2022 SURF

AI-Based Prediction to Protect Jinji Lake Water Quality
Stage 1

Jeongyeong Park(2032801)

Table of Contents

Introduction

Define the Problem

From Jinji lake, massive values corresponding to these 16 variables: longitude, latitude, Time, chlorophyll, electrical conductivity, Low Frequency Water Depth(m), Dissolved Oxygen (% Sat), Dissolved oxygen (mg/L), Ammonia nitrogen, salinity, Phycotin, Total dissolved solids, turbidity, temperature, Ph value, and PH value (mv) were obtained. In this research, variables which are most related to Phycotin are indicated. The relationship is identified with a model and Phycotin can be predicted with it. The quality/performance of a model are also evaluated.

Technique for Solving Regression Problem

The problems dealt with in this research are regression problems since the output variable is a continuous numerical value. This research requires finding mathematical relationships between 'Phycotin' and other variables and predicting Phycotin. Thus, two models are decided to be used. One is multiple linear regression model. There are more than one variable that affects 'Phycotin', multiple linear regression model is used instead of simple linear regression model. The other one is decision trees. It is simple to understand and requires little data preparation. For effective and accurate research, every analysis is done with Python.

Data Analysis

Identify Required Data

Values corresponding to the16 variables: longitude, latitude, Time, chlorophyll, electrical conductivity, Low Frequency Water Depth(m), Dissolved Oxygen (% Sat), Dissolved oxygen (mg/L), Ammonia nitrogen, salinity, Phycotin, Total dissolved solids, turbidity, temperature, Ph value, and PH value (mv) obtained in Jinji Lake are required. The required Data are saved in one table of a csv file. Each value corresponding to each variable is saved in the same column. Values of 16 variables obtained together at once are in the same row and there is a total of 4533 groups of data.

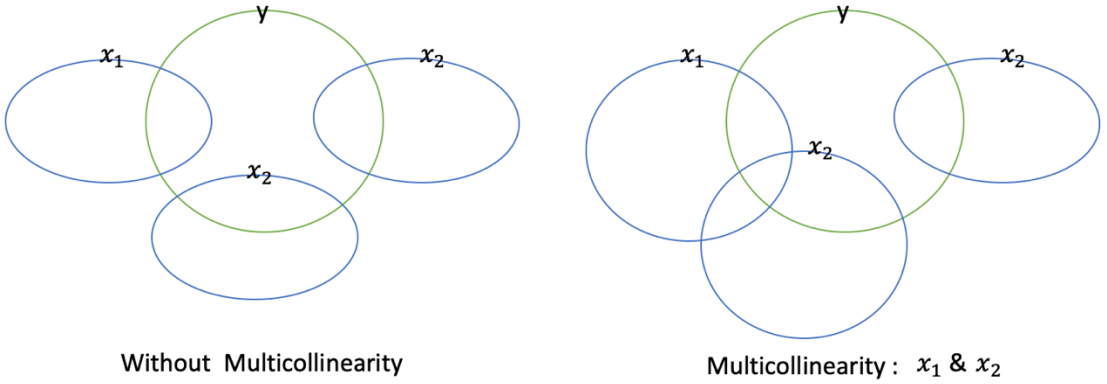
Prepare and Pre-process

By preprocessing data, data becomes easier to interpret and use. This procedure removes inconsistencies or duplicates in data, which can have a negative impact on the accuracy of a model. Data preprocessing also ensures that no incorrect or missing values are present as a result of human error or bugs. For the decision trees model, not much preprocessing is required compared to the multiple linear regression model, however, the same preprocessed data is used in this research.

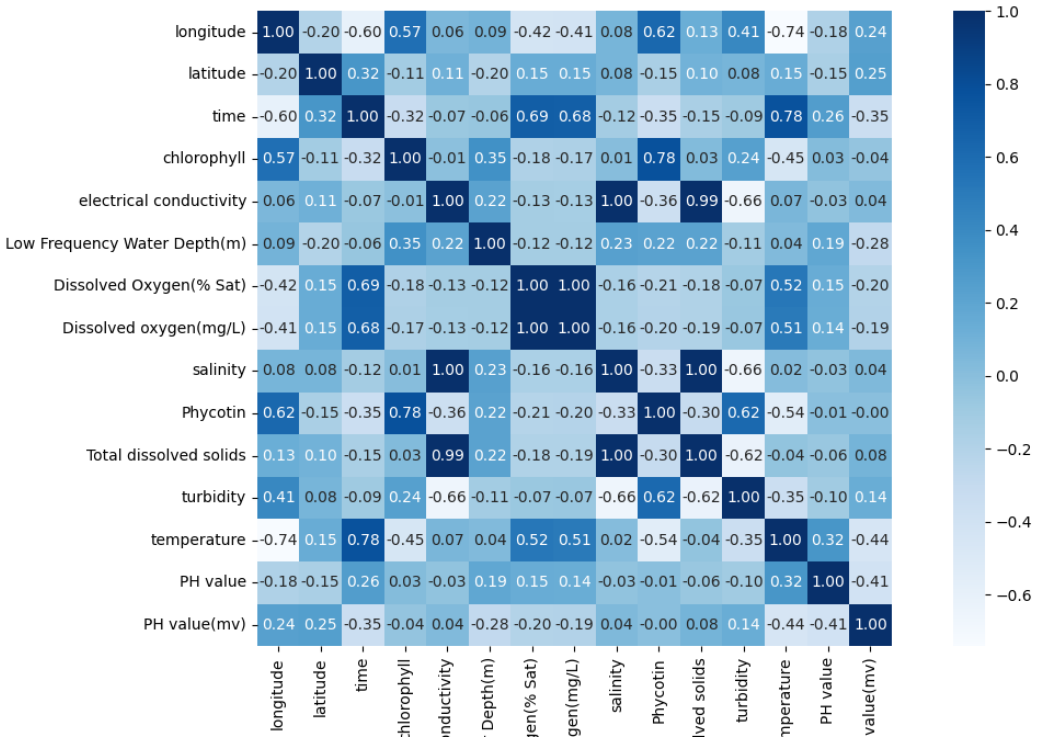
At first, the type of ‘Time’ is changed. The type of it is a string, but other variables are all in numbers so it is impossible to find a relationship between the string ‘Time’ and other variables. Thus, it is changed to an integer. Next, a row with any missing value is removed. In this dataset, there are 5 missing values in ‘Low Frequency Water Depth(m)’ column. ‘Ammonia nitrogen’ column is dropped because every value is 0 so there is no way it can affect Phycotin.

During data preprocessing, multicollinearity is considered. When an independent variable in a multiple regression equation is highly correlated with one or more of the other independent variables, multicollinearity exists.

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3$$



In this research, multicollinearity is considered with co-relationship heatmap and VIF. Through this heatmap, variables having value over |0.9| are first considered. There are 4 pairs: ‘Dissolved Oxygen(% Sat)’ & ‘Dissolved Oxygen(mg/L), ‘Salinity’ & ‘Electrical conductivity’, ‘Salinity’ & ‘Total dissolved Solids’, and ‘Electrical conductivity’ & ‘Total dissolved solids’. Thus, ‘Dissolved Oxygen(% Sat)’, ‘Total dissolved solids’ and ‘Salinity’ are dropped.



After checking correlation coefficient with heatmap, VIF value is also calculated. As it can be seen in the figure, the VIF value of ‘time’ is way over 10. Thus, ‘time’ is also removed.

	VIF_Factor	Feature
0	3.922061e+00	longitude
1	1.530595e+00	latitude
2	1.446200e+09	time
3	1.982417e+00	chlorophyll
4	3.015231e+00	electrical conductivity
5	1.462014e+00	Low Frequency Water Depth(m)
6	1.471851e+00	Dissolved oxygen(mg/L)
7	3.346732e+00	turbidity
8	3.484478e+00	temperature
9	1.274844e+00	PH value
10	1.683619e+00	PH value(mv)

Lastly, since the range of the remaining features is all different, the scaling is done. The mean is removed and each feature/variable is scaled to unit variance using StandardScaler.

Multiple Regression Model

Model the Data

The multiple regression model is built with Scikit-learn tool in Python. The multiple linear regression of this research is:

$$y = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5 + a_6x_6 + a_7x_7 + a_8x_8 + a_9x_9 + a_{10}x_{10}$$

Where y = Phycotin

x_1=Longitude

x_2=Latitude

x_3=Chlorophyll

x_4=Electrical Conductivity

x_5=Low Frequency Water Depth(m)

x_6=Dissolved Oxygen(mg/L)

x_7=Turbidity

x_8=Temperature

x_9=PH value

x_{10}=PH value(mv)

Train and Test

Dataset is split into training dataset and test dataset. 70% of data becomes training data and 30% becomes test data. By using training dataset, model is built. With the Python code, values of a_0 from a_{10} are calculated. Numbers in the array are values for a_1 to a_{10} and the number in the next line is a_0.

```
------------coefficient--------------
[[ 4.45424373 -2.02312228 36.6081806  -5.98081767  2.65102896 -0.62502969
  18.47012865 -9.14936738  1.23110387 -4.78625718]]

------------intercept--------------
[1004.94236018]
```

In order to test the multiple linear regression model, 10 features in test data are used to get each y which is ‘Phycotin’. That calculated values are compared with the ‘Phycotin’ value in the test data to check if this model can predict it properly. Detailed results will be discussed in the result section.

Decision Trees Model

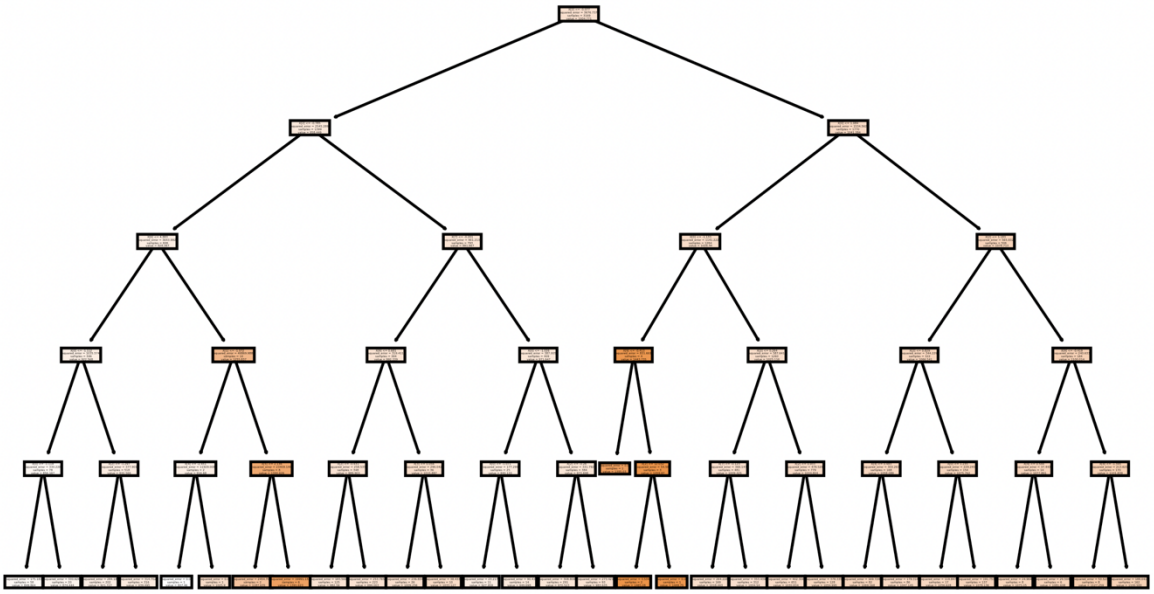
Model the Data

The decision trees model is also built with Scikit-learn tool in Python. For generating decision trees, the max depth of tree is considered. The deeper the tree grows, the more complex the model becomes because it captures more segmentation and more information

about the data. It might cause the overfitting. Also, it is not good to have very low depth because the model is underfitting. Thus, tree with depth 2 was generated first and then the predicted values model generated were considered. After generating trees with depth 3,4, and 5, decision trees model with depth 5 is decided to be used in this research.

Train and Test

Dataset is split into training dataset and test dataset just like dealing with the multiple linear regression model. 70% of data becomes training data and 30% becomes test data. By using training dataset, tree model is built.



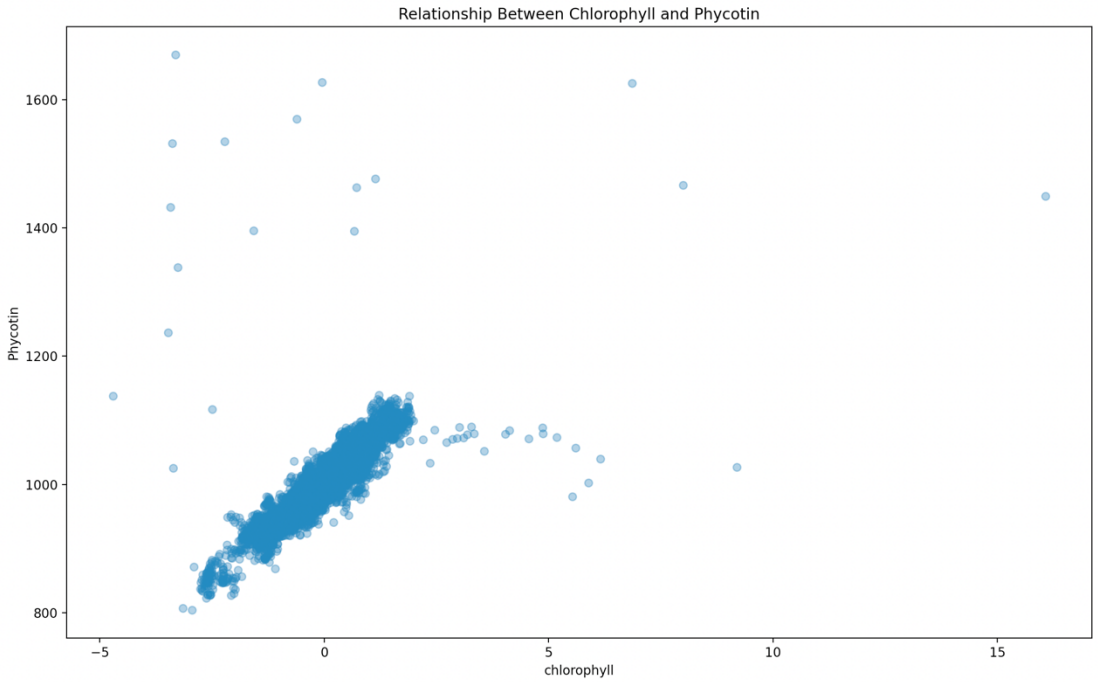
In order to test the model, 10 features in test data are used to get each y value which is 'Phycotin'. That calculated values are compared with the actual 'Phycotin' value in the test data to check if this model can predict it properly.

Result

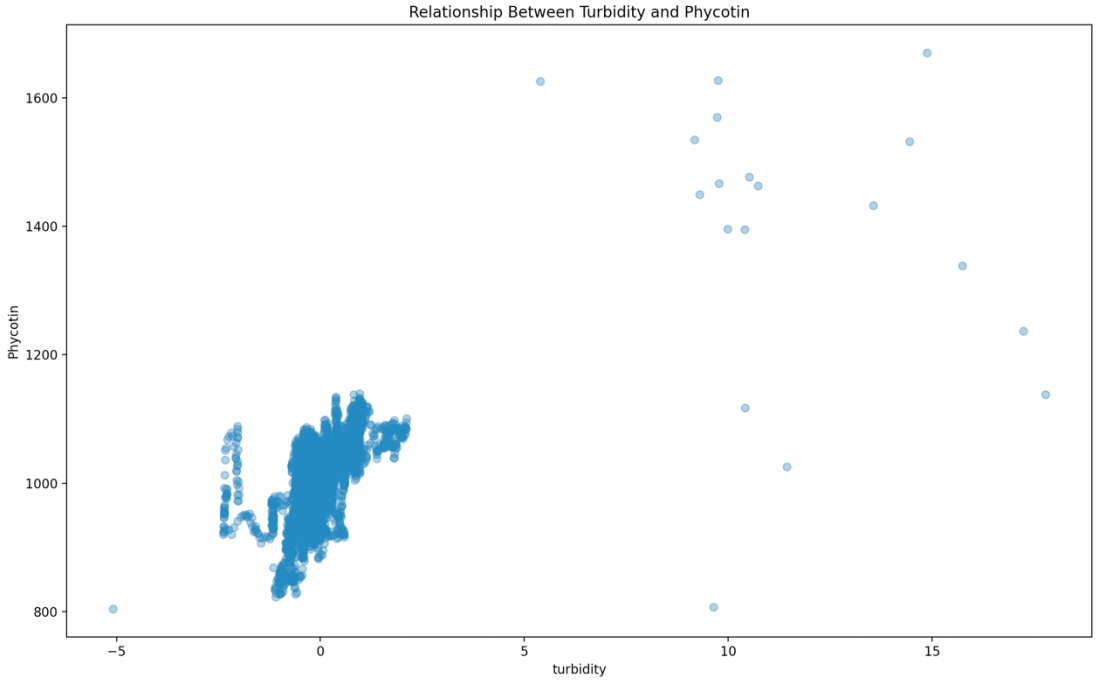
Multiple Regression Model

Variables most related to Phycotin

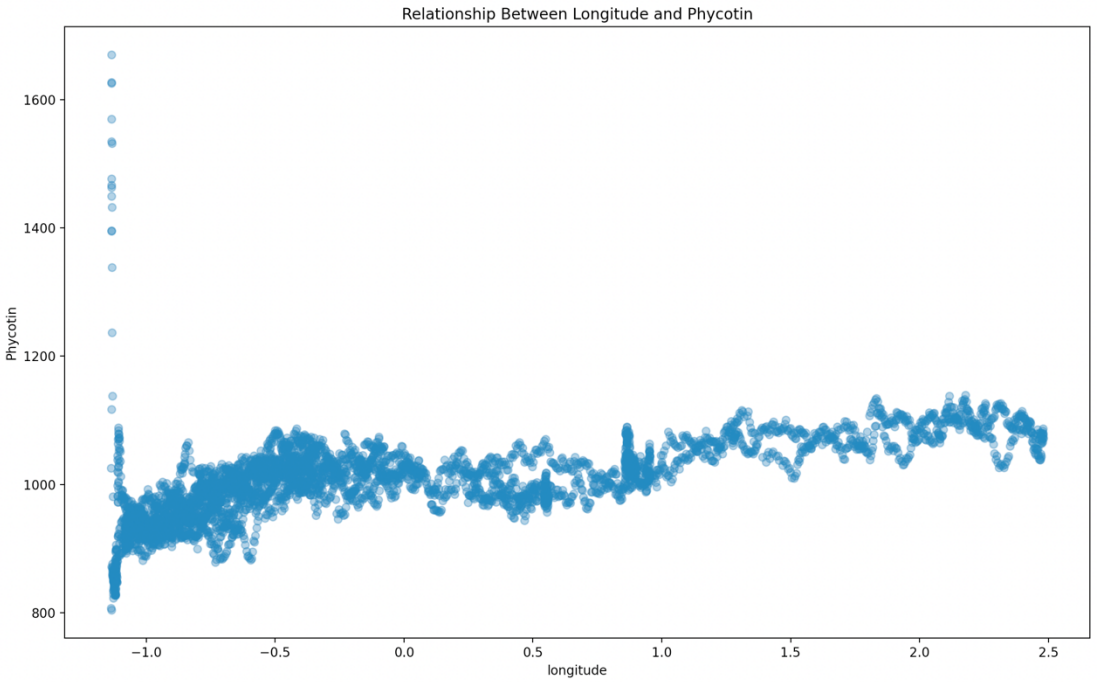
According to the regression coefficients of multiple linear regressions model, 'Chlorophyll' variable, 'Turbidity' variable, and 'Longitude' variable seems like they are most relevant to 'Phycotin'. However, it is difficult to see the relationship between each variable and 'Phycotin' with only coefficients, scatter plots are generated. Every scatter plots graph can be seen in the appendix.



The above graph shows that the value of ‘Phycotin’ tends to increase as the value of ‘Chlorophyll’ grows.



The above graph does not show clear relationship like Chlorophyll & Phycotin graph, it tends to show that the increase in the value of ‘Turbidity’ affects the increase in the ‘Phycotin’ value.



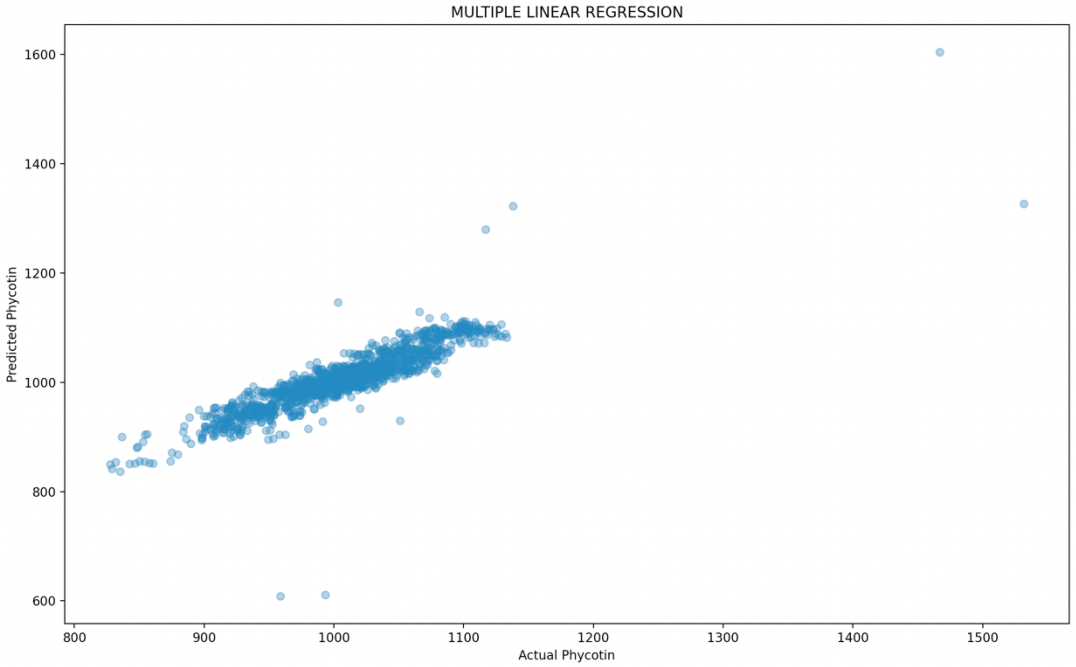
In the above graph, the ‘Phycotin’ value tends to increase as the ‘Longitude’ value increases. Thus, the variables which are most related to ‘Phycotin’ are ‘Chlorophyll’ variable, ‘Turbidity’ variable, and ‘Longitude’ variable.

Predict Phycotin

In order to see if the model can predict ‘Phycotin’ successfully, the predicted y value (Phycotin) is calculated with features(x) in the test dataset. Then, the predicted y values are compared with each of the actual y values of the test dataset. Actual value and predicted value can be compared with this table. It seems the model is predicting ‘Phycotin’ value properly.

Actual Value	Predicted Value
1104.95	1002.752216
1015.39	957.237534
1067.26	973.931800
1086.25	991.456711
1055.61	974.986175
...	...
1074.34	985.098175
1040.13	1101.887007
1056.09	936.839756
1070.57	1092.002563
1085.53	1008.912997

However, since the table only shows a few data and it is difficult to recognize, the scatter plot is used to see the result more clearly.



The more accurately the model predicts the value, the closer this graph will be to the straight line. The above graph is not a straight line but it is relatively close to the straight-line trend.

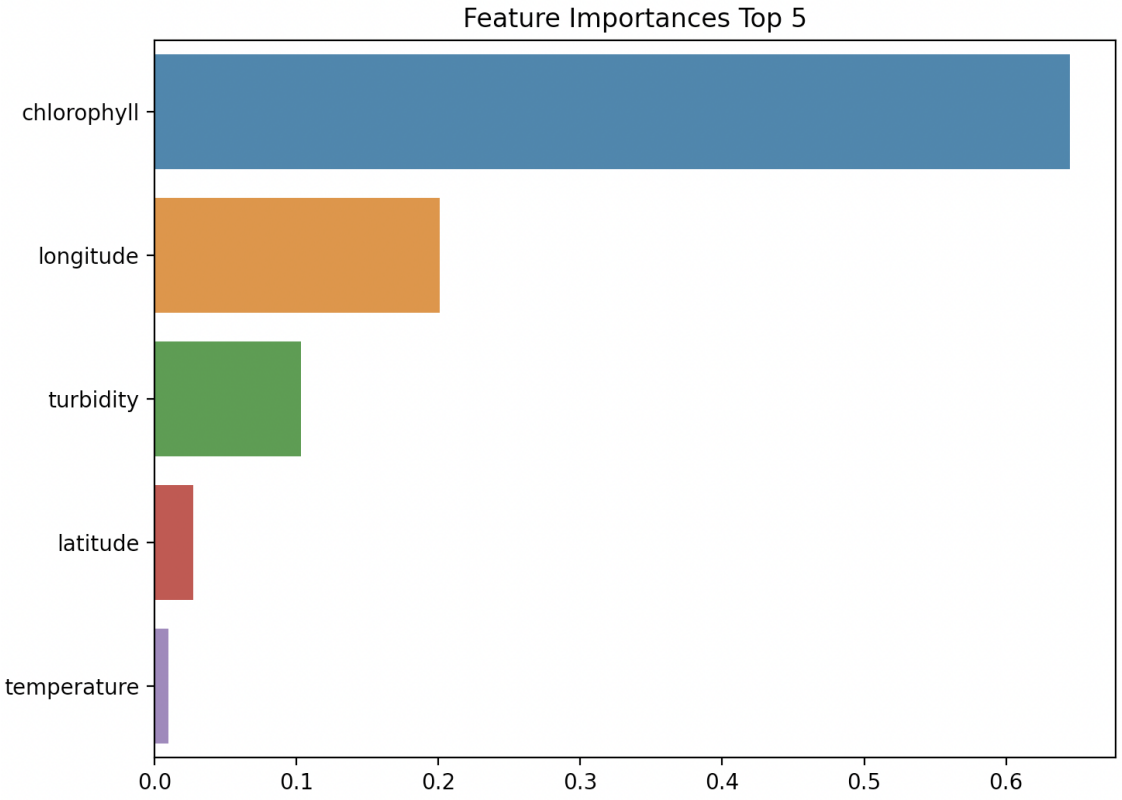
```
--------------Accuracy----------------
0.8356049960922485
83.56049960922485 %
```

Besides this table and graph, the accuracy of the model against the training data is measured. Thus, it shows the model can accurately predict the ‘Phycotin’ value with 10 variables.

Decision Trees Model

Variables most related to Phycotin

In order to find variables most related to ‘Phycotin’, feature importance is calculated. After obtaining values of each feature, they are sorted in descending order. In order to see the feature importance easily, the graph is generated within the code.



According to the model, the variables which are most related to ‘Phycotin’ are ‘Chlorophyll’ variable, ‘Turbidity’ variable, and ‘Longitude’ variable.

Predict Phycotin

The same technique as the multiple linear regression model is used to see if the model can predict ‘Phycotin’ successfully. The predicted y value (Phycotin) is calculated with features(x) in the test dataset. Then, the predicted y values are compared with each of the actual y values of the test dataset.

```
----------Compare Actual Value with Predicted Value--------
  Actual Value  Predicted Value
       1062.68      1078.135912
       1115.03      1030.833840
        963.88      1030.833840
       1084.79       994.003049
       1058.85       939.494873
           ...              ...
       1103.24       994.003049
       1035.27       921.727030
       1056.56      1069.567683
       1036.54       921.727030
       1070.46      1100.583210
```

The data shows the model properly predicts ‘Phycotin’ value with 10 features. For more clear evaluation of predict performance, mean squared error value and R-squared value are calculated.

```
--------------Accuracy----------------
RMSE: 25.99876818700891
R^2: 0.8498092768065261
```

The mean squared error value is relatively low and R-squared value is relatively high, it shows the model’s prediction performance is appropriate.

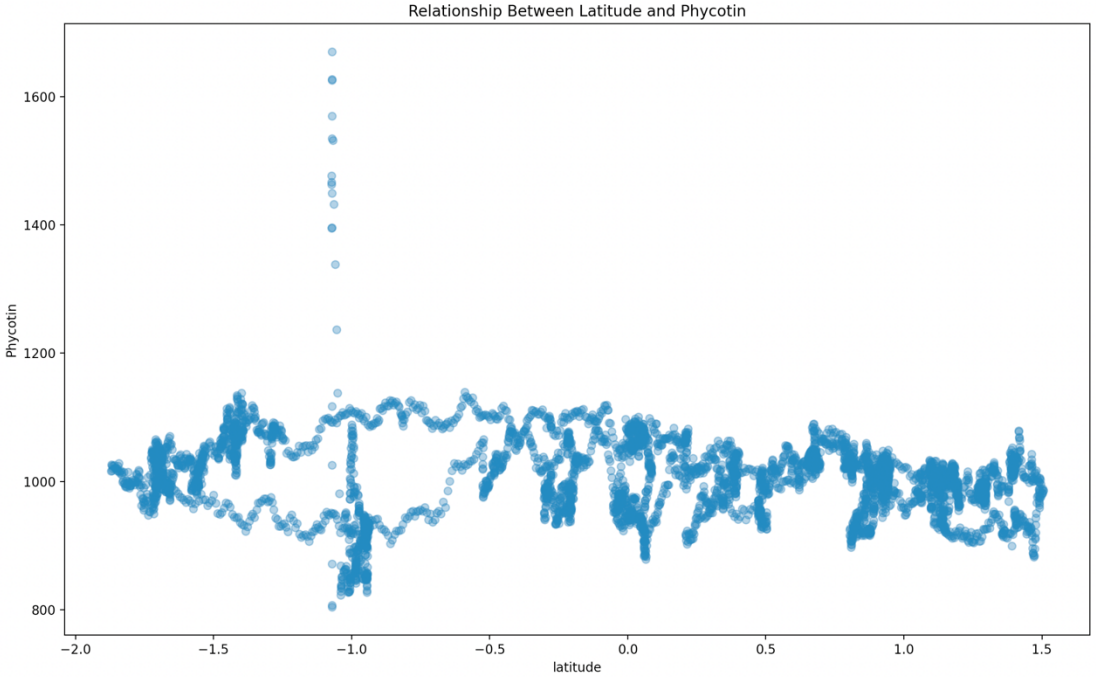
Conclusion

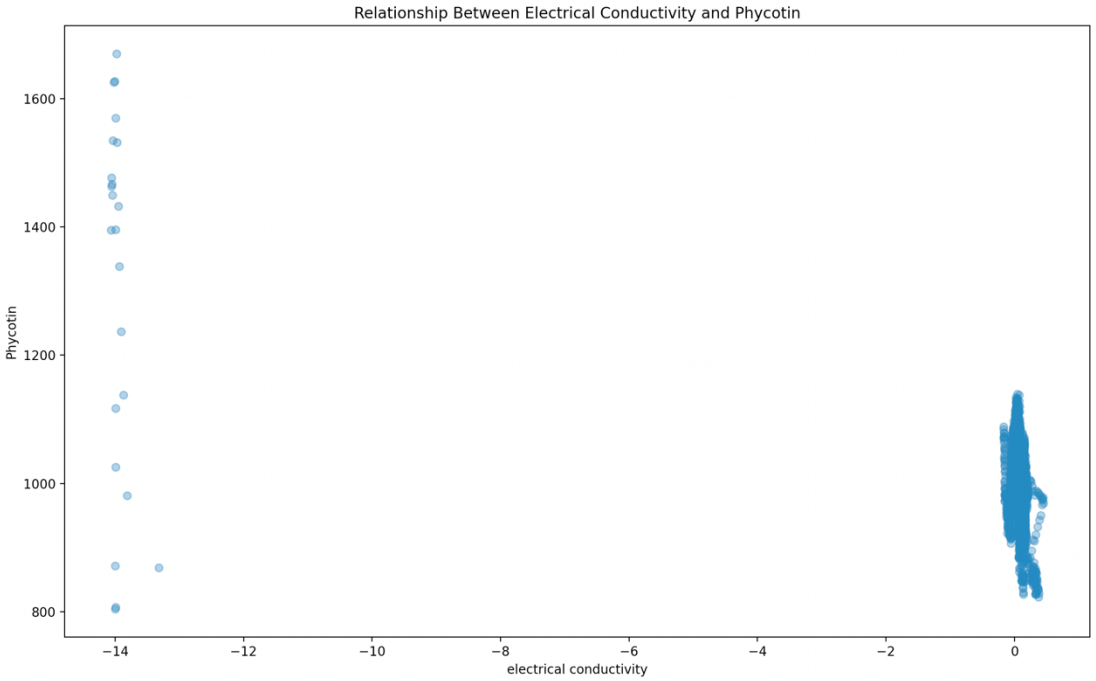
A multiple linear regression model and decision trees model are built in order to find variables most related to ‘Phycotin’ and predict ‘Phycotin’. In the original dataset, there are 16 variables including ‘Phycotin’ which is the target and 15 variables are considered as features. After data pre-processing, 10 variables ‘Longitude’, ‘Latitude’, ‘Chlorophyll’, ‘Electrical Conductivity', ‘Low-Frequency Water Depth(m)’, ‘Dissolved Oxygen(mg/L)’, ‘Turbidity’, ‘Temperature’, ‘PH value’, and ‘PH value(mv)’ are selected for features(x). 70% of the dataset is used for training data and 30% of it is used for test data.

Both models show that the variables most related to ‘Phycotin’ are ‘Chlorophyll’, ‘Turbidity’, and ‘Longitude’. Also, with these models, the ‘Phycotin’ value can be predicted with reasonable accuracy. However, as can be seen in the result of multiple linear regression model, there are some data not following the trend. It seems to be a problem caused by not considering outliers during the data pre-processing. Moreover, in the decision trees model, the result is slightly different every run. It seems to be affected by the values of the training dataset randomly selected each run. In order to build a more accurate model in the future, these problems need to be addressed with further study. Nonetheless, in overall, the accuracy of two models are reasonable. During the data pre-processing, the ‘time’ format is changed from string to integer and missing values are removed. In addition, multicollinearity is eliminated and the data is scaled. Based on testing and checking the performance with a few methods, the model can be concluded as having high accuracy. In conclusion, the built models can well accomplish the purpose of this research.

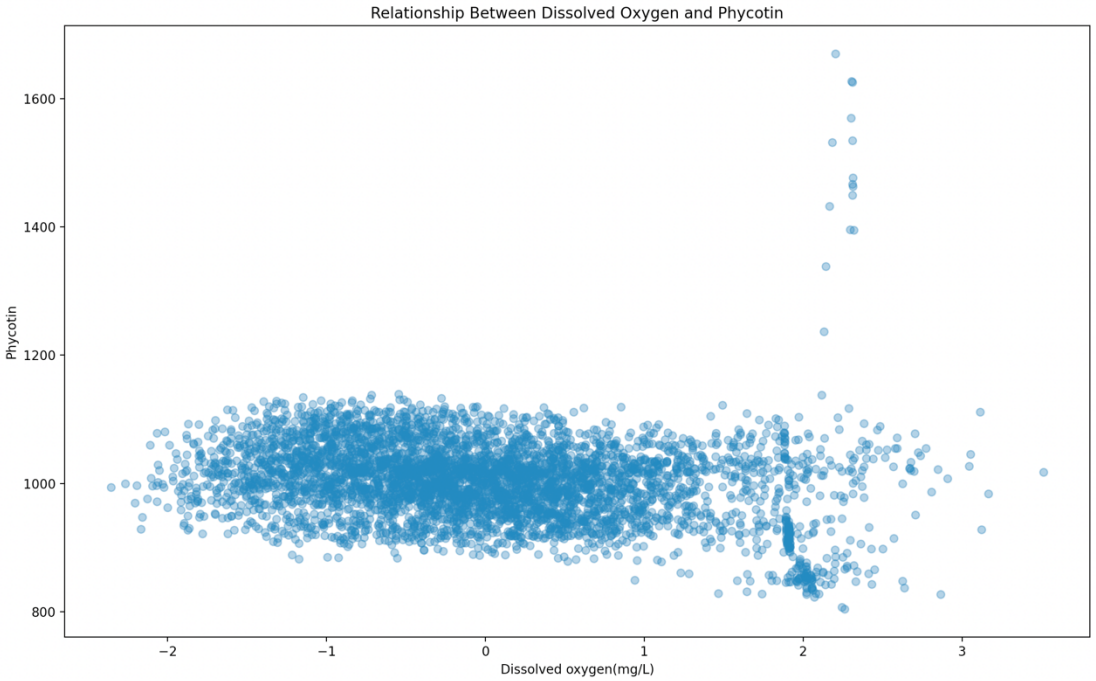
Appendix

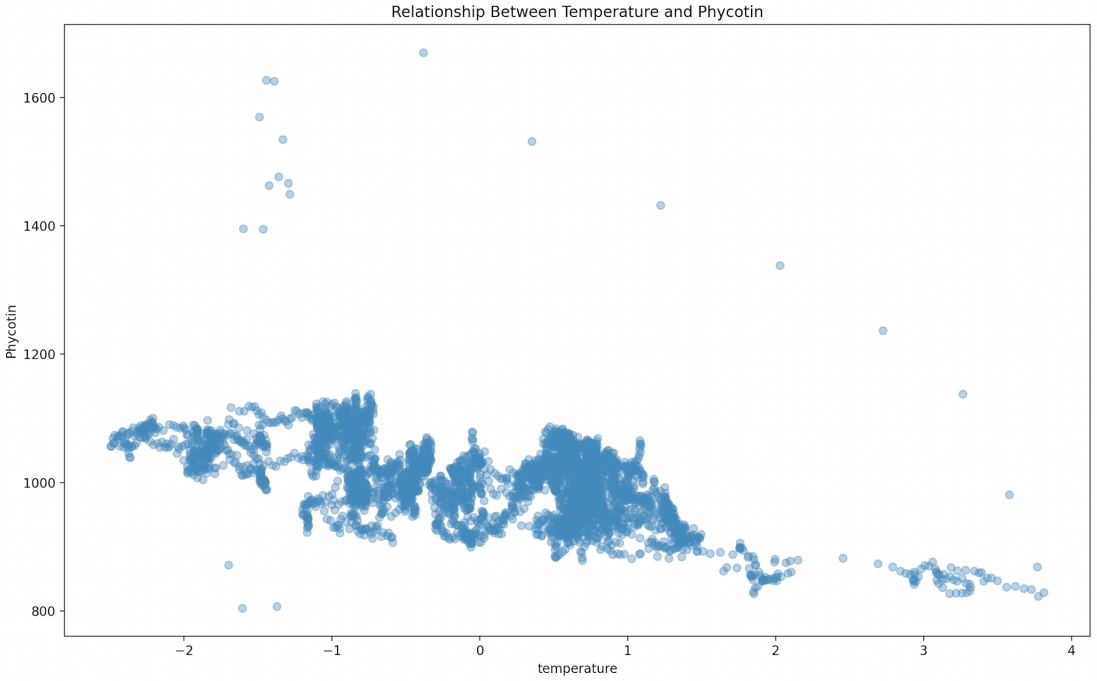
Variables not related to Phycotin

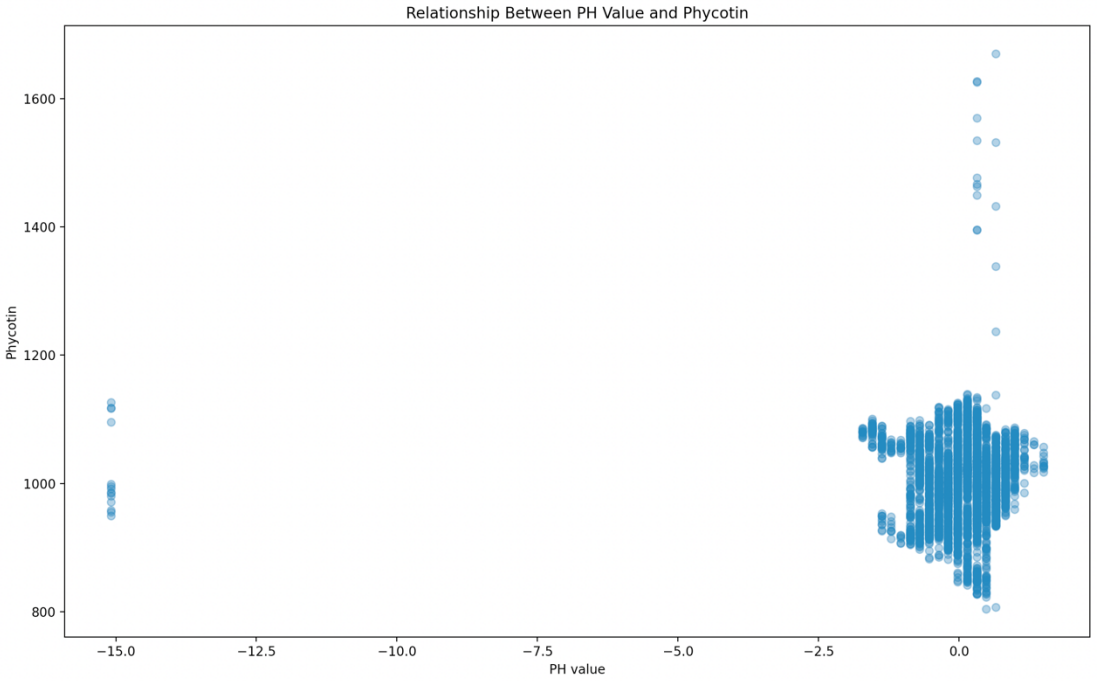


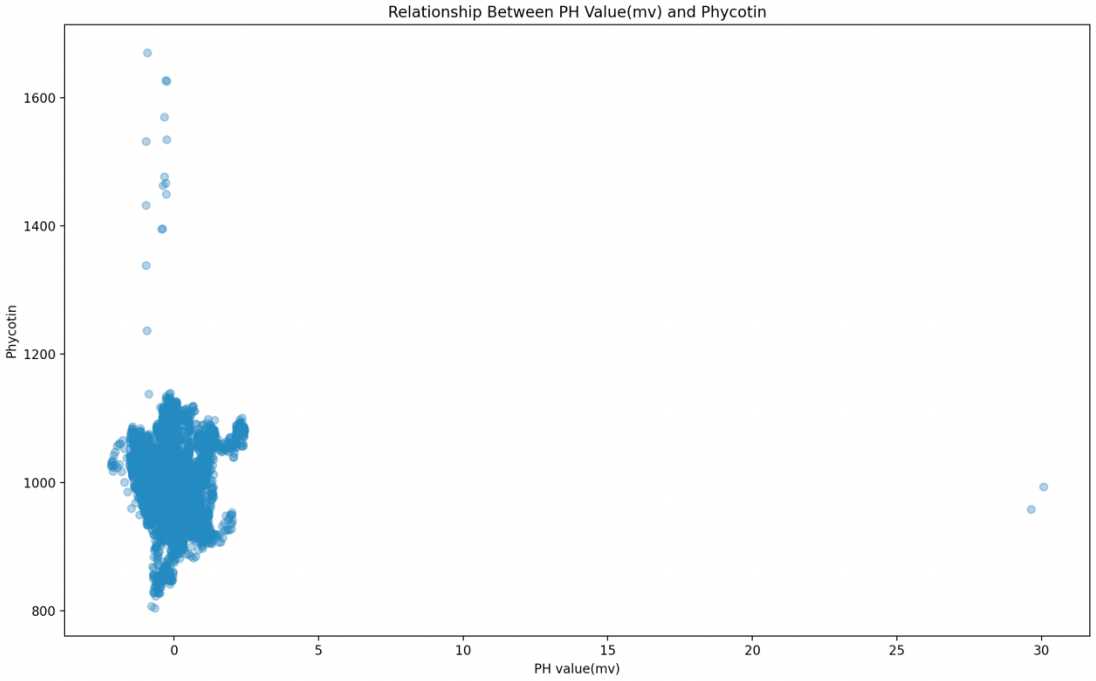



Relationship Between Low Frequency Water Depth and Phycotin
Phycotin
1600
1400
1200
1000
800
−3
−2
−1
0
1
2
Low Frequency Water Depth(m)

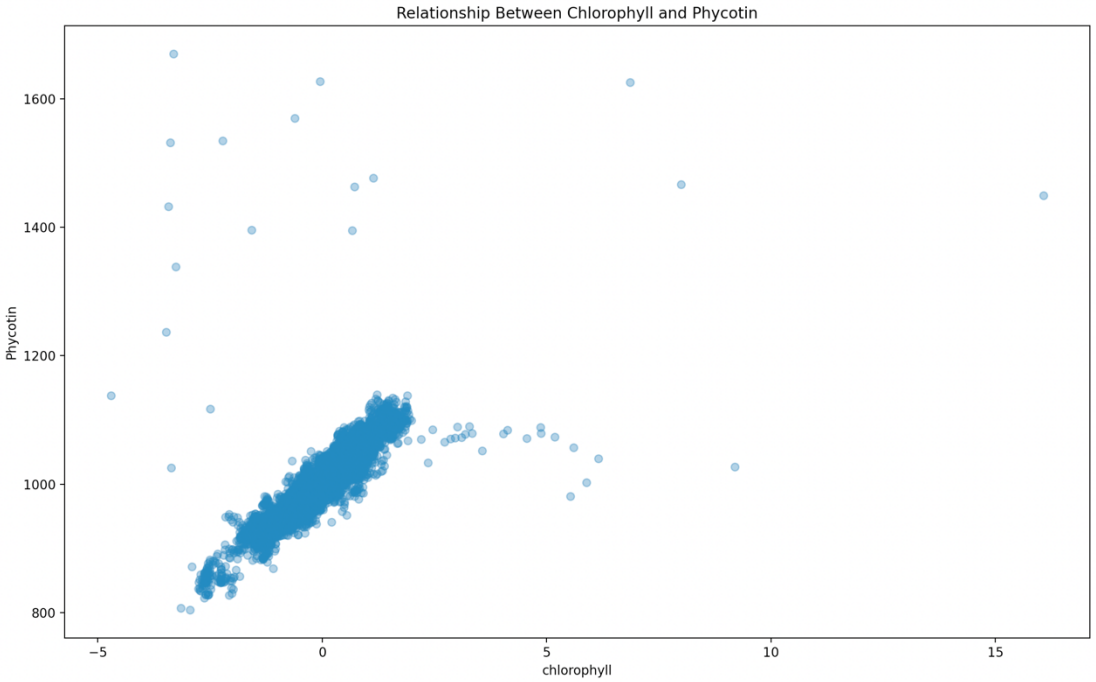

Relationship Between Dissolved Oxygen and Phycotin
Phycotin
1600
1400
1200
1000
800
−2
−1
0
1
2
3
Dissolved oxygen(mg/L)

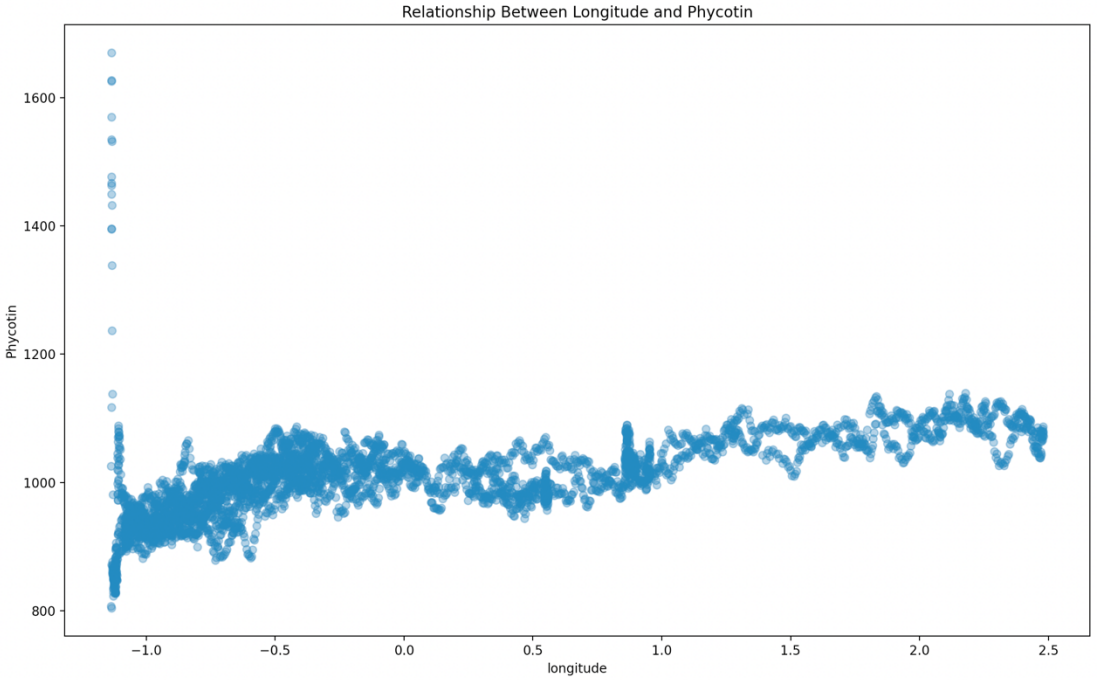

Relationship Between Temperature and Phycotin
Phycotin
temperature
1600
1400
1200
1000
800
−2
−1
0
1
2
3
4

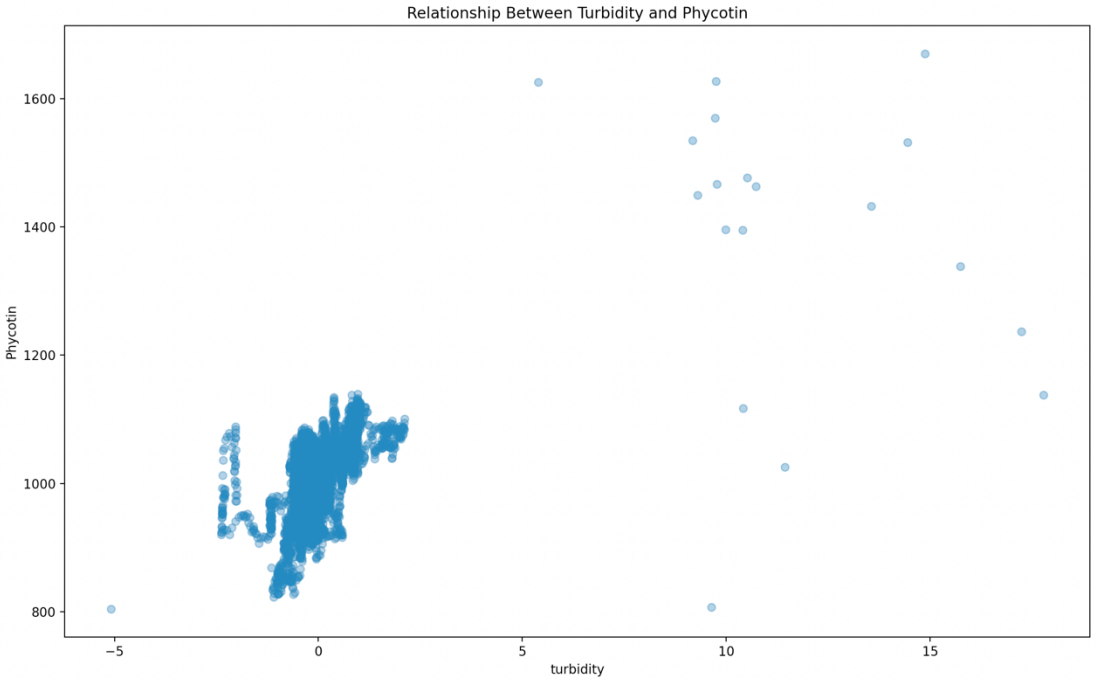

Relationship Between PH Value and Phycotin
Phycotin
PH value
1600
1400
1200
1000
800
−15.0
−12.5
−10.0
−7.5
−5.0
−2.5
0.0



Variables related to Phycotin







Brief Explanation about Python Code

Function	Description
dataPre()	-Data preprocessing -Return final dataframe

dataAnalysis_MLR()	-Data analysis with Multiple Linear Regression model
dataAnalysis_DT()	-Data analysis with Decision Tree model
correlationHeatmap(df)	-Calculate correlation coefficients and draw Heatmap of them
vif(df)	-Calculate VIF value of variables
dataScale(df)	-Scale final data
Model_MLR(x_train,y_train)	-Create multiple linear regression model with training dataset
Model_DT(x_train,y_train)	-Create decision trees model with training dataset
validate_MLR(model, x_train, x_test, y_train, y_test)	-Generate a table and draw a scatter plot for Actual Value & Predicted Value - Calculate accuracy of a multiple linear regression model
validate_DT(model, x_train, x_test, y_train, y_test)	-Generate a table for Actual Value & Predicted Value - Calculate mean squared error value and R-squared value of a decision tree model
featureImpo(model)	-Obtain feature importance of decision tree model
visualization(df_final)	-Draw scatter plot for relationship between each feature and Phycotin respectively

When the code is run, the system will ask:

```
Please choose a model
Type 'MLR' for Multiple Linear Regression
Type 'DT' for Decision Trees
->
```

If the user types MLR, the multiple linear regression model will be used for analysis, and if the user types DT, the decision trees model will be used.