Predictive Analytics for Algae Growth

Predictive Analytics Assignment 1

Thomas Bridgeman
Department of Computing
Atlantic Technological University
Letterkenny, Ireland
L00161227@atu.ie

***Abstract*—Algae are aquatic organisms that photosynthesis. They range from microscopic phytoplankton to large kelp forests and offer both human and marine benefits. However, too much algae of algal blooms destroy ecosystems and drinking supplies. Using predictive models this project aims to predict a way for the population to be regulated to maintain a beneficial level of algae and ensure water sources remain at an optimal level for most efficient algae growth. The methods, model and visualisations will be explored in this report.**

***Index Terms*—Algae, Algal Blooms, Regression, Ensemble**

I. Introduction

Algae is the general term used to refer to aquatic organisms that produce oxygen and energy through photosynthesis. There are vast variations in different species of algae, ranging from large seaweed like kelp to microscopic phytoplankton both of which are some of the most crucial pillars of marine life for shelter for other sea life and food for both sea life and human consumption. See Figure 1.

Fig. 1. Algae [1]

Algae can reside in a multitude of environments including both saltwater and freshwater locations due to the massive variations between different species of algae. Many of the subspecies of algae are used extensively by humans as food, biofuel and medicinal reasons among many others but the most beneficial is the production of oxygen through photosynthesis. Some estimations place algae's oxygen production ranging from over half to seventy per cent of all oxygen produced making it a higher producer than all the trees of the world. While water covers over seventy per cent of the planet's surface, most oxygen replacement programs focus on the reforestation of trees. There are several reasons for this, first, there is still a huge benefit to planting trees as it offers way more incentive than just carrying on the carbon cycle and it is a lot simpler and easier to manage than the algae alternative. However, most importantly too much algae in a location can cause way more problems which would be more detrimental, this is commonly called an algal bloom [2].

Algal bloom is the rapid population explosion of certain species of microalgae which can produce toxins which poison natural marine ecosystems and can make water undrinkable. While an algal bloom can happen naturally, the natural minerals in water and fish usually regulate its growth and it is only with the onslaught of fertilisers and manure which are high in nitrogen and phosphorous content that allows a bloom to occur. See Figure 2.

Fig. 2. Algal Bloom [3]

While human intervention in algae is nothing new and has led to both wonderful and terrible things, a data analytical approach may provide new insight into this complex problem. By examining attributes surrounding water, predictions pertaining to the population of algae could be made which could help maintain a healthy algae level in water for both human and marine life and prevent dangerous blooms.

This research paper aims to examine which feature has the biggest impact on the Algae Population, investigate all relationships present between the population and each of the attributes and finally use this information to predict the algae population using various regression models which will be explored in this paper.

II. Dataset Overview

The 'Research on Algae Growth' [4] dataset is taken from Kaggle and features 8 columns and 9784 rows. Seven of the columns will be taken as features to predict the target population. The seven features include Light(μmol photons/m²/s), Nitrate(mg/L), Phosphate(mg/L), Iron(mg/L), Temperature(°C), pH(1-14 scale) and Carbon Dioxide(mg/L). Initial instinct would place Light as one of the biggest attributes surrounding the population as photosynthesis is a requirement for algae to live. An algal bloom is usually caused by Nitrate and Phosphate content so it would also be expected that these features will affect the high-end populations.

III. Data Exploration

The Algae dataset was first loaded into Google CoLab several measures of central tendency were performed using the Pandas library to ensure data aligned with instinct and get a better understanding of how data may affect the population. Examining the mean, mode and median revealed that they feature similar values for most columns that align with instinct such as pH being around 7 which is the natural pH of water. The standard deviation for all variables is also small when you assume the scale of each item.

The previous measures verified that the data did not need to be cleaned and some visualisations were made to get a better understanding of the data. Figure 3 shows a snippet of the Seaborn Pairplot, this plots a graph for each feature against each other so correlation can be observed. The colour of each dot corresponds to the population in values of 1000. Interestingly, most columns feature no pattern with one another except for light which can be seen to be increasing in population size until a point before it decreases again, which is consistent for all columns. It further adds proof that light and population are correlated.

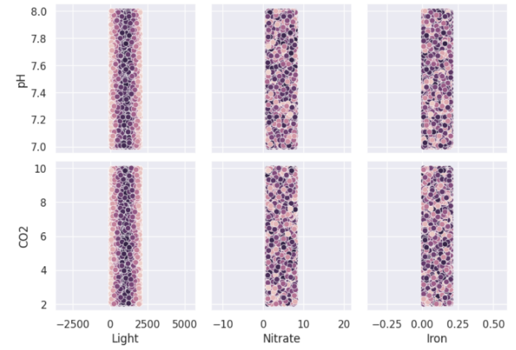
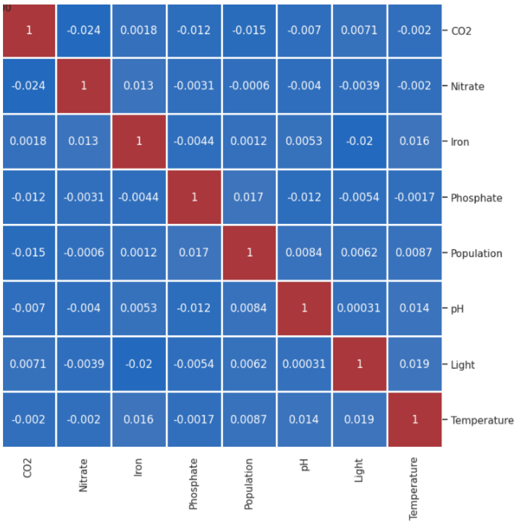


Fig. 3. Seaborn Pairplot Snippet

A correlation map was plotted using the Seaborn Heatmap and Clustermap to further examine the relationship between Light and Population. This however revealed there to be no correlation among any of the features including Light and Population, see Figure 4. Therefore, Figure 5 shows a snippet of a Seaborn PairGrid which showcases how as light increases so does the population until light reaches a measure of roughly 1000 μmol photons/m²/s where it begins to decrease. This would suggest that standard Linear Regression techniques would not work and other models would have to be used to produce a sufficient model for population prediction [5].



	CO2	Nitrate	Iron	Phosphate	Population	pH	Light	Temperature
CO2	1	-0.024	0.0018	-0.012	-0.015	-0.007	0.0071	-0.002
Nitrate	-0.024	1	0.013	-0.0031	-0.0006	-0.004	-0.0039	-0.002
Iron	0.0018	0.013	1	-0.0044	0.0012	0.0053	-0.02	0.016
Phosphate	-0.012	-0.0031	-0.0044	1	0.017	-0.012	-0.0054	-0.0017
Population	-0.015	-0.0006	0.0012	0.017	1	0.0084	0.0062	0.0087
pH	-0.007	-0.004	0.0053	-0.012	0.0084	1	0.00031	0.014
Light	0.0071	-0.0039	-0.02	-0.0054	0.0062	0.00031	1	0.019
Temperature	-0.002	-0.002	0.016	-0.0017	0.0087	0.014	0.019	1

Fig. 4. Seaborn Pairplot Snippet

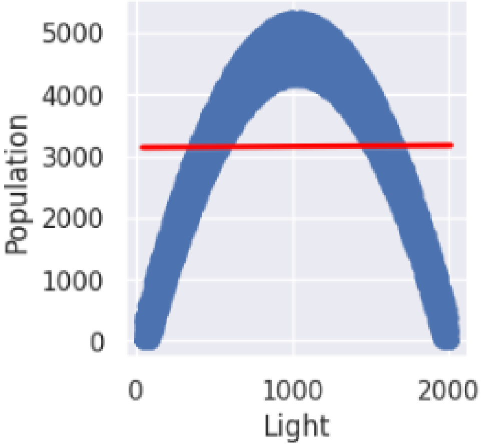


Fig. 5. Seaborn Pairplot Snippet

IV. Regression

Regression models are used for predicting continuous values by finding the relationship between the features and the target variable. There are many types of regression which have different strengths and weaknesses that make them more suited to particular data relationships [6].

A. *Preprocessing and Transformation*

Prior to undergoing work on a regression problem, it is advised to perform some preprocessing and transformation of the data. This can be done through various methods but for this data, Principal Component Analysis or PCA is used. PCA is an unsupervised machine learning technique which reduces dimensionality or quantity of features while minimising loss. The main benefit of performing PCA before regression is to reduce overfitting. The data is also split into training and testing groups in the 60/40 ratio after PCA has been performed.

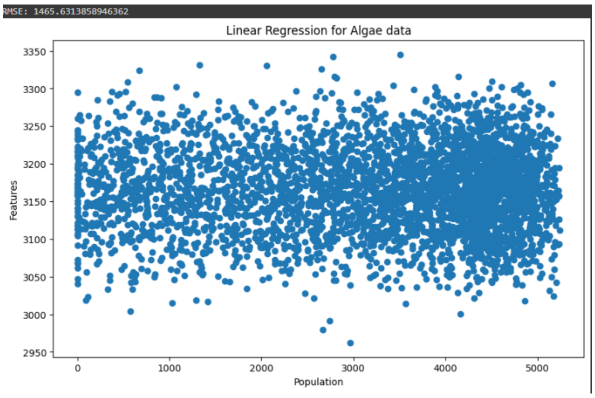
B. Linear, Lasso, Ridge, Elastic Net Regression

Linear Regression is one of the most common forms of regression. It aims to predict the target value based on the value of another variable or group of variables. It relies heavily on the correlation of data points to work effectively but too much correlation can cause overfitting making the model inaccurate when not using the training data. Linear Regression follows the formula:

$$Y_i = f(X_i, \beta) + e_i$$

Where Y_i is the target variable, f is the function, X_i are the feature(s) variables,β are the parameters and e_i are the errors.

Performing linear regression on the Algae data was not successful, see Figure 6. The data does not follow any pattern due to the aforementioned lack of correlation between any of the data. The value of the population does not increase or decrease due to any series of variables. This is further seen when mapping the coefficients to a bar chart, see Figure 7, where it places most of the weight on C02 which was one of the lowest correlated features based on the correlation matrix and Light which appeared to have a bearing on the Population. Examining the metrics like Root Mean Squared Error which judge the performance of a model gives an extremely high score of 1465.63 which should be a value as close to zero as possible.



```
RMSE: 1465.6313858946362
```

Fig. 6. Linear Regression Plot

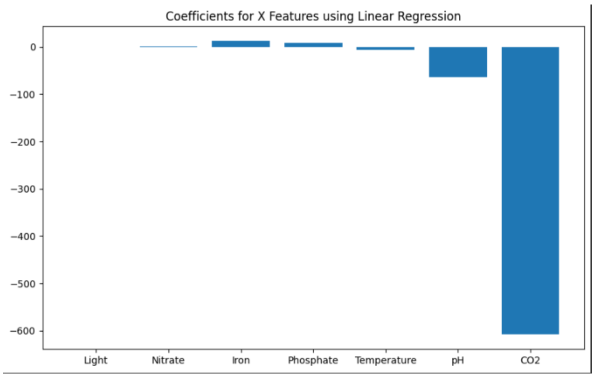


Fig. 7. Linear Regression Coefficient plots

Due to Linear regression's weaknesses, it is rarely used in practice and modified versions are used in its stead. These modified models help reduce overfitting and multicollinearity issues present in data. Lasso Linear Regression or L1 regularisation does this by bringing the associated weights closer to zero if they won't have a significant effect:

$$1/n \sum_{i=1}^{n} (x^i.w - y^i)^2 + \sum_{j=1}^{p} |w^j|$$

By applying the same data as Linear Regression to the L1 model, the RMSE score decreases by one so it technically gets better. Still, the graph of coefficients is practically identical to the previous one. Again, the lack of correlation in the data is showcased. The second modification to Linear Regression is Ridge Linear Regression or L2. Unlike the previous model which aims to push weights to zero, Ridge shrinks all weights to zero and adds a parameter λ_2 multiplied by the norm:

$$1/n \sum_{i=1}^{n} (x^i.w - y^i) + \lambda_2 w^t.w$$

While L2 regression does not affect metrics like RMSE, it does affect the coefficients a lot more than L1, where the weights have been pushed further in the negative direction, further showing the lack of correlation in the data. Combining Lasso and Ridge regression is the most common use case for Linear Regression as it allows the benefits of both to shine through. This is called Elastic Net Regression:

$$1/n \sum_{i=1}^{n} (x^i.w - y^i)^2 + \lambda_1 \sum_{j=1}^{p} |w^j|$$

While this combined formula once again does not feature any improvement to the RMSE score over Lasso, it decreases the range of features, see Figure 8 and massively alters the coefficient graph to be more in line with how you would expect the algae population to be. Affected by temperature and the Nitrate and Phosphate content. Light still has no effect due to the lack of a linear relationship and shows a polynomial relationship may be a much better suit for this regression problem.

C. Polynomial Regression

Polynomial Regression is a form of linear regression that aims to determine a relationship between non-linear features

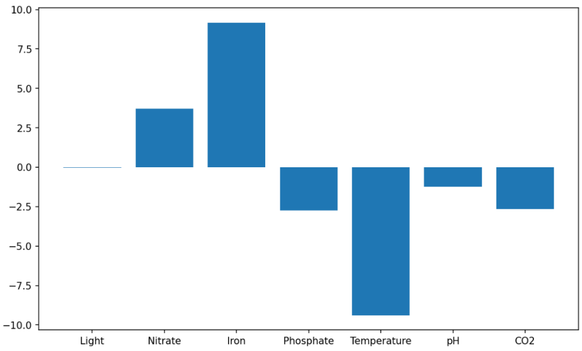
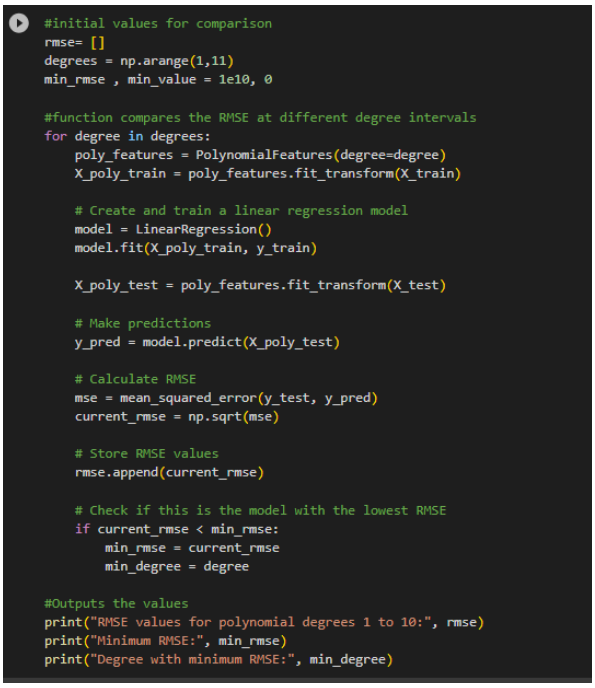


Fig. 8. Elastic Net Coefficient Graph

and a target variable. By using polynomial regression issues of underfitting the data can be solved and make the metrics much better. The polynomial formula takes the linear regression formula and squares, cubes or beyond variables to produce the curve that aims to fit data better.

Performing Polynomial Regression on the algae data, the best polynomial regression degree can be found by looping and testing the model in a loop to find the lowest RMSE value at degrees from 1 to 10 which can be seen in Figure 9 for code and Figure 10 for the graph.



```
#initial values for comparison
rmse= []
degrees = np.arange(1,11)
min_rmse , min_value = 1e10, 0

#function compares the RMSE at different degree intervals
for degree in degrees:
    poly_features = PolynomialFeatures(degree=degree)
    X_poly_train = poly_features.fit_transform(X_train)

    # Create and train a linear regression model
    model = LinearRegression()
    model.fit(X_poly_train, y_train)

    X_poly_test = poly_features.fit_transform(X_test)

    # Make predictions
    y_pred = model.predict(X_poly_test)

    # Calculate RMSE
    mse = mean_squared_error(y_test, y_pred)
    current_rmse = np.sqrt(mse)

    # Store RMSE values
    rmse.append(current_rmse)

    # Check if this is the model with the lowest RMSE
    if current_rmse < min_rmse:
        min_rmse = current_rmse
        min_degree = degree

#Outputs the values
print("RMSE values for polynomial degrees 1 to 10:", rmse)
print("Minimum RMSE:", min_rmse)
print("Degree with minimum RMSE:", min_degree)
```

Fig. 9. Best Poly Degree Code

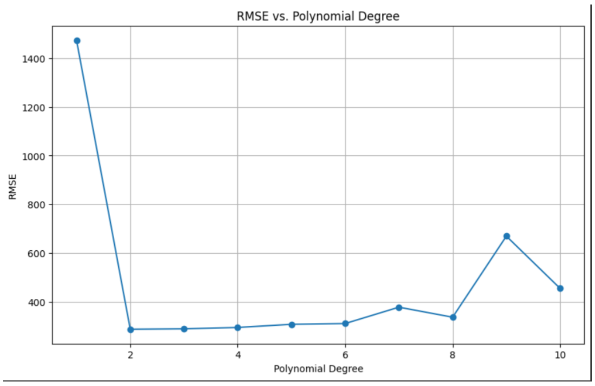


Fig. 10. Best Polynomial Degree Line Graph

A polynomial model as 2 degrees is then applied to the data and it returns an r2 score of 0.96 where an ideal r2-score is 1.0 and a bad score is 0. This is a highly accurate model. By graphing a poly regression line on the light column you can see how the model is achieving a high score for the r2 value, see Figure 11. The predicted values versus the actual values can be graphed and it shows the values aligning accurately, see Figure 12.

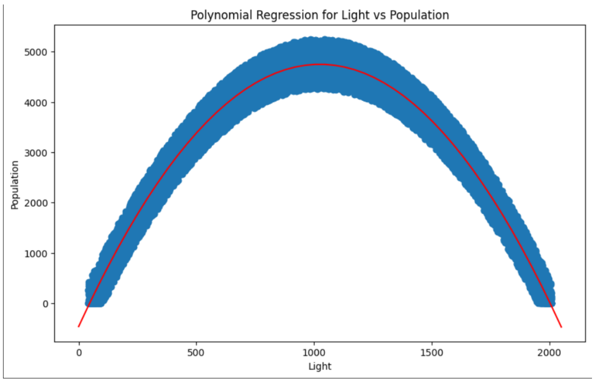


Fig. 11. Polynomial Plot for Light vs Population

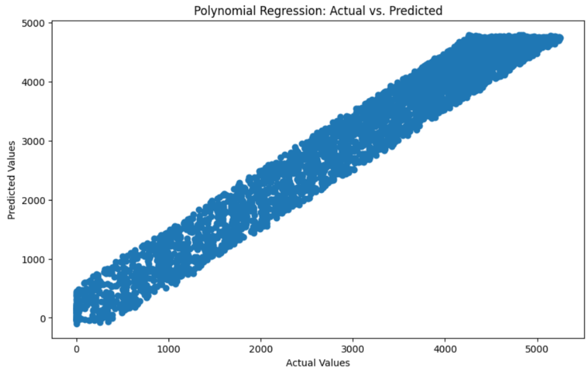


Fig. 12. Scatter Plot of Actual vs Predicted Values for Polynomial Regression

V. Ensemble

Ensemble models is a technique in machine learning that involves the utilisation of multiple models to increase the overall accuracy of a model. This project will use four different models and combine them in a voting regressor model [7].

A. *ExtraTrees Regressor*

ExtraTrees Regressor is a common model in ensemble methods. It is a type of ensemble that makes use of multiple decision trees on various sections of the data and combines the average values to return a higher accuracy than standard decision trees and aims to decrease overfitting.

B. *Random Forest Regressor*

Random Forest Regressor is also a common ensemble method and works in a highly similar way to Extra Trees where it combines multiple decision trees, gets the average to produce a high accuracy and reduces overfitting. Unlike Random Forest it is less random in its selection of cut-off points and it makes use of replica data while ExtraTrees only uses the original samples of data. Random Forest is slightly more computationally expensive.

C. *K Nearest Neighbours*

K Nearest Neighbours aims to find the relationship between a point and the surrounding data points. It works on the principle that similar data points will be located closer together and therefore inferences can made about the data.

D. *Gradient Boosting Regressor*

Gradient Boosting Regression is a powerful tool for finding relationships between non-linear related data which relates well to this data. It is also an ensemble model as it utilises multiple weaker multiples to make the predictions, usually tree-based models.

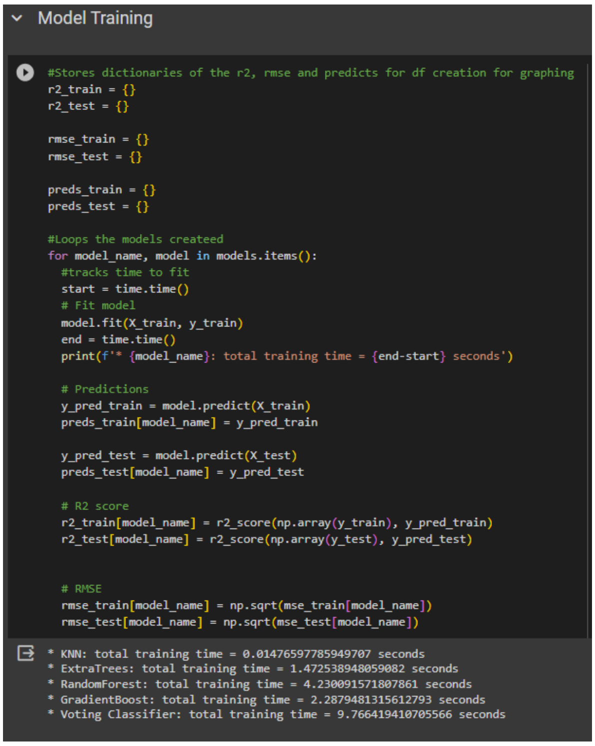
E. *Voting Regressor*

The voting regressor is the main ensemble method, it combines all the specified models into one by averaging the predictions to form the final prediction. The voting regressor in this project uses the four previous models to produce an output that can produce a higher accuracy and minimise the weaknesses of certain models.

F. *Model Comparison*

Each of the models was first created and added to a dictionary. Then each of the models was looped, trained on the training data, and tested for predictions and scores for r2 and RMSE were obtained in terms of both training and testing, see Figure 13 for the code associated.

Upon completion, the dictionaries for the R2-score train and test were combined and sorted with the highest score first. This was then graphed on a double bar chart that shows the comparison between them, see Figure 14. This showed that ExtraTrees was the greatest model as it achieved an r2 score of 100 per cent during training and 96 per cent in testing. The other models achieved similar results of 96 to 99 per cent in training to 95 to 96 in testing. The high results for ExtraTrees could be from overfitting as tree-based models are the most likely to do that but it is quite minimal. Voting Regressor is sat in the middle as it takes the average predictions of the other models to process a more accurate prediction that does not overfit as much as ExtraTrees.



```
#Stores dictionaries of the r2, rmse and predicts for df creation for graphing
r2_train = {}
r2_test = {}

rmse_train = {}
rmse_test = {}

preds_train = {}
preds_test = {}

#Loops the models createed
for model_name, model in models.items():
  #tracks time to fit
  start = time.time()
  # Fit model
  model.fit(X_train, y_train)
  end = time.time()
  print(f'* {model_name}: total training time = {end-start} seconds')

  # Predictions
  y_pred_train = model.predict(X_train)
  preds_train[model_name] = y_pred_train

  y_pred_test = model.predict(X_test)
  preds_test[model_name] = y_pred_test

  # R2 score
  r2_train[model_name] = r2_score(np.array(y_train), y_pred_train)
  r2_test[model_name] = r2_score(np.array(y_test), y_pred_test)


  # RMSE
  rmse_train[model_name] = np.sqrt(mse_train[model_name])
  rmse_test[model_name] = np.sqrt(mse_test[model_name])
```

```
* KNN: total training time = 0.01476597785949707 seconds
* ExtraTrees: total training time = 1.472538948059082 seconds
* RandomForest: total training time = 4.230091571807861 seconds
* GradientBoost: total training time = 2.2879481315612793 seconds
* Voting Classifier: total training time = 9.766419410705566 seconds
```

Fig. 13. Ensemble Model Training and Testing

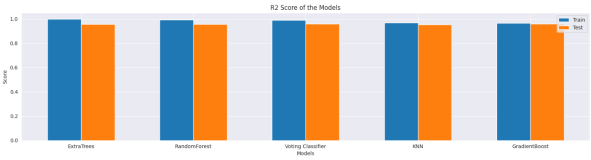


Fig. 14. R2 Score Bar Chart for Ensemble Models

The RMSE dictionaries are combined in the same way sorted with the smallest value at the top as this scenario is more ideal. Figure 15 shows the double bar chart for RMSE. It conveys the same information as the r2 score but in inverse where the lower is better. This is most evident in the training ExtraTrees value being 0 and its testing being 305 as compared to the 100 to 96 in r2. Gradient Boosting Regressor features the most consistent training and testing values as they have minimal differences and the testing data is the lowest of all models.

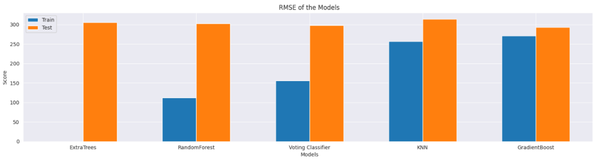


Fig. 15. RMSE Double Bar Chart for Ensemble Models

VI. Results

Various models were used during the completion of this report. PCA first transformed the data to reduce any overfitting issues that may be present. Correlation matrices were examined to show the lack of any linear relationship between the data. Four types of Linear regression were performed on the algae data, linear, lasso, ridge and elastic net. These models were all poor due to the lack of correlation in the data, verified by their high values for RMSE, negative score and lacklustre coefficients.

Examining a relationship of light compared to population revealed that it increased to a point before decreasing so polynomial regression was used. A loop was first constructed to test the ideal number of degrees which was revealed to be 2 and a model created a result. This model produced a nice readable output with a relatively low RMSE and r2-score of 96 per cent, showing how it could be an accurate model for use.

As an alternative to linear regression approaches, an ensemble regression was made. It made use of various ensemble models to create a final ensemble model through the use of the Voting Regressor model. Each of the models supplied came with certain strengths and weaknesses. ExtraTrees and Random Forest produced the best results for training of 100 and 99 per cent respectively for the training set which dropped to just under 96 per cent with the testing data. KNN and Gradient Boost Regressor had lower values of 97 per cent and only dropped to 96 per cent in testing. The voting regressor averaged these out to get a value of 98 for training and 96 for testing.

VII. Conclusion and Future Work

Algae is a very complex issue as it encompasses many facets of the world due to the variety of species of flora and fauna that fall under the term. As a result of this predicting the algae population would be a lot more complicated than just the factors here as external environmental factors such as type of water, local wildlife, season, weather and species of algae would also come into effect. These other factors have the potential to change the output of the model but also these factors could be more unpredictable like the different fish that could be eating algae at a certain time of year. That said, various models have showcased their potential utility in an algae population predictor. With most of the models bar Linear Regression for correlation reasons settling on an approximate r2 score of 96 per cent, the model should be more than sufficient for the development of an application that makes use of an ensemble or polynomial model.

Each of the aims set out at the start of the project has been completed. It has been verified that Light has the biggest effect of the columns on the population as the rest of the columns have little to no effect due to their non-correlated relationship. The population of algae was still able to be predicted under various different models that did not rely on a linear relationship.

In the future this project could expand to a system that monitors features such as Light, pH, etc to predict the population more easily, helping humans monitor and regulate their growth. The accuracy of the model itself could potentially be improved by changing some of the hyperparameters surrounding the models to suit the data and potentially more models to the ensemble for more results. It could help to combine other research that aimed to predict algae growth [8], [9]. The potential final application of this would take the previously mentioned external factors such as weather into account for use outside of a controlled environment and deployment for beneficial use.

References

[1] ryan85209, "How do algae affect water quality?," Aug. 2020.

[2] A. A. El Gamal, "Biological importance of marine algae," *Saudi Pharmaceutical Journal*, vol. 18, pp. 1–25, Jan. 2010.

[3] Utah State University, "Researchers Review Environmental Conditions Leading to Harmful Algae Blooms," Aug. 2019.

[4] R. Missonnier, "Research on Algae Growth," Sept. 2023.

[5] M. N. Metsoviti, G. Papapolymerou, I. T. Karapanagiotidis, and N. Katsoulas, "Effect of Light Intensity and Quality on Growth Rate and Composition of Chlorella vulgaris," *Plants*, vol. 9, p. 31, Dec. 2019.

[6] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*. John Wiley & Sons, Feb. 2021. Google-Books-ID: tCIgEAAAQBAJ.

[7] E. Lutins, "Ensemble Methods in Machine Learning: What are They and Why Use Them?," Aug. 2017.

[8] P. Yu, R. Gao, D. Zhang, and Z.-P. Liu, "Predicting coastal algal blooms with environmental factors by machine learning methods," *Ecological Indicators*, vol. 123, p. 107334, Apr. 2021.

[9] A. E. Brookfield, A. T. Hansen, P. L. Sullivan, J. A. Czuba, M. F. Kirk, L. Li, M. E. Newcomer, and G. Wilkinson, "Predicting algal blooms: Are we overlooking groundwater?," *Science of The Total Environment*, vol. 769, p. 144442, May 2021.