n10463011

**MXN442 Report**

**Link to Github: https://github.com/EelizaD/MXN442-Assessment-3-n10463011**

**Title of paper chosen:** Short-term Lake Erie algal bloom prediction by classification and regression models

**Objective of the study:** Toxic algal blooms are an increasing epidemic which pose threats to human, animal, and ecosystem health. As such, researchers have focused their efforts on understanding the motivators of the toxic algal blooms and predicting the intensity of future blooms. In particular, machine learning has been used to predict algal bloom intensity. Ai et al. (2023) culminated data from the Western Lake Erie Basin, located in the United States of America, to address the following three objectives:

1. Identify the most important features involved in toxic algal blooms in Lake Erie.
2. Determine the best machine learning algorithm to predict bloom intensity in Lake Erie.
3. Predict short-term input features using the machine learning algorithm long short-term memory (LSTM).

**Methods of the study:**

Objective one:

For the first objective of the study, initially a literature review was performed to identify all possible environmental factors impacting the algal blooms in the Western Lake Erie Basin. After identifying all the possible features impacting the intensity of toxic algal blooms in the Western Lake Erie Basin, model reduction was performed to determine the most important factors. The deterioration factor was used to identify the least important factors in the model. To determine the deterioration factor, one variable must be slightly changed, the effect this change has on the model's predictive abilities is the deterioration factor. If the model greatly deteriorates as a result of the change, the variable is considered important, whereas little deterioration indicates an unimportant variable. After the removal of an unimportant variable, the model was fit to the remaining factors and its performance assessed to ensure it still performed adequately. This process of identifying the least important factor and subsequently removing it was repeated until the further removal of parameters inhibited model performance.

Objective two:

For the second objective of the study, nine different machine learning algorithms were used and compared. The nine algorithms consisted of the artificial neural network (ANN), bootstrap aggregating, gradient boosting, gaussian processes, k-nearest neighbour, long short-term memory, random forests, support vector machines, and XGBoost. The collated data included the chlorophyll-a index (CI), which was used to categorise the blooms into four different intensity levels (RF4); light, mild, significant, extreme. Due to the unbalanced nature of the data, two different classifications were also considered, one that split the data into three intensity levels (RF3), and another that split the data into two intensity levels (RF2). For the RF3 levels, the lowest level consisted of the light classification from RF4, the middle level consisted of the mild classification from RF4, and the highest level consisted of the significant and extreme classifications from RF4. For the RF2 levels, the lowest level consisted of the light and mild classifications from RF4, and the second level consisted of the significant and extreme classifications from RF4. Hence, three classification models were considered, one for each data split.

For the implementation of the nine machine learning algorithms, a variety of algorithm scaler values were used. For instance, a learning rate ranging from $10^{-6}$ to $10^{-2}$ was considered for ANN in combination with layers of 1, 2, and 3, nodes stepping from 20 to 200, in steps of 1, and activation functions including 'relu', 'selu', 'sigmoid', and 'linear'. Alternatively, the bootstrap aggregating was considered with 10, 20, 50, 80, 100, 200, 500, and 1000 trees. The gradient boosting algorithm considered the same tree variations and also considered a learning rate of 0.05, 0.1, 0.5, and 1, and maximum depth of 5, 10, 20, 50, 80, and 100. To identify the best performing algorithm, the accuracy of the classification predictions was considered.

Objective three:

The third objective of the study focused on predicting some features that are not available in real time using the long short-term memory (LSTM) algorithm. The LSTM algorithm predicts a new state based on the previous state and the potential state. Initially, a forget gate value, $f_t$, between 0 and 1 is calculated to remove unnecessary information,

$$f_t = \sigma\left(W_f h_{t-1} + U_f I_t + b_f\right),$$

Where $W$ is the weight for gates at different states, $U$ is the weight for the cell at different states, $b$ is the learnable bias, $h_{t-1}$ is the previous output, $I_t$ is the new input, and $\sigma$ is the sigmoid function,

$$\sigma(t) = \frac{1}{1 + e^{-t}}.$$

From this, the potential cell state is calculated,

$$\widetilde{C_t} = \tanh(W_c h_{t-1} + U_c I_t + b_c).$$

Additionally, the feature for $\widetilde{C_t}$ is calculated,

$$i_t = \sigma(W_i h_{t-1} + U_i I_t + b_i).$$

The previous cell state and potential cell state is combined to calculate the new cell state,

$$C_t = f_t C_{t-1} + i_t \widetilde{C_t}.$$

Then, a coefficient for the cell state must be obtained,

$$o_t = \sigma(W_o h_{t-1} + U_o I_t + b_o).$$

Finally, this can be combined to determine the output,

$$h_t = o_t \tanh(C_t).$$

A key component of the third objective was to feed the predicted values from the long short-term memory algorithm into the algal bloom intensity prediction model. For the algal bloom model to maintain accuracy, the long short-term memory algorithm would also need to be accurate. To assess the LSTM algorithm, the $R^2$ value was considered.

**Results of the study:**

For the first objective of the study, the literature review to identify any possible influential yielded over 50 results. The model reduction found that only eight features were key in the prediction of the algal bloom intensity. The eight features fell into four different subcategories, physiochemical, temporal, hydrodynamic, and meteorological. In the physiochemical subcategory, the important factors were 5TN, N30, SSRP, and STN. The factor 5TN is the 5-year cumulative total nitrogen load, N30 is the total nitrogen, SSRP is the spring soluble reactive phosphorus loading, and STN is the spring total nitrogen loading. The temporal factor was the time of year, denoted as time period. The hydrodynamic subcategory included WLM10 and WLD20, where WLM10 is the water level near the Maumee River, and WLD20 is the water level near the Detroit River. Finally, the important meteorological factor was the solar irradiance, SOL30.

For the second objective, multiple performance metrics were considered to determine the best algorithm for predicting the intensity of algal blooms in the Western Lake Erie Basin. The classification models were evaluated by their accuracy in correctly predicting the classification of a bloom. On the other hand, the regression models were assessed using their $R^2$ value and root mean-squared error (RMSE). As seen in Table 1 which provides a summary of the performance metrics for the nine algorithms considered, the random forests algorithm performed the best.

*Table 1: Performance metrics for the nine machine learning algorithms considered. The bolded line, RF (random forests), was identified as the best performing algorithm.*

| | **Classification model** | | | **Regression model** | |
|---|---|---|---|---|---|
| | RF2 | RF3 | RF4 | $R^2$ | RMSE |
| **ANN** | 87.5% | 68.8% | 58.3% | 0.16 | 6.5 |
| **BA** | 87.7% | 76.5% | 71.4% | 0.50 | 5.3 |
| **GB** | 87.4% | 73.5% | 68.4% | 0.49 | 5.4 |
| **GP** | 77.6% | 61.3% | 61.3% | -0.69 | 9.8 |
| **KNN** | 84.9% | 70.9% | 64.8% | 0.39 | 5.9 |
| **RF** | **88.9%** | **77.2%** | **70.7%** | **0.52** | **5.3** |
| **SVM** | 81.0% | 62.9% | 59.7% | -0.13 | 8.1 |
| **XGB** | 86.9% | 76.1% | 70.6% | 0.39 | 5.8 |
| **LSTM** | 52.3% | 47.3% | 48.4% | -0.07 | 9.56 |

For the final objective, only short-term factors that were deemed important in objective one were considered using the long short-term memory algorithm. Hence, solar irradiance (SOL10), water level near the Detroit River (WLD10), water level near the Maumee River (WLM10), and the total nitrogen (N10) were considered. Three of these factors performed well, with $R^2$ values of 0.75, 0.92, and 0.89 for the parameters SOL10, WLD10, and WLM10 respectively. The parameter N10 performed very poorly, with an $R^2$ of 0.22.

**Aims of my work:**

The code for the research is unavailable, however, the data used is available, hence, I have attempted to replicate the results of the paper. However, I had a limited amount of time to complete the replication, so chose to reduce the scope of the objectives I addressed. Rather than address all three objectives, I focused on the first two objectives of determining the most important features and the best algorithm for predicting the algal bloom intensity in the Western Lake Erie Basin. More specifically, rather than consider all nine algorithms, a subset of three algorithms was considered which covered neural network algorithms and tree-based algorithms. The three algorithms considered was the k-nearest neighbour (KNN), random forests, and artificial neural network (ANN). By choosing this subset, a variety of algorithm types could be considered while maintaining an achievable project scope. For the first objective, the same features and data was considered, meaning no literature review was required to identify possible important features. Additionally, regardless of which algorithm I find to perform best, the model reduction will be performed on the random forests model for two reasons. The first reason is to directly compare my results with the results from the paper, and the second reason is so that I can use the same method of model reduction.

**Results of my work:**

The k-nearest neighbour (KNN) classification model was considered with k values of 3, 5, 10, 20, 30, as was used in Ai et al. (2023). A confusion matrix was created for each k value, and the results averaged to determine the accuracy and kappa coefficients for the RF2, RF3, and RF4 models. For the RF2 model, an average accuracy of 83.33% was obtained, and an average kappa coefficient of 0.5097. The RF3 model achieved an average accuracy of 72.92% and average kappa coefficient of 0.4765. Finally, the RF4 model had an average accuracy of 65.42% and average kappa coefficient of 0.3526. In comparison to the results from Ai et al. (2023), the accuracy remains relatively similar, with the RF2 model being most accurate, followed by the RF3 model and the RF4 model being least accurate. Figures 1 – 3 provide plots of the accuracy of the model at each k value. In all cases, there is a general trend of the accuracy decreasing as the k value increases. However, a k value of 5 resulted in the highest accuracy for the RF2 model, whereas a k value of 3 yielded the highest accuracy for the RF3 and RF4 models.

The number of trees considered in the random forest (RF) classification model ranged from 10 to 100, increasing by 10, as in Ai et al. (2023). Similar to the KNN model, a confusion matrix was created to assess the accuracy of each model. The overall accuracy for the RF4, RF3, and RF2 model was 66.73%, 78.16%, and 88.16% respectively. The level of accuracy increased as the number of classification levels reduced, which was also seen using the KNN model, and also in Ai et al. (2023). Unlike the KNN classification model, the accuracy of the RF model remained similar to the results found in Ai et al. (2023). Figures 4 – 6 plot the accuracy of each model against the number of trees used. Unlike the KNN models, there is no clear pattern associating the accuracy with the number of trees. For the RF2 and RF3 models, the optimal number of trees is around 60, whereas the RF4 model reached the optimal accuracy at around 30 trees.

The artificial neural network considered various learning rates, nodes, and layers. As per Ai et al. (2023), the learning rate ranged from $10^{-6}$ to $10^{-2}$, the number of nodes ranged from 20 to 200, increasing in steps of 20, and 1, 2, and 3 layers were considered. Similar to previous results, the RF4 model performed the worst, with an accuracy of 63.64%, followed by the RF3 model with an accuracy of 75.51%, and finally the RF2 model with an accuracy of 87.14%. In comparison to the results of Ai et al. (2023), the RF2 model performed very similarly, however, the RF3 and RF4 model performed better in the replication than in the original study.

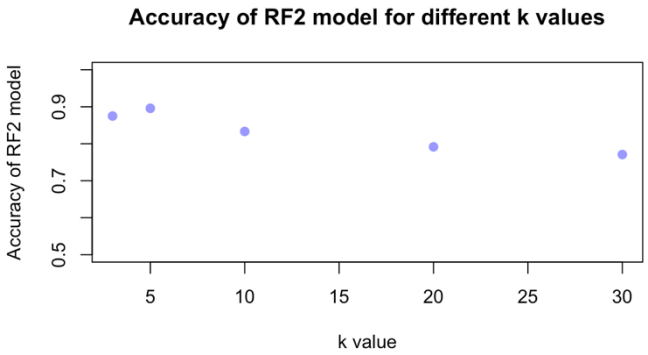


*Figure 1: Plot of RF2 model accuracy against k values. The k values used were 3, 5, 10, 20, and 30. It can be seen that a k of 3 obtained the highest accuracy, whereas a k of 20 resulted in the lowest accuracy.*

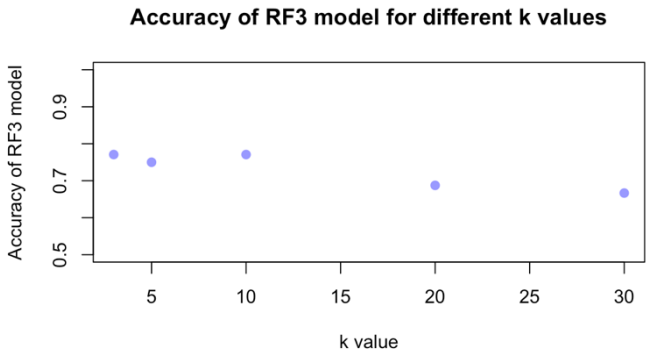


*Figure 2: Plot of RF3 model accuracy against k values. The k values used were 3, 5, 10, 20, and 30. It can be seen that a k of 3 obtained the highest accuracy, whereas a k of 20 resulted in the lowest accuracy.*

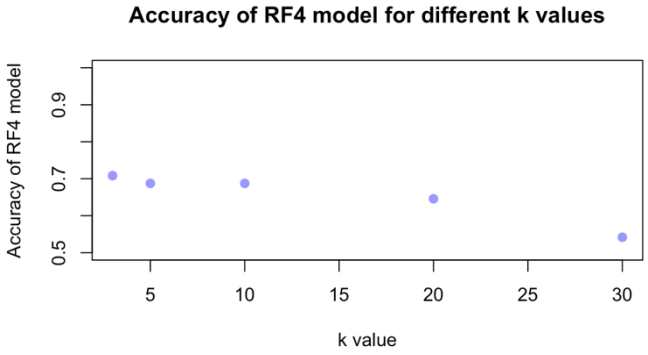


*Figure 3: Plot of RF4 model accuracy against k values. The k values used were 3, 5, 10, 20, and 30. It can be seen that a k of 3 obtained the highest accuracy, whereas a k of 20 and 30 resulted in the lowest accuracies.*

Model reduction was performed on the random forest 4, 3, and 2 level models. The RF4 model was reduced to seven most important factors, with an accuracy of 68.37%. The most important factors identified were SOL20, 9SRP, TN3-6, time period, WLD20, WLD30, and EWIND30. In comparison to the results from Ai et al. (2023), three of the identified most important factors matched. Specifically, solar irradiance (SOL20), time period, and water level near Detroit River (WLD30). The 9-year cumulative soluble reactive phosphorus (9SRP), total nitrogen (TN3-6), and easterly and westerly wind speed (WIND30) was also identified as being the most important factors for the RF4 model. The RF3 model was also reduced to seven most important factors, with an accuracy of 76.33% The most important factors included Q30, SOL20, SOL30, time period, 5TN, WLD20, and WLD30. Of these, the solar irradiance (SOL20), time period, 5-year cumulative total nitrogen (5TN), and water level near the Detroit River (WLD30) were also identified as the most important factors in the RF3 models in Ai et al. (2023). Finally, the RF2 model was reduced to using only five variables. However, at the

point of three variables, the accuracy was still acceptable, at 87.35%. It was chosen to keep five variables as further reduction seemed unreasonable despite the maintained accuracy. The five most important features identified were Q30, time period, 5TN, WLD10, and WLD30. In Ai et al. (2023), the 5-year cumulative total nitrogen (5TN), time period, and water level near Detroit River (WLD10), were identified as important features. However, Ai et al. (2023), did not identify the flowrate (Q30) as an important feature.

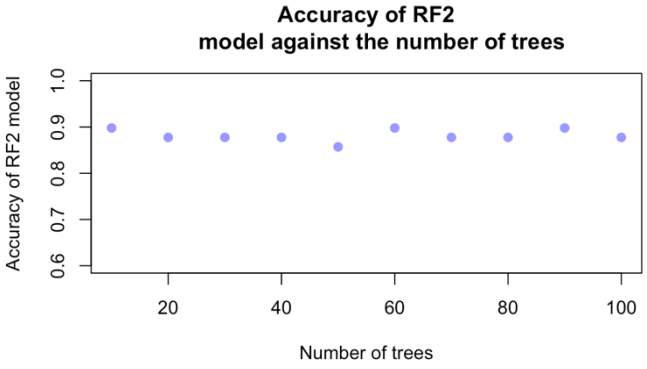


*Figure 4: Plot of RF2 model accuracy against number of trees. The number of trees ranged from 10 to 100, increasing by 10. The accuracy remains relatively constant throughout, however, 60 trees appears to result in the highest accuracy.*

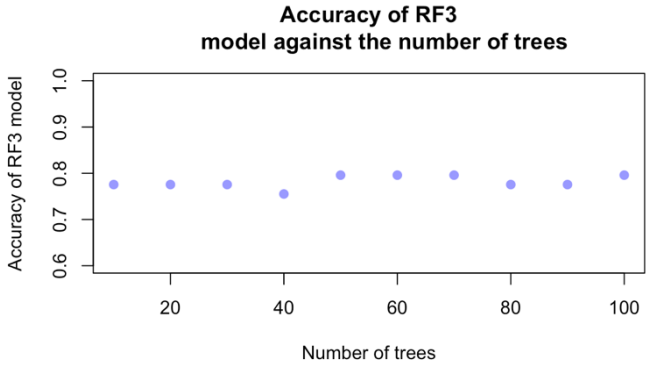


*Figure 5: Plot of RF3 model accuracy against number of trees. The number of trees ranged from 10 to 100, increasing by 10. The accuracy remains relatively constant throughout, however, 40-60 trees appears to result in the highest accuracy.*

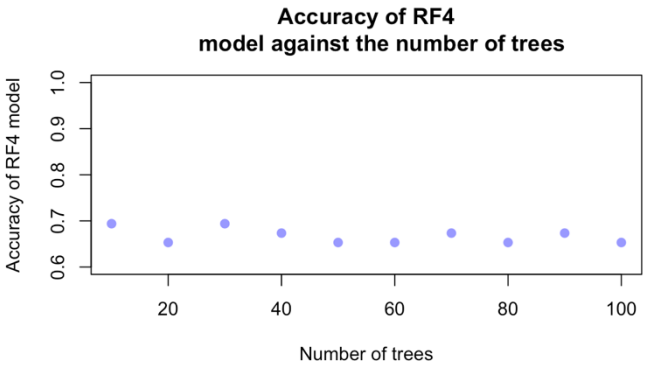


*Figure 6: Plot of RF4 model accuracy against number of trees. The number of trees ranged from 10 to 100, increasing by 10. The accuracy remains relatively constant throughout, however, 30 trees appears to result in the highest accuracy.*

Considering the three algorithms used for the classification model, the random forests models had the highest accuracy, which matches the result of Ai et al. (2023). The model reduction results differed from the original study. The RF4 and RF3 models identified seven factors to be most important, with three and four of these factors matching the identified factors in Ai et al. (2023). The RF2 model was reduced to five most important factors, however, further reduction was possible without the loss of accuracy. Of these five most important factors, three were also identified as most important factors in Ai et al. (2023). Overall, the results of identifying the best performing algorithm remained very similar to that of Ai et al. (2023), however, the results of the model reduction differed significantly more.

*Table 2: Accuracy of ANN, KNN, and RF algorithms for the RF2, RF3, and RF4 models.*

| | RF2 | RF3 | RF4 |
|---|---|---|---|
| **ANN** | 87.14% | 75.51% | 63.64% |
| KNN | 83.33% | 72.92% | 65.42% |
| **RF** | 88.16% | 78.16% | 66.73% |

## References

Ai, H., Zhang, K., Sun, J., & Zhang, H. (2023). Short-term Lake Erie algal bloom prediction by classification and regression models. *Water Research, 232*. https://doi.org/10.1016/j.watres.2023.119710