International Journal of Engineering, Science & Mathematics
Vol.13Issue01,January2024,
ISSN:2320-0294ImpactFactor:6.765
JournalHomepage:http://www.ijmra.us,Email:editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed &Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell’s Directories of Publishing Opportunities, U.S.A

EVALUATING THE LEVELS OF DISSOLVED OXYGEN IN US LAKES – AN ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING USE CASE

Olabode Soetan

Abstract

The dissolved oxygen (DO) levels in lakes play a critical role in assessing environmental conditions and minimizing water treatment costs. This research investigates the impact of various factors, including temperature, salinity, and atmospheric pressure, on DO concentrations. Maintaining healthy water, vital for aquatic life, requires DO concentrations above 6.5-8 mg/L and between 80-120%. The study explores the rise in sea level and temperature, emphasizing the need for research in this changing environment. The literature review delves into trace metal contamination in coastal waters, highlighting the influence of non-point sources, water temperature, and DO on metal cycling. Utilizing linear regression and logistic regression analyses, the study establishes a comprehensive model. The results reveal that DO saturation, conductivity, and pH positively affect DO levels, while depth, temperature, and pH negatively impact them. Logistic regression indicates significant regressors, demonstrating the model's efficacy in classification. The research incorporates machine learning techniques such as neural networks, Classification Tree, Random Forest, Bagging, and Boosting. Among these, the Classification Tree method exhibits outstanding performance with zero misclassification error, 100% sensitivity, specificity, and accuracy. In conclusion, this research employs a multidimensional approach, combining traditional statistical methods with advanced machine learning techniques. The findings provide valuable insights for understanding and managing dissolved oxygen levels in lakes, especially in the context of changing environmental conditions.

1.0 Background

The level of dissolved oxygen (DO) in lakes is important for assessing environmental conditions as well as reducing water treatment costs. High DO often precede toxic algal blooms, while low DO causes carcinogenic metals to precipitate in water treatment (Durell et al., 2022). When dissolved oxygen becomes too low, fish and other aquatic organisms cannot survive. The colder water is, the more oxygen it can hold. As the water becomes warmer, less oxygen can be dissolved in the water.

The amount of oxygen that can be dissolved in water depends on several factors, including water temperature, the number of dissolved salts present in the water (salinity), and atmospheric pressure. Healthy water should generally have dissolved oxygen concentrations above 6.5-8 mg/L and between about 80-120 %.Going by the rise in sea level and temperature, there is need to carry out research of this nature.

2.0 Literature Review

An increasing body of evidence suggests that much of the trace metal contamination observed in coastal waters is no longer derived from point-source inputs, but instead originates from diffuse, non-point sources. Previous research has shown that water temperature and dissolved oxygen regulate non-point source processes such as sediment diagenesis; however, limited information is available regarding the effect of these variables on toxic trace metal cycling and speciation in natural waters (Beck &Sañudo-Wilhelmy, 2007).

Overall, for every 1 °C River Water Temperature (RWT) increase, there will be about 2.3% decrease in DO saturation level concentrations over Indian catchments under climate signals (Rajesh &Rehana, 2022).

Furthermore, in a study to verify the usefulness of water quality indices, as the indicators of water pollution, for assessment of spatial-temporal changes and classification of river water qualities, results revealed the serious negative effects of the city urban activity on the river water quality.In the studied section of the river, the water quality index (WQI) was 71 units (classified as good) at the entry station and 47.6 units (classified as bad) at the outlet station. For the studied period, a significant decrease in water quality (mean WQI decrease = 11.6%, p = 0.042) was observed in the rural areas. A comparative analysis revealed that the urban water quality was significantly bad as compared with rural. The analysis enabled to classify the water quality stations into three groups: good water quality, medium water quality and bad water quality. Four water quality indices were investigated: WQI (considering 18 water quality parameters), WQI(min) and WQI(m) (considering five water quality parameters: temperature, pH, DO, Electrical Conductivity, and Total Suspended Solids) and WQI(DO) (considering a single parameter, DO) (Kannel et al., 2007).

3.0 Exploratory Data Analysis

The target variable in this dataset is Dissolved Oxygen (DO). This dataset contains 5 regressors which includes: Depth, Temperature, DO Saturated, pH, and Conductivity.

Depth: depth from the lake surface (0) to bottom (10)

Temperature (Temp): this is the temperature of the water at the time of collating data

DO Saturated (DOsat): dissolved oxygen saturated is a measure of how much oxygen is dissolved in the water - the amount of oxygen available to living aquatic organisms.

pH: potency of hydrogen ion. Between 0-14. It is a measure of how acidic or basic an aqueous solution is.

Conductivity: conductivity is a measure of the ability of water to pass an electrical current.

The boxplot below revealed that DO Saturated and Temperature are negatively skewed, Conductivity and Depth shows a normal distribution, whilst the pH values are positively skewed. Positively skewed data will have a mean that is higher than the median. The mean of negatively skewed data will be less than the median, which is the exact reverse of what happens in a positively skewed distribution. No matter how long or fat the tails are, the distribution exhibits zero skewness if the data is graphed symmetrically (Chen, 2022). The mean, which is the average of all the values in a positively skewed distribution, is higher than the median because the data is more heavily weighted toward the lower side. The median, on the other hand, is the data's middle value. The entire data is occupied with a lot of outliers to an extent, more evidently in DO Saturated and Conductivity.

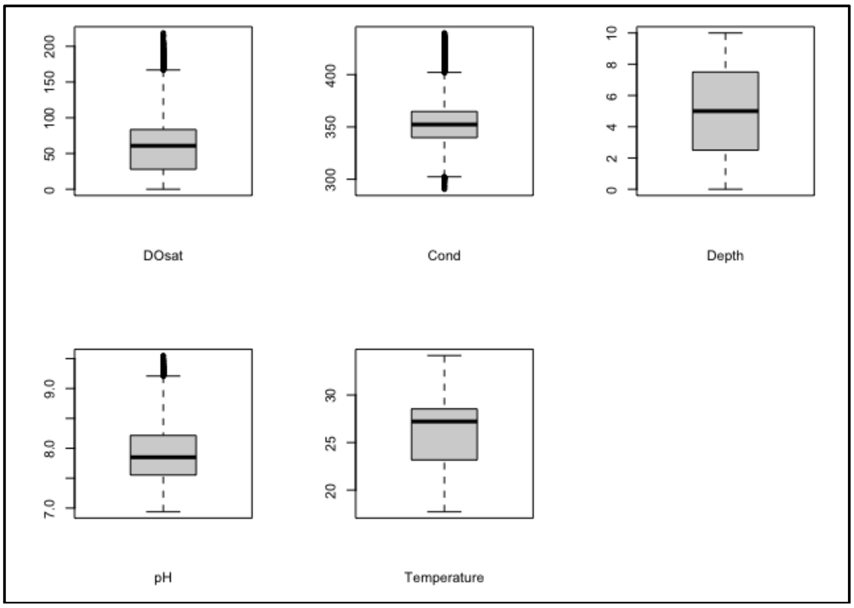


fig.1: boxplot

On the other hand, the scatterplot below shows the relationship between the regressors and the response variable DO.DOsat, pH, and Conductivity all have a positive relationship with DO although Conductivity has a low positive relationship with DO. This implies that the higher the pH, DOsat and

Conductivity, the higher the DO level. In contrast, Temperature has a low negative relationship with DO whilst depth has no correlation with DO.

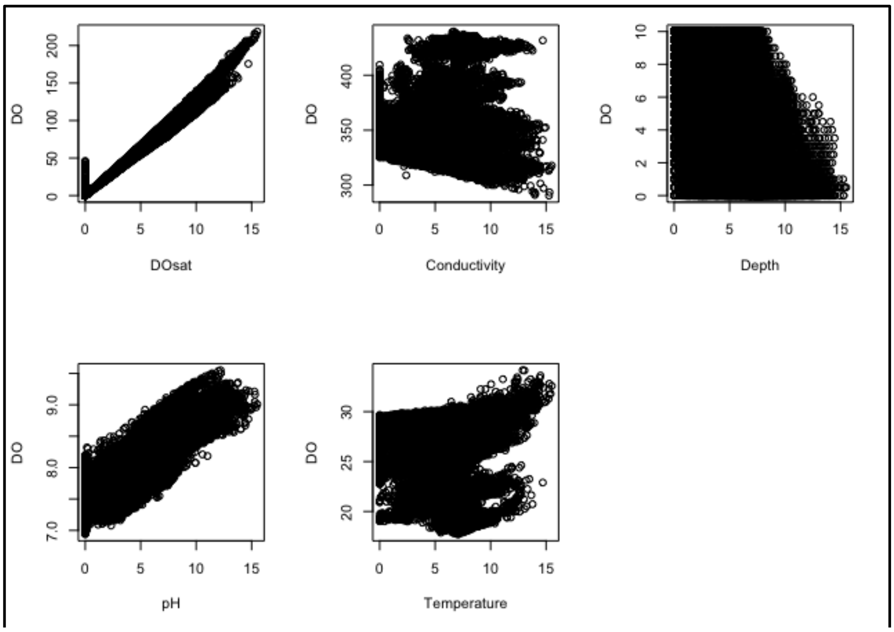


fig. 2: scatterplot

The table below further affirms the relationship between DO and the regressors. DOsat and pH are highly positively correlated to DO whilst Conductivity is slightly positively correlated. On the other hand, Depth and Temperature have a negative correlation with DO as evident in fig. 2 above.

	DO	DEPTH	TEMP	DOSAT	PH	COND
DO	1	-0.49	-0.31	0.99	0.83	0.19
DEPTH	-0.49	1	-0.13	-0.52	-0.4	0.07
TEMP	-0.31	-0.13	1	-0.2	-0.1	-0.59
DOSAT	0.99	-0.52	-0.2	1	0.85	0.11
PH	0.83	-0.4	-0.1	0.85	1	0.14
COND	0.19	0.07	-0.59	0.11	0.14	1

table 1: correlation matrix

The chart below further supports the relationship between DO and the regressors. This is simply a graphical representation of the data in table 1 above. DOsat and pH are highly positively correlated to DO whilst Conductivity is slightly positively correlated. On the other hand, Depth and Temperature have a negative correlation with DO.

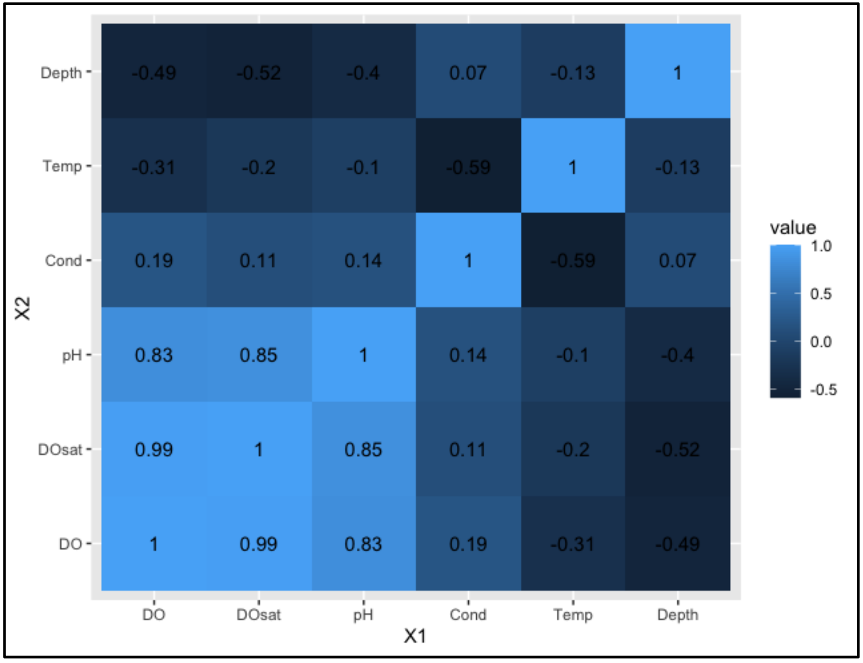


fig. 3: heatmap

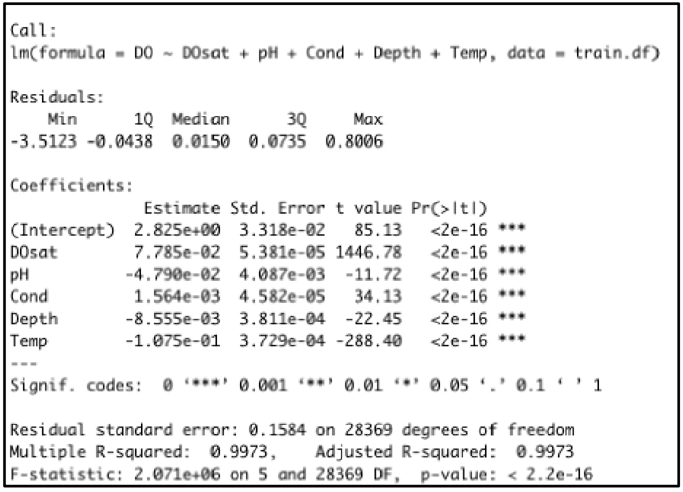
4.0 **Analysis**

4.1 **Linear Regression**

The dataset has 47,292 rows of data. For the purpose of a regression analysis, the dataset was partitioned in a 60% by 40% ratio with 60% of the data assigned to training dataset and 40% assigned to validation dataset.

A linear regression model was built on the training dataset. The image in fig 4 below shows that all the regressors are highly significant even in the lowest significance level. Furthermore, with a **residual standard error** of **0.16** the model is said to be very useful for predicting the response variable. The residual standard error is used to measure how well a regression model fits a dataset. The smaller the residual standard error, the better a regression model fits a dataset. Also, R-squared measures the goodness of fit of a regression model. Hence, a higher R-squared indicates the model is a good fit while a lower R-squared indicates the model is not a good fit. In this case, an **R squared** and **Adjusted R squared** values of **99.7%** tells how well fit the model is to the data. The linear regression formula for this model (model 1) would thus be written as:

$\hat{Y}$=2.825 + 0.0779DOsat – 0.0479pH + 0.00156Cond – 0.00856Depth – 0.108Temp



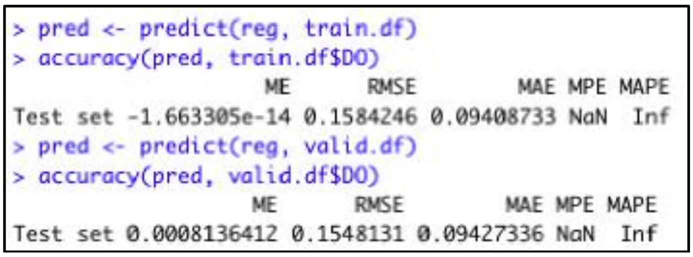
```
Call:
lm(formula = DO ~ DOsat + pH + Cond + Depth + Temp, data = train.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5123 -0.0438  0.0150  0.0735  0.8006 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.825e+00  3.318e-02   85.13   <2e-16 ***
DOsat        7.785e-02  5.381e-05 1446.78   <2e-16 ***
pH          -4.790e-02  4.087e-03  -11.72   <2e-16 ***
Cond         1.564e-03  4.582e-05   34.13   <2e-16 ***
Depth       -8.555e-03  3.811e-04  -22.45   <2e-16 ***
Temp        -1.075e-01  3.729e-04 -288.40   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1584 on 28369 degrees of freedom
Multiple R-squared:  0.9973,	Adjusted R-squared:  0.9973 
F-statistic: 2.071e+06 on 5 and 28369 DF,  p-value: < 2.2e-16
```

fig. 4: details of the result of a regression analysis on the training data

The linear regression formula above implies that a unit increase in *DOsat*, and *Cond* would cause the level of dissolved oxygen to increase by 0.0779mg/L and 0.00156 mg/L respectively. On the other hand, a unit increase in *pH*, *Depth*, and *Temp* would cause the level of dissolved oxygen to reduce by 0.0479 mg/L, 0.00856 mg/L, and 0.108 mg/L respectively.

Sum of Squared Errors (SSE) tells how much of Y is left unexplained. It tells how much cannot be attributed to a linear relationship. The Mean Square Error (MSE) on the other hand tells how close a regression line is to a set of points. It takes the distance from these points to the regression line and squares them. A high SSE/MSE suggests that the data have a reasonable degree of differences between them and may not usable. Root Mean Squared Error (RMSE) is simply the square root of the MSE. Lower values of RMSE indicate a better fit. In this case, a low **RMSE** of **0.16** for both the training and validation data as seen in fig. 5 below indicates a good fit. A low Mean Absolute Error (**MAE**) of **0.09** for both the training and validation data also further confirms that the model is well fit. This tells that the distance between the real data and the predicted data is quite minute.

Since both the training and validation data perform in the same manner, then there is no case of overfitting in this test.



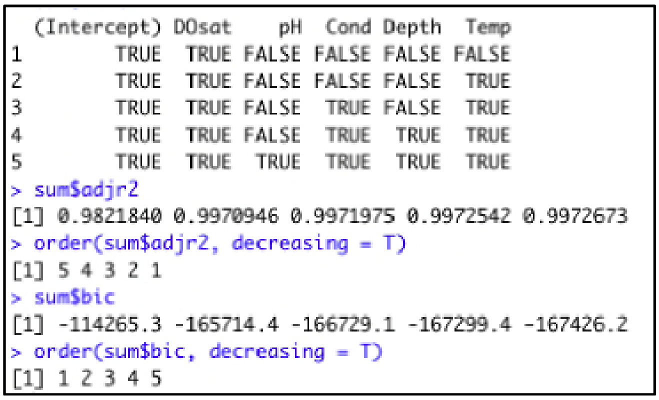
```
> pred <- predict(reg, train.df)
> accuracy(pred, train.df$DO)
                    ME      RMSE        MAE MPE MAPE
Test set -1.663305e-14 0.1584246 0.09408733 NaN  Inf
> pred <- predict(reg, valid.df)
> accuracy(pred, valid.df$DO)
                   ME      RMSE        MAE MPE MAPE
Test set 0.0008136412 0.1548131 0.09427336 NaN  Inf
```

fig. 5: details of the result of a regression analysis on the training data

To identify the best variables fit for this model, three variable selection procedures were employed (i.e.

exhaustive search, forward selection, and backward elimination). Results from the exhaustive search approach revealed that a model with all the regressors (subset 5) would perform better than any other model in terms of Bayesian Information Criterion (BIC) and Adjusted R Squared performance. An **Adjusted R squared** and **BIC** of **0.9973** and **-167426** respectively gave subset 5 an edge over other subsets as displayed in fig. 6 below.

BIC is a metric that is used to compare the goodness of fit of different regression models. As complexity of the model increases, BIC value increases and as likelihood increases, BIC decreases. Hence, lower BIC is better.



```
  (Intercept) DOsat    pH  Cond Depth  Temp
1        TRUE  TRUE FALSE FALSE FALSE FALSE
2        TRUE  TRUE FALSE FALSE FALSE  TRUE
3        TRUE  TRUE FALSE  TRUE FALSE  TRUE
4        TRUE  TRUE FALSE  TRUE  TRUE  TRUE
5        TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
> sum$adjr2
[1] 0.9821840 0.9970946 0.9971975 0.9972542 0.9972673
> order(sum$adjr2, decreasing = T)
[1] 5 4 3 2 1
> sum$bic
[1] -114265.3 -165714.4 -166729.1 -167299.4 -167426.2
> order(sum$bic, decreasing = T)
[1] 1 2 3 4 5
```

fig. 6: details of an exhaustive search variable selection approach on the training data

Similarly, the forward selection and backward elimination procedures gave identical and similar results. A graphical representation of this is as shown in fig. 7 below.

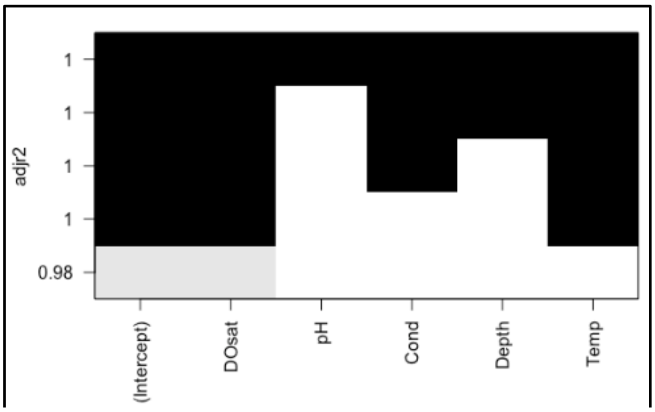


fig. 7: graphical representation of variable selected from the forward and backward selection methods

From the evaluation of the results of the variable selection procedures as well as results from a regression analysis on the entire training data set, it is in my opinion based on the structure of the dataset and the objective of the research topic to utilize a model with all 5 regressors present. Asides the fact that a model with all 5 regressors has performed better than others, the regressors are also equally significant individually. This model meets all the requirements for a good model in terms of the results from metrics like RMSE, MAE, BIC, R squared, Adjusted R squared and other analytical concepts.

Taking a look at the residuals, it is further established that the dataset fits well to the model. A larger chunk of the residuals is closer to zero as previously indicated with a **MAE** of **0.09**. Only a very minute portion of the residuals are outliers. The histogram in fig. 8 below also shows a normal distribution of the residuals. The basic assumption of regression model is normality of residual. If your residuals are not normal then there may be problem with the model fit, stability, and reliability (Roy, 2020).

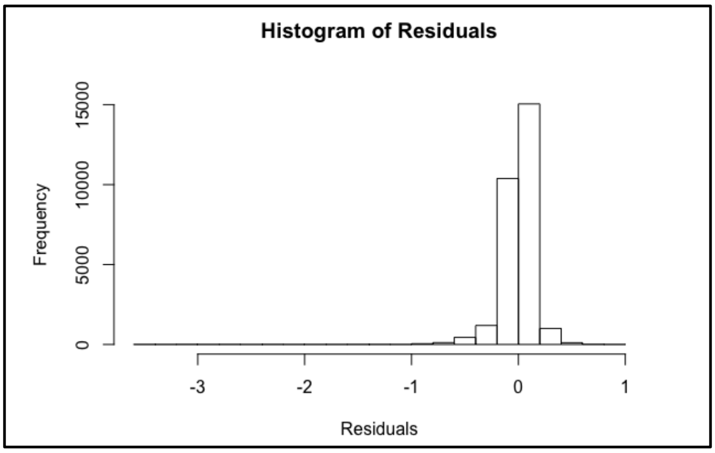


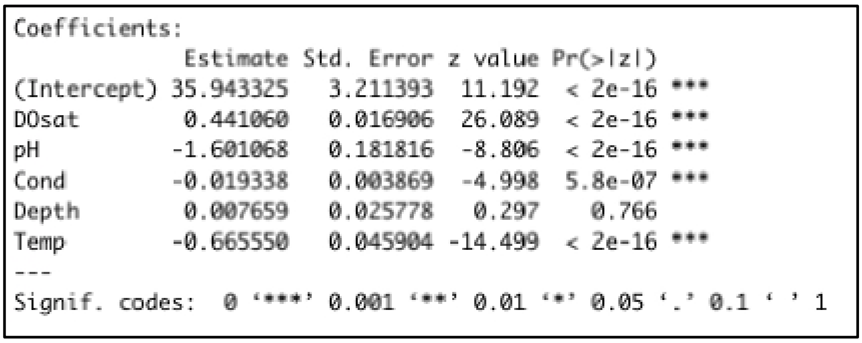
fig. 8: histogram of residuals.

4.2 **Logistic Regression**

As previously mentioned, the dataset has 47,292 rows of data. For the purpose of a regression analysis, the dataset was partitioned in a 60% by 40% ratio with 60% of the data assigned to training dataset and 40% assigned to validation dataset.

Based on the exhaustive search report as in fig. 6 above, a logistic regression model was built on the top 3 subsets (referred to as model 1, model 2, model 3 in subsequent sentences) on the training and validation dataset. The image in fig. 9 below shows that all the regressors except for 'Depth' are highly significant even in the lowest significance level. Depth is only significant at 77% significance level.

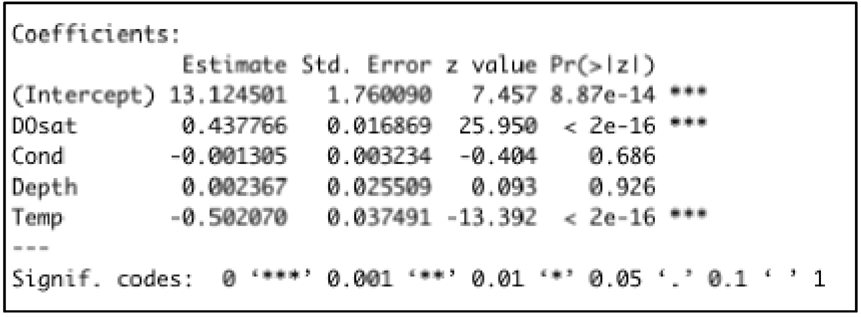
$$P = \frac{1}{1+e^{-(35.94+0.44DOsat\ -1.6pH\ -0.16Cond\ +0.008Dept\ h-0.666Temp\ )}}$$



```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept) 35.943325   3.211393  11.192  < 2e-16 ***
DOsat        0.441060   0.016906  26.089  < 2e-16 ***
pH          -1.601068   0.181816  -8.806  < 2e-16 ***
Cond        -0.019338   0.003869  -4.998 5.8e-07 ***
Depth        0.007659   0.025778   0.297    0.766    
Temp        -0.665550   0.045904 -14.499  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

fig. 9: model 1 variable significance.

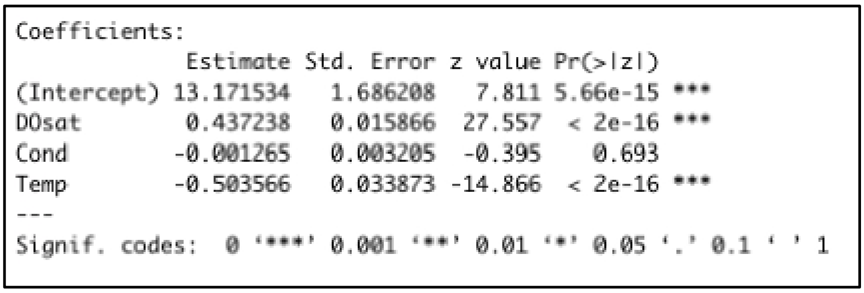
In the same vein, 'Depth' is also only significant in model 2 at 92% significance level as well as conductivity at 69% significance level. This is as shown in fig. 10.



```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 13.124501   1.760090   7.457 8.87e-14 ***
DOsat        0.437766   0.016869  25.950  < 2e-16 ***
Cond        -0.001305   0.003234  -0.404    0.686
Depth        0.002367   0.025509   0.093    0.926
Temp        -0.502070   0.037491 -13.392  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

fig. 10: model 2 variable significance.

Also, in model 3 with three regressors, conductivity is seen to be significant only at 69% significance level as in fig. 11 below.



```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 13.171534   1.686208   7.811 5.66e-15 ***
DOsat        0.437238   0.015866  27.557  < 2e-16 ***
Cond        -0.001265   0.003205  -0.395    0.693
Temp        -0.503566   0.033873 -14.866  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

fig. 11: model 3 variable significance.

Table 2 below gives a snapshot of the results of a logistic regression from the models mentioned above. Results of the training and validation classification data on model 1 gave a better result in terms of Misclassification Error, Accuracy, True Positive Rate (TPR, a.k.a. Sensitivity), False Positive Rate (FPR), and Specificity when all the regressors were selected compared to when the regression was executed on the other two models. The model with all regressors (i.e. model 1) had an overall **Misclassification Error** of **1.8%**, **Accuracy** of **98.2%**, **Sensitivity** of **98.8%** and **Specificity** of **91.8%**. The validation data also had a similar performance.

With near 100% performance in terms of specificity, sensitivity, and accuracy, this shows that the model is quite useful in correctly classifying the true positives, true negatives, and resulting in more correct predictions respectively. This is evident in the details presented in table 3 below. It is important to note that the aim of every good model however is to have the true negatives and true positives correctly

MODEL	AIC	OPTIMAL CUTOFF	MISCLASSIFICATION ERROR	TPR	FPR	SPECIFICITY	ACCURACY
MODEL 1 TD	3853.4	0.32	0.018	0.988	0.082	0.918	0.982
MODEL 1,2,3 VD	3853.4	0.33	0.019	0.988	0.085	0.915	0.981
MODEL 2 TD	3932.3	0.37	0.0215	0.983	0.065	0.935	0.978
MODEL 3 TD	3930.3	0.37	0.0215	0.983	0.065	0.935	0.978

predicted.

table 2: logistic regression model performances

For model 1, there were three hundred and nine cases of false positives and two hundred and ten cases of false negatives. This is a negligible portion of the entire training data set. Similar performance was observed in both model 2 and model 3 as well all though they had more cases of the wrong classifications than in model 1.

CONFUSION MATRIX MODEL 1 TD		
	0	**1**
0	2358	309
1	210	25498

CONFUSION MATRIX MODEL 2 TD		
	0	1
0	2401	444
1	167	25363

CONFUSION MATRIX MODEL 1,2,3 VD		
	0	**1**
0	1581	213
1	147	16976

CONFUSION MATRIX MODEL 3 TD		
	0	**1**
0	2401	444
1	167	25363

table 3: confusion matrix report for all logistic regression models

Furthermore, ROC curves and lift charts are other ways to evaluate the performance of a model. Evaluating the ROC curve and lift charts, it was observed that when all the available regressors were featured in the model, as well as other models, it appeared that the models are quite fitting with an AUC of **99%** and a significant lift in the cumulative actual of the predicted values. For a good model, the AUC is aimed to be closer to **100%** (the optimum classifier) and a lift chart showing evident lift in the lift chart curve. Details are in the figures below.

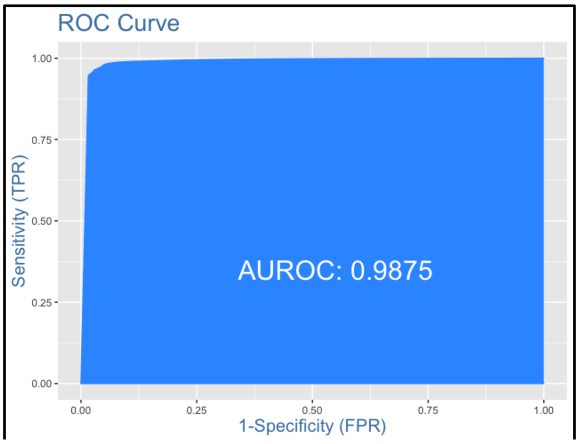


fig. 12: model 1 (training data) ROC curve

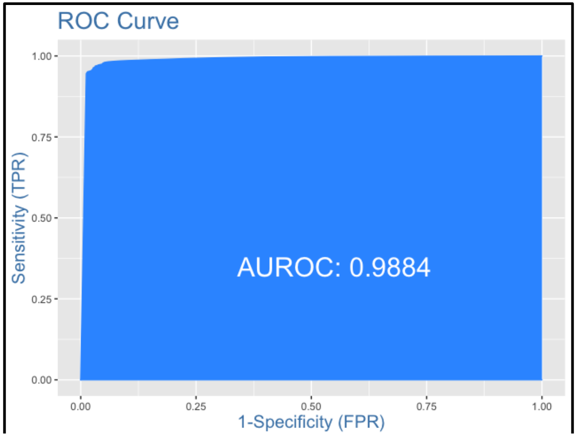


fig. 13: model 1 (validation data) ROC curve

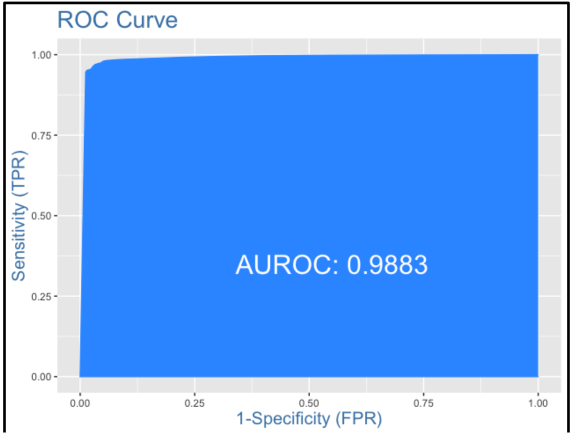


fig. 14: model 2 ROC curve

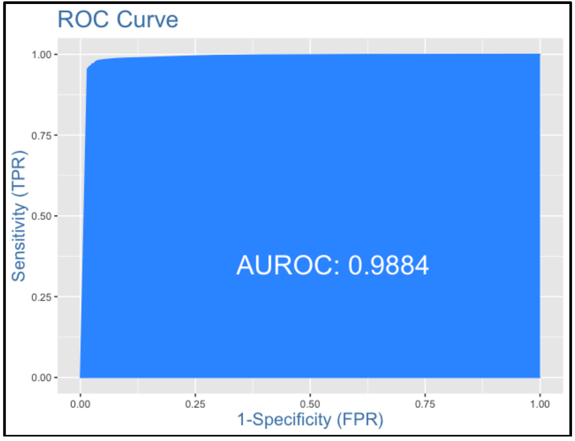


fig. 15: model 3 ROC curve

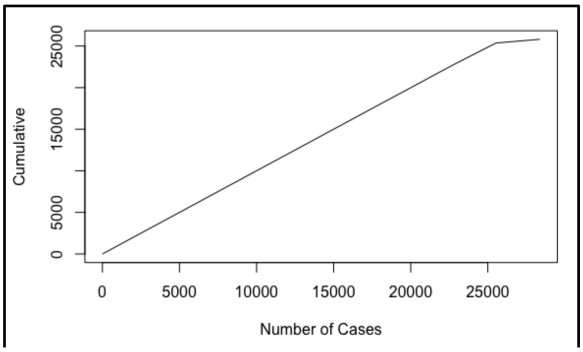


fig. 16: lift chart for all models

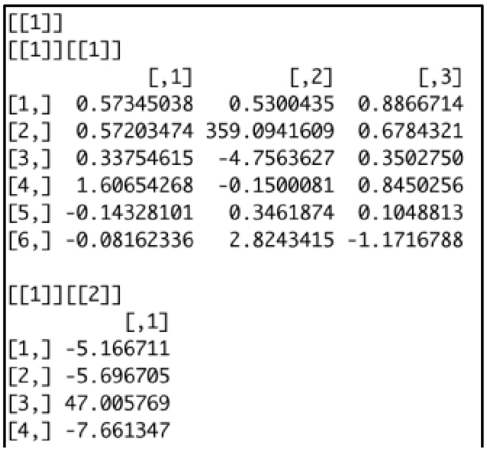
4.3 Neural Network

A neural network is an artificial intelligence technique that instructs computers to analyze data in a manner modeled after the human brain. It is a kind of artificial intelligence technique known as deep learning that makes use of interconnected neurons or nodes in a layered structure to mimic the human brain. To further analyze the data, I developed a Neural Network model on the dataset. Results and observations are described in the following paragraphs.

Similarly, in this analysis, the dataset was partitioned in a 60% by 40% ratio with 60% of the data assigned to training dataset (TD) and 40% assigned to validation dataset (VD).

In order to compare results, I built three models (model 1, model 2, and model 3) with the combination of regressors as in the Logistic Regression approach. The images below show the weights of the regressors for each model built. The preferred model (i.e. model 1) has the probability formula written as thus:

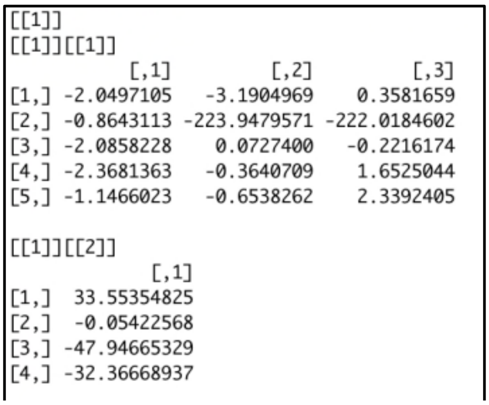
$$P = \frac{1}{1+e^{-(0.57+0.57DOsat\ +0.33pH\ +1.60Cond\ -0.14Dept\ h-0.08Temp\ )}}$$



```
[[1]]
[[1]][[1]]
            [,1]        [,2]       [,3]
[1,]  0.57345038   0.5300435  0.8866714
[2,]  0.57203474 359.0941609  0.6784321
[3,]  0.33754615  -4.7563627  0.3502750
[4,]  1.60654268  -0.1500081  0.8450256
[5,] -0.14328101   0.3461874  0.1048813
[6,] -0.08162336   2.8243415 -1.1716788

[[1]][[2]]
          [,1]
[1,] -5.166711
[2,] -5.696705
[3,] 47.005769
[4,] -7.661347
```

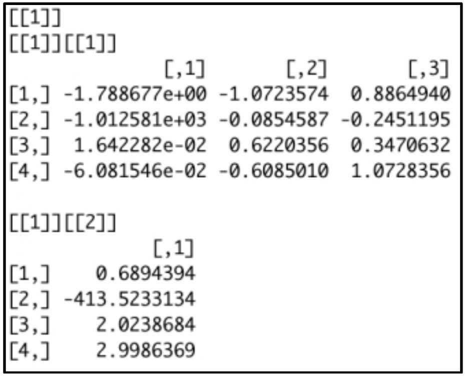
fig. 17: model 1 regressor coefficients.

The coefficients and weights for model 2 and model 3 are as displayed in fig. 18 and fig. 19 below respectively.



```
[[1]]
[[1]][[1]]
           [,1]         [,2]         [,3]
[1,] -2.0497105   -3.1904969    0.3581659
[2,] -0.8643113 -223.9479571 -222.0184602
[3,] -2.0858228    0.0727400   -0.2216174
[4,] -2.3681363   -0.3640709    1.6525044
[5,] -1.1466023   -0.6538262    2.3392405

[[1]][[2]]
            [,1]
[1,]  33.55354825
[2,]  -0.05422568
[3,] -47.94665329
[4,] -32.36668937
```

fig. 18: model 2 regressor coefficients.



```
[[1]]
[[1]][[1]]
              [,1]       [,2]       [,3]
[1,] -1.788677e+00 -1.0723574  0.8864940
[2,] -1.012581e+03 -0.0854587 -0.2451195
[3,]  1.642282e-02  0.6220356  0.3470632
[4,] -6.081546e-02 -0.6085010  1.0728356

[[1]][[2]]
             [,1]
[1,]    0.6894394
[2,] -413.5233134
[3,]    2.0238684
[4,]    2.9986369
```

fig. 19: model 3 regressor coefficients.

The outcomes of the neural network models stated above are shown in Table 4 below. Results of the training and validation classification data on model 1, as well as model 2 and model 3 gave a new perfect result in terms of Misclassification Error, True Positive Rate (TPR, a.k.a. Sensitivity), False Positive Rate (FPR), Specificity, and Accuracy.

The model with all regressors (i.e. model 1) had an overall **Misclassification Error** of 0.3% compared to **1.8%** for Logistic Regression, **Accuracy** of **99.7%** compared to **98.2%** for Logistic Regression, **Sensitivity** of **100%** compared to **98.8%** for Logistic Regression**,** and **Specificity** of **99.7%** compared to **91.8%** for Logistic Regression**,** it is obvious the neural Network models performed better. With near **100%** performance in terms of specificity, sensitivity, and accuracy, this shows that the model is quite useful in correctly classifying the true positives, true negatives, and resulting in more correct predictions respectively. It is evident that the neural network machine learning approach is a more effective way of analyzing the data as a much better result is realized using this machine learning approach.

MODEL	OPTIMAL CUTOFF	MISCLASSIFICATION ERROR	TPR	FPR	SPECIFICITY	ACCURACY
MODEL 1 TD	0.32	0.003	1	0.033437	0.997	0.997
MODEL 1,2,3 VD	0.33	0.0027	0.999942	0.029002	0.997	0.997
MODEL 2	0.37	0.003	1	0.033437	0.997	0.997
MODEL 3	0.37	0.003	1	0.033437	0.997	0.997

table 4: neural network model performances

Evaluating the confusion matrix data, there was zero case of false positives and eighty-six cases of false negatives. This is a negligible portion of the entire training data set. Identical performance was observed in both model 2 and model 3 as well.

CONFUSION MATRIX MODEL 1 TD		
	0	**1**
0	2486	0
1	86	25803

CONFUSION MATRIX MODEL 2 TD		
	0	**1**
0	2486	0
1	86	25803

CONFUSION MATRIX MODEL 1,2,3 VD

CONFUSION MATRIX MODEL 3 TD

	0	1
0	1674	1
1	50	17192

	0	1
0	2486	0
1	86	25803

table 5: confusion matrix report for all neural network models

Assessing the ROC curve and lift charts, it was observed that when all the available regressors were featured in the model, as well as other models, it appeared that the model is a bit less fitting with an AUC of **97%** when compared to model 2 and 3, although there is a significant lift in the cumulative actual of the predicted values. Model 2 and model 3 had AUC scores of **98%**. For a good model, the AUC is aimed to be closer to **100%** (the optimum classifier) and a lift chart showing evident lift in the lift chart curve. Details are in the figures below.

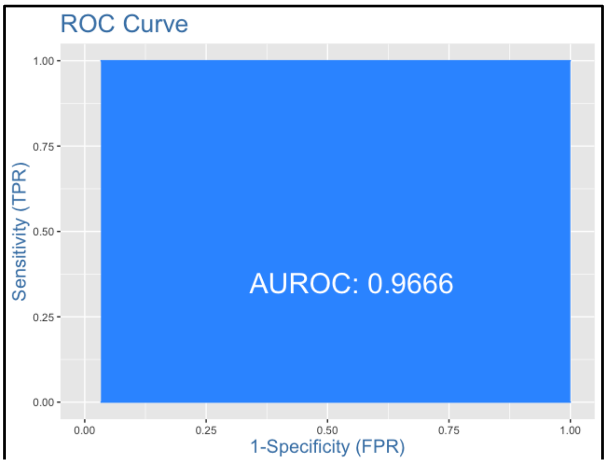


fig. 20: neural network model 1 ROC curve

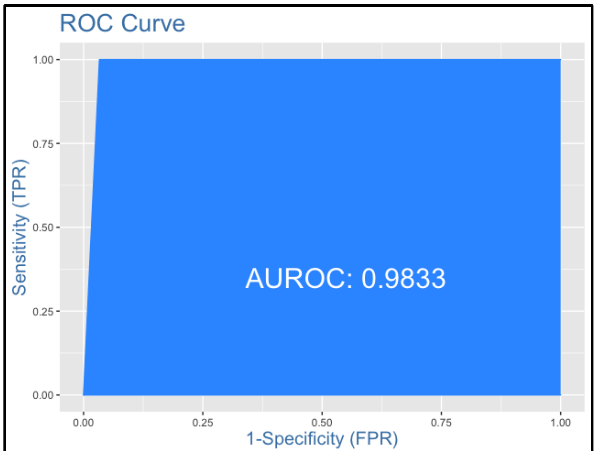


fig. 21: neural network model 2 and 3 ROC curve

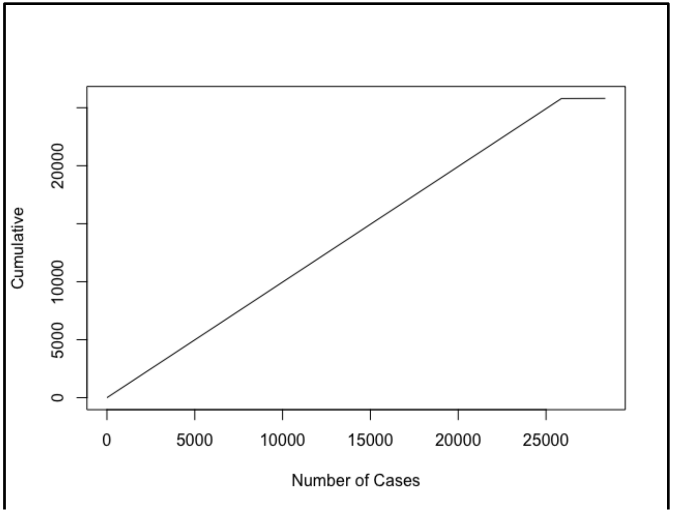


fig. 22: lift chart for all neural network models

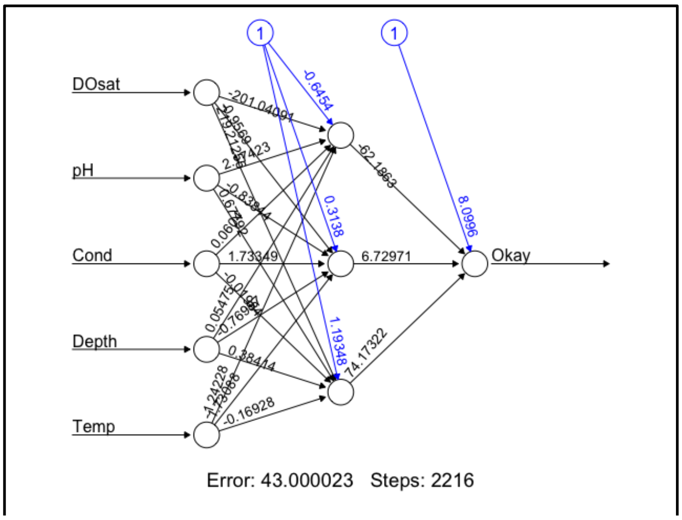


fig. 23: neural network plot for model 1

4.4 Classification Tree, Random Forest, Bagging and Boosting

As is with Neural Networks and Logistic Regression models, Classification Tree, Random Forest, Bagging, and Boosting are machine learning techniques for classification purposes. Overall, the Neural Network approach has produced the best (classification) result thus far, hence I would be comparing the results from the Neural Network with that of the other approaches enumerated above in this section of the report.

Likewise, in this analysis, the dataset was partitioned in a 60% by 40% ratio with 60% of the data assigned to training dataset and 40% assigned to validation dataset. To compare results, I evaluated the confusion matrix report from the Classification Tree, Random Forest, Bagging, and Boosting techniques, as well as results from the Neural Network report. All regressors were selected for all the techniques. The Neural Network report was carried over to this section because it has been selected to be the best classification method thus far.

Table 6 below shows the results of the different classification techniques used in this section. The best performing approach was identified to be the Classification Tree method. The Classification Tree method resulted in a perfect performance in terms of Misclassification Error of **0%**, True Positive Rate (TPR, a.k.a. Sensitivity) of **100%**, False Positive Rate (FPR) of **0%**, Specificity of **100%**, and Accuracy of **100%**. This shows that the model is quite useful in correctly classifying the true positives (specificity), true negatives (sensitivity), and resulting in more correct predictions (accuracy).

This was closely followed by the Bagging technique, then Random Forest, Boosting, and finally the Neural Network process.

MODEL	MISCLASSIFICATION ERROR	TPR	FPR	SPECIFICITY	ACCURACY
NEURAL NETWORK	0.003031	1.0	0.033437	0.997000	0.996969
CLASSIFICATION TREE	**0.0**	**1.0**	**0.0**	**1.0**	**1.0**
RANDOM FOREST	0.000070	0.999961	0.000387	0.999613	0.999930
BAGGING	0.000035	1.0	0.000387	1.0	0.999965
BOOSTING	0.002890	1.0	0.031697	1.0	0.997110

table 6: results of the different classification techniques

The data in table 6 above were all generated from the details in the confusion matrix for each technique. The Classification Tree approach had zero case of false positives and false negatives. The Random Forest, Bagging, and Boosting techniques also had very minute misclassifications. Details of the confusion matrix are in table 7 below.

CLASSIFICATION TREE		
	0	**1**
0	2587	0
1	0	25788

RANDOM FOREST		
	0	**1**
0	2586	1
1	1	25787

BOOSTING		
	0	**1**
0	2505	0
1	82	25788

BAGGING		
	0	**1**
0	2586	0
1	1	25788

table 7: confusion matrix for the different classification techniques

In terms of variable importance, DOsat was observed to have the most correlation with the level of dissolved oxygen (DO) as in figure 24 below. It is a replica of the illustration in figure 3 above.

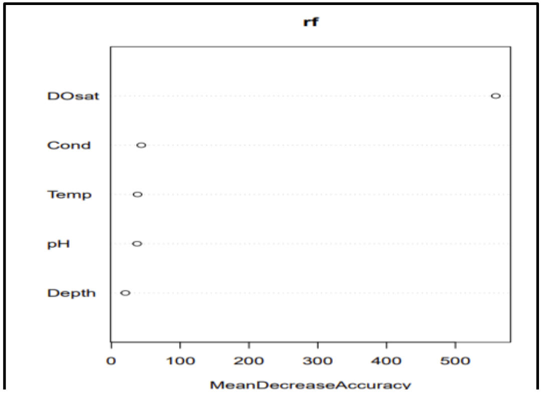


fig. 24: random forest variable importance

The image below is a snapshot of the decision tree. The tree started with a root node of DOsat. DOsat being the most positively correlated variable to DO. The branches help carry out the classification in terms of the result of the node. A leaf without a subsequent branch ends the process for any specific node. The figure below basically explains how the classifications are done.

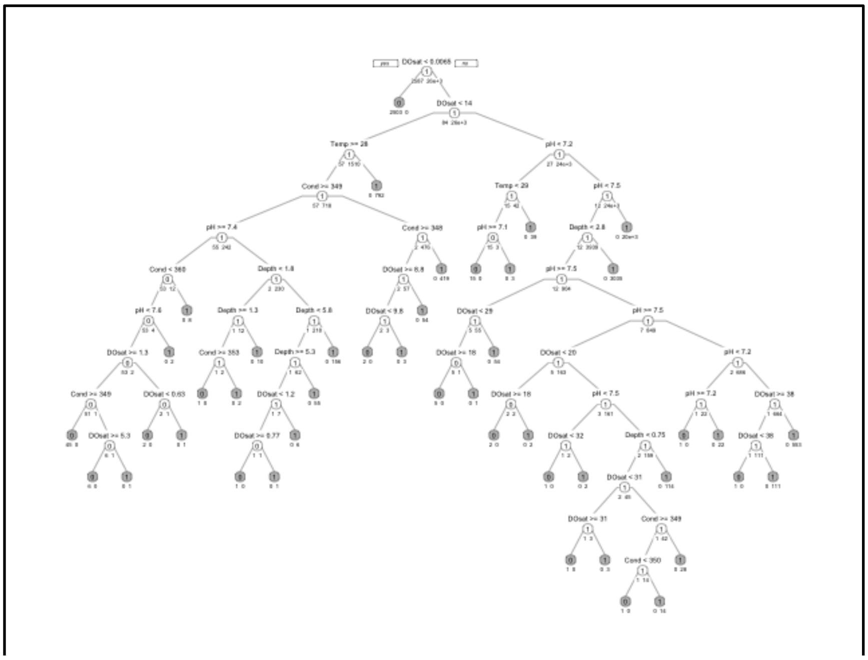


fig. 25: decision tree

5.0 Summary and Conclusion

With respect to a linear regression analysis, this dataset is often quite helpful for analysis purposes. It functioned admirably in every way. The model and the data both fit together well. Regarding the findings from metrics like RMSE, MAE, BIC, R squared, Adjusted R squared as detailed above, and other analytical concepts, this model satisfies every condition for a good model. It is further proven that the dataset fits the model effectively by examining the residuals. The linear regression formula above implies that a unit increase in *DOsat*, and *Cond* would cause the level of dissolved oxygen to increase by 0.0779mg/L and 0.00156 mg/L respectively. On the other hand, a unit increase in *pH*, *Depth*, and *Temp* would cause the level of dissolved oxygen to reduce by 0.0479 mg/L, 0.00856 mg/L, and 0.108 mg/L respectively.

Furthermore, as it relates to logistic regression, for model 1, all regressors are significant, although 'Depth' is only significant at 77% significance level. pH, Conductivity, and Temperature have negative effects on Dissolved Oxygen (DO) whilst DO Saturated and Depth have positive effects. The model's performance in terms of specificity, sensitivity, and accuracy is almost perfect, proving its worth in terms of correctly classifying true positives, true negatives, and making more accurate predictions, respectively. Also, evaluating the ROC curve and lift charts, it was observed that when all the available regressors were featured in the model, as well as other models, it appeared that the models are quite fitting with an AUC of 99% and a significant lift in the cumulative actual of the predicted values.

Also, a neural network is an artificial intelligence method that gives instructions to computers on how to interpret data in a way that is similar to the way the human brain does. Results from the neural network generally produced a better performance than the logistic regression. The model's performance in terms of specificity, sensitivity, and accuracy is perfect, proving its worth in terms of correctly classifying true positives, true negatives, and making more accurate predictions, respectively. The lift chart and AUC scores also further affirm this.

Additionally, as is with Neural Networks and Logistic Regression models, Classification Tree, Random Forest, Bagging, and Boosting are machine learning techniques for classification purposes. The best performing approach was identified to be the Classification Tree method. The Classification Tree method resulted in a perfect performance in terms of Misclassification Error of **0%**, True Positive Rate (TPR, a.k.a. Sensitivity) of **100%**, False Positive Rate (FPR) of **0%**, Specificity of **100%**, and Accuracy of **100%**. This shows that the model is quite useful in correctly classifying the true positives (specificity), true negatives (sensitivity), and resulting in more correct predictions (accuracy). The Classification Tree approach had zero case of false positives and false negatives. The Random Forest, Bagging, and Boosting techniques also had very minute misclassifications. Overall, it has been observed that the Classification Tree approach is the best machine learning technique for this dataset.

References

Durell, L., Scott, J. T., Nychka, D., &Hering, A. S. (2022). Functional forecasting of dissolved oxygen in high- frequency vertical lake profiles. *Environmetrics*. https://doi.org/10.1002/env.2765

Rajesh, M., &Rehana, S. (2022). Impact of climate change on river water temperature and dissolved oxygen: Indian riverine thermal regimes. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-12996-7

Beck, A. J., &Sañudo-Wilhelmy, S. A. (2007).Impact of water temperature and dissolved oxygen on copper cycling in an urban estuary.*Environmental Science & Technology*, *41*(17), 6103–6108. https://doi.org/10.1021/es062719y

Kannel, P. R., Lee, S., Lee, Y.-S., Kanel, S. R., & Khan, S. P. (2007).Application of water quality indices and dissolved oxygen as indicators for river water classification and urban impact assessment.*Environmental Monitoring and Assessment*, *132*(1-3), 93–110. https://doi.org/10.1007/s10661-006-9505-1

Chen, J. (2022). Skewness: Positively and negatively skewed defined with formula. Investopedia.Retrieved from https://www.investopedia.com/terms/s/skewness.asp#:~:text=The%20mean%20of%20positively%20skewed,or%20fat%20the%20tails%20are.

Zach. (2021). How to calculate BIC in R. Statology.Retrieved from https://www.statology.org/bic-in-r/#:~:text=The%20Bayesian%20Information%20Criterion%2C%20often,that%20best%20fits%20the%20data.

Roy, S. (2020). Re: How important are normal residuals for regression? Retrieved from: https://www.researchgate.net/post/How-important-are-normal-residuals-for-regression/5e133405a4714bab0e3096d2/citation/download.