**Comparative Analysis of Machine Learning Algorithms Efficacy in Water Resources Management**

**Abstract**

Recently, many water quality prediction models have been developed so as to improve the water resources management system. Amongst them, the models based on the application of artificial intelligence are prevalent. In this study, a comparative analysis of the efficiency of different machine learning models in Dissolved Oxygen prediction was carried out. The General Linear Model, the Decision Tree, the Deep Learning (DL), the Random Forest, the Gradient Boosted Tree and the Support Vector Machine (SVM) were analyzed. The accuracy degree was being determined based on the root mean squared error, absolute error, relative error, mean squared error, correlation and squared correlation values. The values of the largest number of the indicators pointed to the fact that the SVM method was the most reliable in DO prediction, and the same was also characterized by the longest total scoring time. According to the performance indicators, the DL method only slightly lags behind the SVM method, simultaneously gaining in significant advantage in the form of the shortest time needed for that method to create a model.

Key words: Water resources, Machine Learning algorithms, Support Vector Machine, Deep learning

**Introduction**

In recent decades, significant efforts made towards the renewal and protection of bodies of water have been recorded at a global level. In order to prevent accidents, it is necessary to recognize and identify all the hydrological changes that refer to potential negative consequences in time. The best way to do that is to develop and improve water quality prediction models as an important step towards establishing reliable water resources management. These models can timely warn about the urgency of undertaking adequate measures. Water quality modelling is an inevitable activity in water resources engineering. Supportive of that is the fact that increasingly greater attention is being paid to this field by environmental scientists and engineers.

Contemporary water quality management models are most frequently based on artificial intelligence methods, i.e. the application of machine learning techniques. Machine learning models and data mining methods are numerical tools whose advantage reflects in the potential to process huge datasets. Machine learning algorithms are particularly useful in the cases of complex ecosystems, in which traditional techniques cannot be successfully applied (Valerio et al. 2021). Machine learning methods offer a precise prediction without any need for complex programming, whereas their flexibility enables the development of better and more effective models, which serve to overcome difficulties in water quality parameter monitoring (Ho et al. 2019). They enable finding patterns to be further used to predict future outcomes based on historical data about water quality by applying new datasets. Limitations of machine learning algorithms reflect in the fact that creating an optimal model is a more complicated procedure, whereas an inadequate choice of variables may endanger the model performances (Granata et al. 2017).

The motivation for this study reflects in a wish to define a method based on machine learning algorithms, which will enable the most efficient water quality prediction in real time and reduce the number of input variables as much as possible, simultaneously not decreasing the reliability of such prediction. The research goals are to i) determine the significance of each selected water quality parameter in the prediction of dissolved oxygen (DO); ii) make a comparison of the efficiency and accuracy level of different supervised machine learning algorithms for DO prediction; iii) define the minimum number of input variables that leads to the most reliable prediction of DO values for each method individually.

**Literature Review**

So far, artificial intelligence methods have intensively been studied and generally accepted as an alternative in modeling complex non-linear systems, such as water systems (Najah et al. 2012; Xu and Liu 2013; Parsaie and Haghiabi 2015). When solving a problem in the hydrology field, non-linearity very frequently occurs. Some shortcomings of the ANNs are also over-fitting, convergence to a local minimum and slow learning. Therefore, apart from ANN variations, different methods based on machine learning algorithms such as tree-based and vector-based models, whose successfulness in prediction was different, were examined. Numerous studies have recognized the Support Vector Machine as a reliable modeling and prediction method in the field of water resources management (Xiaoliang et al. 2017; Alizadeh et al., 2020). The advantages of the SVM in relation to the ANN reflect in the unique and optimal architecture and the ability to provide a quick training process, simultaneously eliminating subjectivity (Hipni et al. 2013). The authors Olyaie et al. (2017) applied two versions of neural networks for the prediction of the same water quality parameter – Multilinear Perceptron (MLP) and RBFNN; Linear Genetic Programming (LGP) and the SVM. According to the results, the most reliable prediction could be made by applying the SVM model.

Studying the efficiency of the prediction of surface water quality parameters by applying different modern neurocomputing models is increasingly more present. On the example of Laguna Lake among machine learning algorithms such as Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), Gradient Boost (GB) and Deep Learning (DL), the DT achieved the highest accuracy and precision (Lerios and Villarica 2019). In comparison with the classical approaches such as the MLP and Linear Regression (LR), the DL method showed to be more reliable in the prediction of water quality parameters, such as pH, dissolved oxygen, turbidity and chloride (Solanki et al. 2015). The authors Chen et al. (2020) carried out a comparison of the performances of the seven traditional and the three ensemble learning models for the prediction of water quality by applying a large dataset generated by monitoring the quality of the waters of the main river flows and lakes in China. Their study proved that the DT, the RF and the Deep Cascade Forest (DCF) were the most efficient (Chen et al., 2020).

In spite of the proven potentials of the machine learning approach in the field of ecological management, there are still an insufficient number of research studies and the insufficient application of machine learning models in the sphere of water resources management (Yajima and Derot 2018; Chen et al. 2020), as well as the comparisons of the performances of different ensemble learning models. This methodology, however, is in full swing and an increasingly larger number of studies on their application can be expected in this environmental field.

**Experimental**

In this study, the representative examples of supervised machine learning algorithms are applied for the prediction of the DO value. The modelling processes were conducted using the Rapid Miner Software version 8.9. The methods General Linear Model, Deep Learning, Decision Tree, Random Forest, Gradient Boosted Trees and Support Vector Machine are considered. The proposed methodology is depicted in Figure 1.

**Figure 1.** Methodology flow

*Study Area*

The research area of this study is the Tisa River on its flow through Serbia, with all the related tributaries of this river. The Tisa River is one of the biggest rivers in Central and East Europe. It springs in Ukraine, then runs through Romania and Hungary in a length 966 km, with the water area of around 156,087km2, only to end its flow in Serbia, where it flows into the Danube River. It contributes about 13% of the Danube's total runoff. Taking into consideration the fact that the Tisa River belongs to the basin of the Danube River, the second in length European river, the monitoring and effective management of the water of this river system is a question of international significance (Babić et al. 2019). The Tisa river lies approximately between 45˚8'17" of the North latitude, and 20˚16'39" of the East longitude. In this study, the investigation of the water quality of the Tisa River included 11 measuring stations situated on the main course of the Tisa and its tributaries.

*Dataset*

Apart from the dataset size, the reliability of the prediction model was also determined by the parameters used in training the learning models. The selection of the parameters analyzed in this paper was made based on the literature review (Nemati et al. 2015; Lerios and Villarica 2019; Ren et al. 2020). It was noticed that the river Temperature (T), Electrical Conductivity (EC), pH, Ammonium Ion (NH4-N) and Orthophosphates (PO4-P) had most frequently been used in the studies of this type. In this study, the mean monthly values of the mentioned parameters were used as the input data for DO prediction.

The physicochemical parameters were retrieved from the water quality monitoring system by the Serbian Environmental Protection Agency (SEPA) that samples river quality monthly, whereas the analysis covered the period from January 2011 to December 2018. After the elimination of the missing values, an extensive dataset was generated. The results of the descriptive statistics are shown in Table 1.

**Table 1.** Descriptive statistics

*Performance Indicators*

Model evaluation and selection are an important segment of machine learning. Quality measures are used to quantify the model's ability to predict. In this research study, four statistical formulas were selected, with the aim of determining prediction efficiency, namely Root Mean Square Error, Mean Absolute Error, Relative Error, Squared Error, Correlation, Squared Correlation. These measures were selected based on the frequency of their application in similar research studies (Zhang et al. 2017; Olyaie et al. 2017; Valerio et al. 2021).

**Results and Discussion**

*Variable Importance*

In this paper, the initial parameters for examining a possibility of the prediction of the DO value were pH, T, EC, NH4 and PO4. The initial dataset was subjected to a variable importance analysis, which defines the share of each mentioned parameter in DO prediction. In order to obtain the parameters that, have an influence on the final result on the same scale, the Z-score standardization of the dataset was first performed. According to the results, it can be noticed that the parameter T has the greatest significance for DO prediction. It is only followed by the parameters: pH, PO4 and EC, and ultimately NH4, as the parameter with the least influence.

According to the results obtained by the variable importance analysis, the four sets of the input variables were defined, while DO was the target variable in all the cases. The variables were being selected based on their significance in DO prediction. The first set consisted of all the analyzed parameters (T, pH, PO4, EC and NH4). In every next step, the set was reduced by removing the variable that had the least significance at that moment. So, the parameter NH4 was removed in the second dataset as the parameter of the least significance, whereas the remaining parameters T, pH, PO4 and EC were used for a prediction. The third set consists of the parameters T, pH and PO4. Finally, the last input variables set comprised the measured values of the two parameters (T and pH) rated as the most significant for DO prediction.

Every created dataset individually was processed by using the most representative machine learning techniques, namely the General Linear Model (GLM), Deep Learning (DL), Decision Tree (DT), Random Forest (RF), Gradient Boosted Trees (GBT) and the Support Vector Machine (SVM).

The goal was to define an optimal dataset and the model that would be enable reduction in costs, simultaneously maximizing prediction reliability. Based on the obtained values of various performance indicators, the efficiency of each model in DO prediction, as well as the most optimal number of the parameters, were assessed.

In the following section of the paper, various input data combinations are investigated, and the model is assessed in terms of prediction performance using numerical indices and graphical comparisons. In all the cases, the GLM, DL, DT, RF, GBT and SVM algorithms are used for modeling.

*DO prediction based on the T, pH, EC, PO4 and NH4 parameters*

In the first case, the dataset consisted of all the selected parameters (T, pH, EC, PO4 and NH4) was used as the input. Based on the Model Performance Metrics, a comparison of the predictions with the actual data (Table 2) was made.

**Table 2**. Comparison of model performances for the first data set

The obtained results indicate that three of the five selected performance indicators recognize the DL algorithm that is the most reliable in DO prediction. The use of the selected supervised machine algorithms for DO prediction using the first dataset rendered the graphs presented in Figure 4. The determination coefficient (r2) that estimates 0.763 indicates that the DL renders the best prediction of the DO value.

**Figure 2**. Scatter plots of actual versus predicted DO values for the first data set.

(1) GLM, (2) DL, (3) DT, (4) RF, (5) SVM

The total time needed for each algorithm to process data was taken into consideration as well. From this point of view, the DL model has satisfactory characteristics, given the fact that it ranked second as per the data processing speed. Contrary to said, the SVM model took the most time when speaking about this dataset.

*DO prediction based on the T, pH, PO4 and EC parameters*

In the second dataset, the T, pH, PO4 and EC parameters were taken as the input variables. After this dataset had been processed, certain values were obtained for the performance indicators for each machine learning technique (Table 3).

**Table 3.** Comparison of model performances for the second data set

**Figure 3**. Scatter plots of actual versus predicted DO values for the second data set.

1. GLM, (2) DL, (3) DT, (4) RF, (5) SVM

The values of all the performance indicators indicate that the SVM model is the most reliable in DO prediction based on the second set of the input variables. The value of the determination coefficient (r2), which is 0.774, is considered as a good value. It should be highlighted, however, that the values obtained for the DL model are almost equal to the values of the SVM model, so it can also be distinguished as a significantly reliable model. In Figure 3, the graphs indicating the accuracy level obtained for all the analyzed models are presented.

On the example of the second dataset, it can be noticed that modeling takes the most amount of time if an SVM algorithm is applied. Considerably less time is required by the other techniques, among which the DT, DL and GLM techniques can be distinguished.

*DO prediction based on the T, pH and PO4 parameters*

The next dataset processed by applying a set of different supervised machine learning techniques consists of the values of the three water quality indicators (T, pH and PO4). According to the results shown in Table 3, the SVM is again distinguished as a model whose performance indicators show that it achieves the highest accuracy level. As in the previous case, the DL technique is almost equally efficient in DO prediction based on this combination of the input variables.

**Table 4**. Comparison of model performances for the third data set

The graphical demonstration of the comparison of the actual values versus the predicted values with the calculated determination coefficient in this case is presented in Figure 4.

Comparing the model building time, the SVM sets an example with respect to the longest period needed for it to process data. Contrary to that, the DT, GLM and DL algorithms have the shortest total scoring time.

**Figure 4**. Scatter plots of actual versus predicted DO values for the third data set.

(1) GLM, (2) DL, (3) DT, (4) RF, (5) SVM

*DO prediction based on the T and pH parameters*

Finally, the last group of the variables comprises only the two parameters that distinguished themselves based on the variable importance analysis as the parameters with the biggest share in DO prediction. In this case, too, the performances of the different models were measured by means of the six indicators (Table 4).

**Table 5.** Comparison of model performances

The comparative analysis of the obtained variables shows that the SVM is distinguished as more reliable as per all the indicators than the remaining five models based on different supervised machine learning algorithms, only to be followed by DL, GLM, GBT, RF and DT.

**Figure 5**. Scatter plots of actual versus predicted DO values for the fourth data set.

1. GLM, (2) DL, (3) DT, (4) RF, (5) SVM

Based on the total scoring time, it can be concluded that, as in the previous cases, the SVM-algorithm-based modeling convincingly takes the most time in relation to the modeling based on the remaining algorithms does.

**Discussion**

This research explores an alternative machine learning method in order to predict water quality using minimal and easily available water quality parameters. A set of representative machine learning algorithms were employed. Dissolved oxygen is one of the most significant indicators of surface water quality. Therefore, finding the most reliable model for the prediction of the values of this parameter is quite significant. In this paper, the pH, T, PO4, EC and NH4 parameters were selected as the input variables. The first step in the set model was to determine the significance of each of the mentioned variables in DO prediction by carrying out variable importance analysis. The results showed that their significance decreased in the following order: T, pH, PO4, EC and NH4.Apart from the initial dataset, yet three more datasets were created, simultaneously excluding the parameter of the least significance from each next dataset. Each dataset was processed by the GLM, DL, DT, RF and SVM machine learning techniques, which was aimed at defining the dataset and the technique whose combination led to the most reliable prediction and the lowest costs of monitoring. The reliability of prediction was assessed based on the values of the six different indicators and graphically presented. The scoring time of all the models was also compared.

The calculated values of the indicators pointed to the fact that, in the largest number of the cases, the SVM learning technique had the best performances, this being particularly true when the input dataset comprised the T, pH, PO4, and EC (r2 = 0.774) parameters. According to the voluminous literature, the advantages of the SVM reflect in global optimization, the sparsity of the solution, and non-linearity and generalization (Ding et al. 2018). The understandable mechanism and the prediction accuracy of this method make it preferable among the other methods.

The values of the coefficients obtained in the case of the DL-algorithm-based modeling are insignificantly lower. On the other hand, the difference in the scoring time is significant and is to the advantage of the DL technique that takes much shorter time for modeling. Many authors intensively work on achieving significant improvements in accuracy by amending and modifying the different functions and procedures of the deep learning procedures, with the aim of overcoming the existing shortcomings by exploiting unlabeled data so as to improve performance (Sajjadi et al. 2016; Hailat et al. 2018).

In this study, the DT model achieved the lowest degree of the DO parameter prediction reliability, whereas the advantage of this technique reflects in the fast calculation speed.

**Conclusions**

Surface water quality parameter modeling is considered as a very significant issue, which determines the efficient and effective water resources management system. A significant reduction in costs can be achieved by finding reliable prediction models and applying the techniques capable of determining the related variables of water quality values in the future.

The results presented in this research study have indicated that prediction of the water quality (expressed through the DO values) of the Tisa River on its flow through Serbia can be made with the highest degree of precision by applying the SVM and DL machine learning methods, namely in the case when an input dataset consist of the values of the T, pH, PO4 and EC parameters. Should the scoring time also be taken into consideration, a significant advantage can be given to the DL algorithm, which takes much shorter modeling time than the SVM does. The result analysis allows us to conclude that, in this particular case, the DL technique has advantage if there is a need for a quick reaction which will not jeopardize the prediction ability to a great extent.

The recommendations for further research studies relate to the selection of a larger number of input variables and the creation of a hybrid model in which an integration of different machine supervised techniques would be performed, which would help overcome shortcomings and take an advantage of each one of them.

**Funding**

The research presented in this paper was done with the financial support of the Ministry of Education, Science and Technological Development of the Republic of Serbia, within the funding of the scientific research work at the University of Belgrade, Technical Faculty in Bor, according to the contract with registration number 451-03-9/2021-14/ 200131.

**References**

Alizadeh, F., Gharamaleki, A.F., Jalilzadeh, M. *et al.* 2020. “Prediction of River Stage-Discharge Process Based on a Conceptual Model Using EEMD-WT-LSSVM Approach.” *Water Resour.* 47**,**41–53 (2020). <https://doi.org/10.1134/S0097807820010066>

Babić, G., M. Vuković, D. Voza, Lj. Takić and I. Mladenović-Ranisavljević. 2019. "Assessing Surface Water Quality in the Serbian Part of the Tisa River Basin." *Pol. J. Environ. Stud*., 28(6): 4073-4085 <https://doi.org.10.15244/pjoes/95184>

Chen, K., H. Chen, C. Zhou, Y. Huang, X. Qi, R. Shen, R., F. Liu, M. Zuo, X. Zou, J. Wang, Y. Zhang, D. Chen, X. Chen, Y. Deng and H. Ren. 2020. "Comparative analysis of surface water quality prediction performance and identification of key parameters using different machine learning models based on big data." *Water Res*. 171: 115454. <https://doi.org/10.1016/j.watres.2019.115454>

Ding, S., Z. Zhu and X. Zhang. 2017. "An overview on semi-supervised support vector machine." *Neural. Comput. Appl*. 28: 969–978. <https://doi.org/10.1007/s00521-015-2113-7>

Everaert G., I. Pauwels, E. Bennetsen and P. Goethals. 2016. "Development and Selection of Decision Trees for Water Management: Impact of Data Preprocessing, Algorithms and Settings." *AI Commun*. 29 (6): 711-723. <https://doi.org.10.3233/AIC-160711>

Granata, F., S. Papirio, G. Esposito and G. de Marinis. 2017. "Machine learning alghoritms for the forecasting of wastewater quality indicators." *Water*. 9(2): 105. <https://doi.org/10.3390/w9020105>

Hailat, Z., A. Komarichev and X. Chen. 2018. "Deep Semi-Supervised Learning." In: 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 2018 pp. 2154-2159. <https://doi.org.10.1109/ICPR.2018.8546327>

Hipni, A., A. El-Shafie, A. Najah, O. A. Karim, A. Hussain and M. Mukhlisin. 2013. "Daily Forecasting of Dam Water Levels: Comparing a Support Vector Machine (SVM) Model With Adaptive Neuro Fuzzy Inference System (ANFIS)." *Water Resour. Manag.* 27: 3803–3823 <https://doi.org.10.1007/s11269-013-0382-4>

Ho, J. H., H. A. Afan, H. Amr, S. El-Shafie, B. Koting, N. Syuhadaa et al. 2019. "Towards a time and cost effective approach to water quality index class prediction. "*J. Hydrol*. 575: 148-165 <https://doi.org/10.1016/j.jhydrol.2019.05.016>

Liaw, A. and M. Wiener M. 2002. "Classification and Regression by Random Forest. "R News, 2/3: 18-22.

Lerios, L. J. and V. M. Villarica. 2019. "Pattern extraction of water quality prediction using machine learning algorithms of water reservoir. "*Int. J. Mech. Eng. Robot. Res.* 8(6). <https://doi.org.10.18178/ijmerr.8.6.992-997>

Najah, A. A., A. El-Shafie and O. A. Karim. 2012. "Water quality prediction model utilizing integrated wavelet-ANFIS model with cross-validation. "*Neural Comput. Appl.* 21: 833-841. <https://doi.org/10.1007/s00521-010-0486-1>

Nemati, S., H. M. Fazelifard, O. Terzi, A. M. Ghorbani. 2015. "Estimation of dissolved oxygen using data-driven techniques in the Tai Po River, Hong Kong. "*Environ. Earth Sci.* 74: 4065–407. <https://doi.org.10.1007/s12665-015-4450-3>

Olyaie, E., Z. H. Abyanaeh and D. A. Mehr. 2017. "A comparative analysis among computational intelligence techniques for dissolved oxygen prediction in Delaware River. "*Geosci. Front.* 8: 517-527 <http://dx.doi.org/10.1016/j.gsf.2016.04.007>

Parsaie, A. and A. H. Haghiabi. 2015. "The effect of predicting discharge coefficient by neural network on Increasing the Numerical Modeling Accuracy of Flow Over Side Weir. "*Water Resour. Manag.* 29: 973–985 <https://doi.org/10.1007/s11269-014-0827-4>

Ren, Q., X. Wang, W. Li, Y. Wei, D. An. 2020. "Research of dissolved oxygen prediction in recirculating aquaculture systems based on deep belief network. "*Aquac. Eng.* 90: 102085. <https://doi.org/10.1016/j.aquaeng.2020.102085>

Sajjadi, M., M. Javanmardi and T. Tasdizen. 2016. "Regularization With Stochastic Transformations and Perturbations for Deep Semi-Supervised Learning. "NIPS.

Valerio, C., L. De Stefano, G. Martinez-Munoz and A. Garrido 2021. "A machine learning model to assess the ecosystem response to water policy measures in the Tagus River Basin (Spain)." *Sci. Total Environ.* 750: 141252. <https://doi.org/10.1016/j.scitotenv.2020.141252>

Yajima, H. and J. Derot. 2018. "Application of the Random Forest model for chlorophyll-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases." *J. Hydroinform.* 20: 206–220 <https://doi.org/10.2166/hydro.2017.010>

Xiaoliang, Ji and X. Shan A. R. Dahlgren and M. Zhang. 2017. "Prediction of dissolved oxygen concentration in hypoxic river systems using support vector machine: a case study of Wen-Rui Tang River, China." *Environ. Sci. Pollut. Res.* 24: 16062–16076. <https://doi.org/10.1007/s11356-017-9243-7>

Xu L. and S. Liu 2013. "Study of short term water quality prediction model based on wavelet neural network." *Math. Comput. Model.* 58: 807-813. <https://doi.org/10.1016/j.mcm.2012.12.023>

Zhang, L., Z. Zou and W. Shan. 2017. "Development of a method for comprehensive water quality forecasting and its application in Miyun reservoir of Bejing, China." *J. Environ. Sci*. 56: 240-246. <https://doi.org/10.1016/j.jes.2016.07.017>

**Tables**

**Table 1.** Descriptive statistics

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | N | Min | Max | Mean | Std. Deviation | Skewness | |
| Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Std. Error |
| T | 1006 | ,1 | 30,2 | 14,343 | 8,5210 | -,039 | ,077 |
| pH | 981 | 7,20 | 9,10 | 8,0492 | ,27980 | ,304 | ,078 |
| EC | 1006 | 0 | 2,450 | 679,03 | 410,324 | 1,445 | ,077 |
| NH4 | 1006 | ,01 | 3,28 | ,1073 | ,23135 | 7,236 | ,077 |
| PO4 | 1005 | ,000 | 2,460 | ,14979 | ,244130 | 3,640 | ,077 |
| DO | 1004 | ,9 | 93,0 | 9,300 | 4,2511 | 7,643 | ,077 |

**Table 2**. Comparison of model performances for the first data set

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | RMSE | Absolute error | Relative error | Mean Squared error | Correlation | Squared correlation |
| General Linear Model | 1,755 | 1,293 | 13,6% | 3,134 | 0,835 | 0,699 |
| Deep learning | **1,573** | 1,149 | 12,3% | **2,505** | **0,873** | **0,763** |
| Decision tree | 2,832 | 1,636 | 16,0% | 10,374 | 0,698 | 0,514 |
| Random forest | 1.964 | 1,324 | 13,4% | 3,985 | 0,793 | 0,635 |
| Gradient Boosted Trees | 2,12 | 1,455 | 14,4% | 4,746 | 0,779 | 0,613 |
| Support Vector Machine | 1,58 | **1,112** | **11,9%** | 2,517 | 0,869 | 0,756 |

**Table 3.** Comparison of model performances for the second data set

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | RMSE | Absolute error | Relative error | Squared error | Correlation | Squared correlation |
| General Linear Model | 1,755 | 1,293 | 13,6% | 3,134 | 0,835 | 0,699 |
| Deep learning | 1,569 | 1,172 | 12,7% | 2,489 | 0,875 | 0,766 |
| Decision tree | 2,797 | 1,63 | 15,5% | 8,954 | 0,687 | 0,495 |
| Random forest | 2,011 | 1,348 | 13,5% | 4,217 | 0,789 | 0,629 |
| Gradient Boosted Trees | 2,122 | 1,433 | 14,3% | 4,812 | 0,78 | 0,616 |
| Support Vector Machine | **1,525** | **1,082** | **11,8%** | **2,349** | **0,88** | **0,774** |

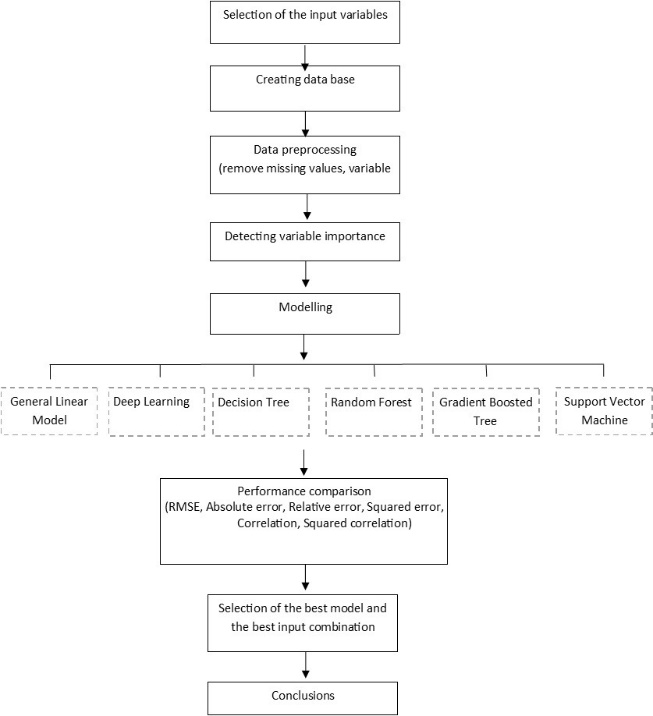
**Table 4**. Comparison of model performances for the third data set

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | RMSE | Absolute error | Relative error | Squared error | Correlation | Squared correlation |
| General Linear Model | 1,777 | 1,316 | 13,9% | 3,201 | 0,829 | 0,689 |
| Deep learning | 1,685 | 1,258 | 13,6% | 2,867 | 0,854 | 0,73 |
| Decision tree | 3,028 | 1,672 | 15,5% | 11,108 | 0,679 | 0,486 |
| Random forest | 1,951 | 1,325 | 13,5% | 3,942 | 0,795 | 0,637 |
| Gradient Boosted Trees | 2,352 | 1,477 | 14,5% | 6,304 | 0,746 | 0,572 |
| Support Vector Machine | **1,633** | **1,158** | **12,4%** | **2,704** | **0,859** | **0,739** |

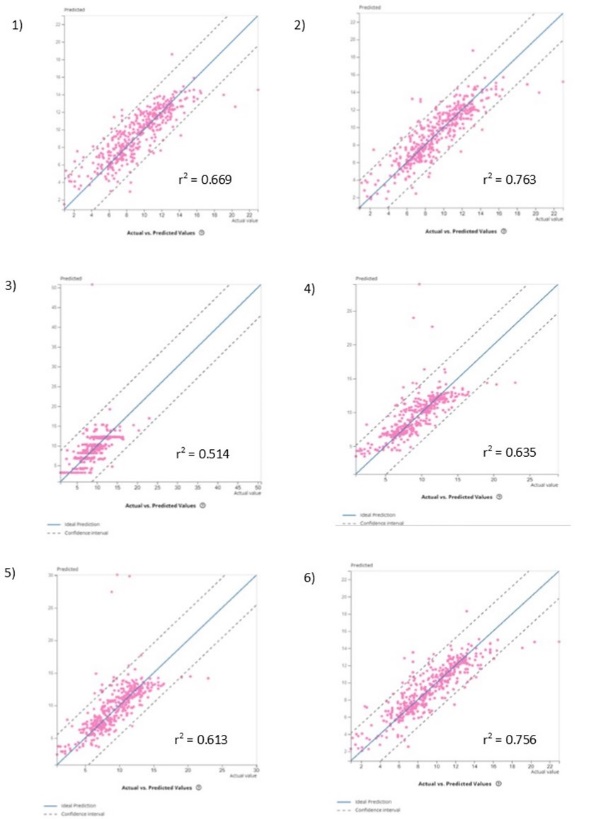
**Table 5.** Comparison of model performances

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | RMSE | Absolute error | Relative error | Squared error | Correlation | Squared correlation |
| General Linear Model | 1,963 | 1,465 | 15,5% | 3,891 | 0,788 | 0,623 |
| Deep learning | 1,947 | 1,496 | 16,2% | 3,809 | 0,798 | 0,637 |
| Decision tree | 3,074 | 1,9 | 17,3% | 10,426 | 0,567 | 0,355 |
| Random forest | 2,529 | 1,668 | 16,1% | 7,16 | 0,68 | 0,488 |
| Gradient Boosted Trees | 2,105 | 1,518 | 15,5% | 4,464 | 0,766 | 0,588 |
| Support Vector Machine | **1,917** | **1,353** | **14,3%** | **3,701** | **0,806** | **0,651** |

**Figures**

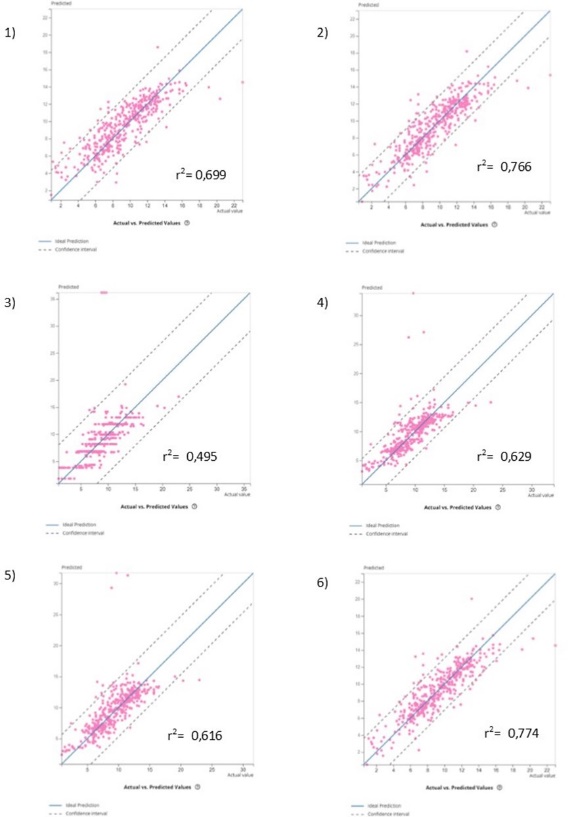
****

**Figure 1.** Methodology flow

****

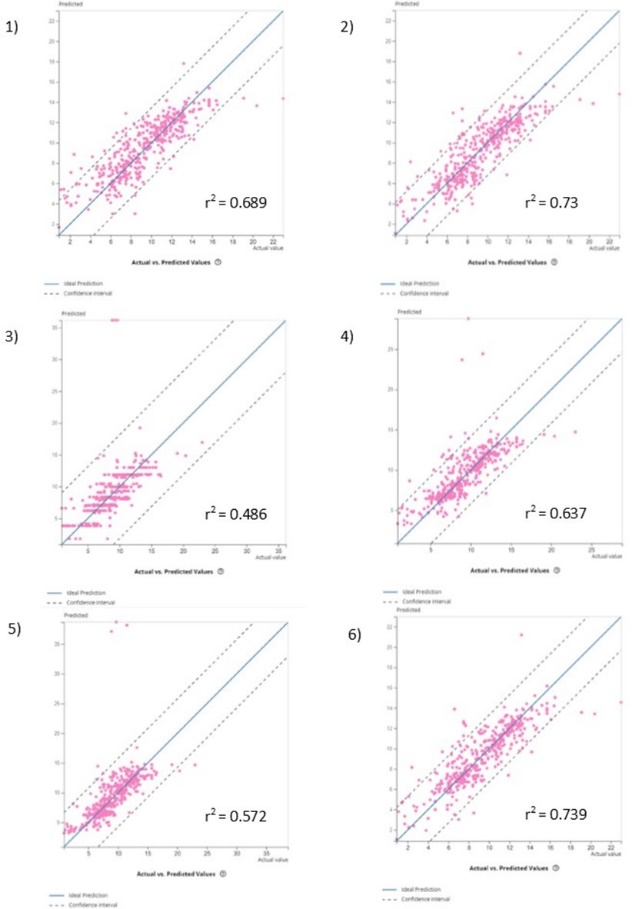
**Figure 2**. Scatter plots of actual versus predicted DO values for the first data set.

(1) GLM, (2) DL, (3) DT, (4) RF, (5) SVM

****

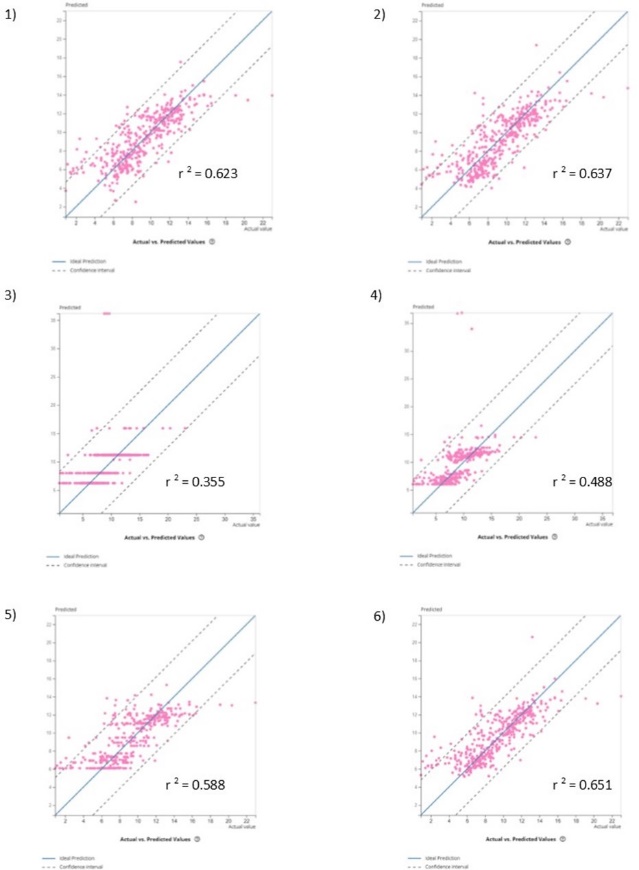
**Figure 3**. Scatter plots of actual versus predicted DO values for the second data set.

1. GLM, (2) DL, (3) DT, (4) RF, (5) SVM

****

**Figure 4**. Scatter plots of actual versus predicted DO values for the third data set.

(1) GLM, (2) DL, (3) DT, (4) RF, (5) SVM

****

**Figure 5**. Scatter plots of actual versus predicted DO values for the fourth data set.

1. GLM, (2) DL, (3) DT, (4) RF, (5) SVM