aid in solving some of the navigation problems when there is no network and if satellite relay systems are impaired due to solar storms. This work shows that the prediction accuracy of the geolocation can be effectively improved if more data related with groundwater quality is available. The works cited in the introduction section may be viewed to know the navigation challenges. Though artificial networks and deep learning frameworks were used in estimation of groundwater and watershed components, the location-based applications were less studied and this work will add some information in this research area. AutoML tools like H_2O can be used across various areas of scientific studies and cloud-based analytics can also aid us in lowering the cost of research.

5. Conclusion

This work aimed to predict latitude, longitude, and haversine distance, and it is concluded that the Extra Trees Regressor model performed better than other models in all the three cases. This work can be used in various applications and preferably larger datasets with location components. This research helps solve location-based problems if for any reason satellite-based navigation becomes impaired due to unpredictable natural disasters. The present work can be also used to address several issues that helps in both ways i.e., in understanding the role of groundwater quality variables in locational intelligence and vice versa. Through the evolution of artificial intelligence and cloud-based analytics, the cost of understanding the local hydrology has drastically reduced, however the field-based investigations are still consuming excess human labor and money. The researchers in the fields of hydrology, pollution, and GIS can get benefitted from this work.

Acknowledgement

The authors would like to thank Central Groundwater Control Board, Government of India for making data available for this research work.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Afzaal, H., Farooque, A. A., Abbas, F., Acharya, B., & Esau, T. (2019). Groundwater estimation from major physical hydrology components using artificial neural networks and deep learning. Water, 12(1), 5. https://doi.org/10.3390/w12010005
2. Azdy, R. A., & Darnis, F. (2020, April). Use of haversine formula in finding distance between temporary shelter and waste end processing sites. In Journal of Physics: Conference Series, 1500(1), 012104. https://doi.org/10.1088/1742-6596/1500/1/012104
3. Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient water quality prediction using supervised machine learning. Water, 11(11), 2210. https://doi.org/10.3390/w11112210
4. Alizamir, M., Kisi, O., & Zounemat-Kermani, M. (2018). Modelling long-term groundwater fluctuations by extreme learning machine using hydro-climatic data. Hydrological sciences journal, 63(1), 63-73. https://doi.org/10.1080/02626667.2017.1410891
5. Alkan, H., & Celebi, H. (2019). The Implementation of Positioning System with Trilateration of Haversine Distance. 2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), 1–6. Istanbul, Turkey: IEEE. https://doi.org/10.1109/PIMRC.2019.8904289
6. Bowes, B. D., Sadler, J. M., Morsy, M. M., Behl, M., & Goodall, J. L. (2019). Forecasting groundwater table in a flood prone coastal city with long short-term memory and recurrent neural networks. Water, 11(5), 1098. https://doi.org/10.3390/w11051098
7. Dwivedi, P., Khan, A. A., Mudge, S., & Sharma, G. (2022). Explainable AI (XAI) for Social Good: Leveraging AutoML to Assess and Analyze Vital Potable Water Quality Indicators. In Computational Intelligence in Data Mining: Proceedings of ICCIDM 2021 (pp. 591-606). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-16-9447-9_45
8. Gupta, P. K., Yadav, B., & Yadav, B. K. (2019). Assessment of LNAPL in subsurface under fluctuating groundwater table using 2D sand tank experiments. Journal of Environmental Engineering, 145(9), 04019048. https://doi.org/10.1061/(ASCE)EE.1943-7870.0001560
9. Izady, A., Davary, K., Alizadeh, A., Ziaei, A. N., Alipoor, A., Joodavi, A., & Brusseau, M. L. (2014). A framework toward developing a groundwater conceptual model. Arabian Journal of Geosciences, 7, 3611-3631. https://doi.org/10.1007/s12517-013-0971-9
10. Jamin, P., Cochand, M., Dagenais, S., Lemieux, J. M., Fortier, R., Molson, J., & Brouyère, S. (2020). Direct measurement of groundwater flux in aquifers within

the discontinuous permafrost zone: an application of the finite volume point dilution method near Umiujaq (Nunavik, Canada). Hydrogeology Journal, 28(3), 869-885. https://doi.org/10.1007/s10040-020-02108-y
11. Kim, G. B. (2020). A study on the establishment of groundwater protection area around a saline waterway by combining artificial neural network and GIS-based AHP. Environmental Earth Sciences, 79(5), 117. https://doi.org/10.1007/s12665-020-8862-3
12. Kuhn, M., & Johnson, K. (2013). Applied predictive modeling, 26, 13. New York: Springer.
13. Larsen, E., Noever, D., MacVittie, K., & Lilly, J. (2021). Overhead-MNIST: Machine Learning Baselines for Image Classification. https://doi.org/10.48550/ARXIV.2107.00436
14. Mallikarjuna, B., Sathish, K., Venkata Krishna, P., & Viswanathan, R. (2021). The effective SVM-based binary prediction of ground water table. Evolutionary Intelligence, 14(2), 779–787. https://doi.org/10.1007/s12065-020-00447-z
15. Meyer, H., Reudenbach, C., Wöllauer, S., & Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications–Moving from data reproduction to spatial prediction. Ecological Modelling, 411, 108815. https://doi.org/10.1016/j.ecolmodel.2019.108815
16. Moez, A. (2020). PyCaret: An open source, low-code machine learning library in Python. PyCaret, Apr.
17. Mukherjee, A., Duttagupta, S., Chattopadhyay, S., Bhanja, S. N., Bhattacharya, A., Chakraborty, S., ... & Sahu, S. (2019). Impact of sanitation and socio-economy on groundwater fecal pollution and human health towards achieving sustainable development goals across India from ground-observations and satellite-derived nightlight. Scientific Reports, 9(1), 15193.
18. Omar, P. J., Gaur, S., Dwivedi, S. B., & Dikshit, P. K. S. (2019). Groundwater modelling using an analytic element method and finite difference method: an insight into Lower Ganga River basin. Journal of Earth System Science, 128, 195. https://doi.org/10.1007/s12040-019-1225-3
19. Pant, R. R., Zhang, F., Rehman, F. U., Wang, G., Ye, M., Zeng, C., & Tang, H. (2018). Spatiotemporal variations of hydrogeochemistry and its controlling factors in the Gandaki River Basin, Central Himalaya Nepal. Science of the Total Environment, 622, 770-782. https://doi.org/10.1016/j.scitotenv.2017.12.063
20. Pham, B. T., Jaafari, A., Prakash, I., Singh, S. K., Quoc, N. K., & Bui, D. T. (2019). Hybrid computational intelligence models for groundwater potential mapping. Catena, 182, 104101. https://doi.org/10.1016/j.catena.2019.104101
21. Xin, L., & Mou, T. (2022). Research on the Application of Multimodal-Based Machine Learning Algorithms to Water Quality Classification. Wireless Communications and Mobile Computing, 2022, 9555790. https://doi.org/10.1155/2022/9555790
22. Xue, J., Huo, Z., Wang, F., Kang, S., & Huang, G. (2018). Untangling the effects of shallow groundwater and deficit irrigation on irrigation water productivity in arid region: New conceptual model. Science of The Total Environment, 619, 1170-1182. https://doi.org/10.1016/j.scitotenv.2017.11.145
23. Çubukçu, E. A., Demir, V., & Sevimli, M. F. (2022). Digital elevation modeling using artificial neural networks, deterministic and geostatistical interpolation methods. Turkish Journal of Engineering, 6(3), 199-205. https://doi.org/10.31127/tuje.889570
24. Demir, V., & Citakoglu, H. (2023). Forecasting of solar radiation using different machine learning approaches. Neural Computing and Applications, 35(1), 887-906. https://doi.org/10.1007/s00521-022-07841-x
25. Demir, V., & Yaseen, Z. M. (2023). Neurocomputing intelligence models for lakes water level forecasting: a comprehensive review. Neural Computing and Applications, 35(1), 303-343. https://doi.org/10.1007/s00521-022-07699-z
26. Citakoglu, H., & Demir, V. (2023). Developing numerical equality to regional intensity–duration–frequency curves using evolutionary algorithms and multi-gene genetic programming. Acta Geophysica, 71(1), 469-488. https://doi.org/10.1007/s00521-022-07699-z
27. https://indiawris.gov.in/wris/
28. http://cgwb.gov.in/GW-data-access.html
29. https://scikit-learn.org/stable/supervised_learning.html#supervised-learning
30. https://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html