Predicting Water Temperature in Ponds Using Machine Learning: A Random Forest Approach

PRANESH M[1], Dr. Nithyanandh Selvam[2]

[1] Student, Department of Computer Applications (MCA), PSGCAS, Coimbatore.

[2] Associate Professor, Department of Computer Applications (MCA), PSGCAS, Coimbatore.

Abstract:

The preservation of ecological balance and water quality depends on the precise estimation of aquatic parameters. One important element that affects biological activity, chemical reactions, and the general health of an ecosystem is water temperature. Conventional monitoring methods are useful for observations made in real time, but they frequently can't forecast future changes. The use of machine learning models, specifically the Random Forest Regressor, and time-series analysis to predict water temperature in a pond setting is investigated in this study. Key factors impacting temperature fluctuations are identified through correlation analysis, lagged features, and historical data. The findings show that machine learning techniques improve predicted accuracy greatly when compared to traditional statistical methods. This study demonstrates the potential of predictive modeling in improving environmental monitoring and aquaculture management by allowing for proactive decision-making and early detection of anomalies.

Keyword:

[Pond water, Analysis, Correlation, Data, Accuracy]

1. Introduction:

Water temperature is critical in aquatic habitats, impacting chemical processes, biological activity, and overall water quality. Temperature changes influence the metabolic rates of aquatic organisms, the solubility of gases like oxygen, and the multiplication of algae and bacteria. Maintaining ideal temperature conditions in aquaculture and water resource management is critical for aquatic life survival and environmental preservation. Traditional monitoring systems rely on periodic sampling and sensor-based readings, which, while valuable, do not provide predictive information. The increasing availability of large-scale environmental data creates an opportunity to use advanced data analytics for forecasting. Machine learning algorithms provide an effective alternative by recognizing patterns and trends in historical data to effectively forecast future temperature swings. The purpose of this work is to create an effective predictive model for water temperature using machine learning methods, notably Random Forest Regressor, by adding time-series information such lagged variables and correlation-based feature selection. The outcomes of this study help to improve environmental monitoring by allowing for early interventions and informed decision-making in water resource management.

2. Literature Review:

The Data analysis is done to check weather the pond condition is suitable for the cultivation of algae growth, Machine learning algorithms are helpful for us in maintaining the pond condition and helpful for the decision making process easier.

In book of "Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32" helpful for the understanding the algorithm.

In book of “Li, J., Zhang, Q., & Guo, X. (2019). Predicting Water Quality using Machine Learning Models: A Case Study on Aquaculture Ponds”

3. Aim and Objective:

This journal’s main goal is to show how the machine learning algorithms are useful for us in the understanding in the pond water analysis.

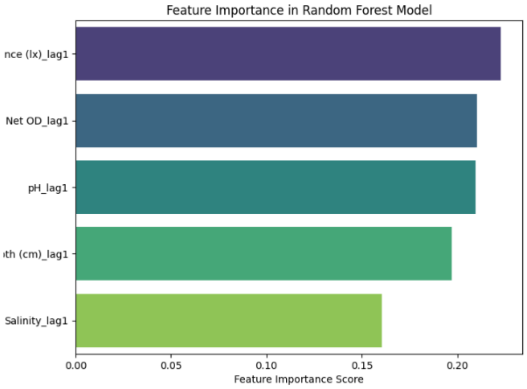
The objective is to utilize various analyzation methods to see how the analyzation results by showing various methods such as correlation matrix & time-series to analyze the pond water data,

4. Methodology:

During this analyzation process there are various methods are used for the analysis of the pond water data’s. There are various parameters are present in the data this are processed through the various data analysis process and use random forest regression algorithm for the temperature analysis.

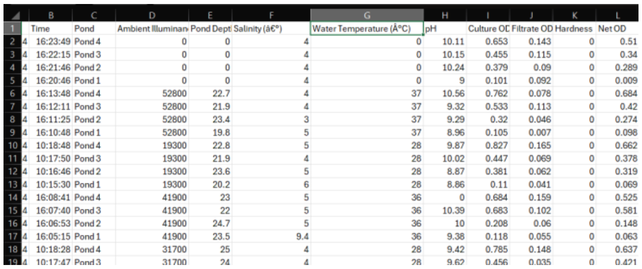
4.1. Selection of the algorithm:

The choosing of suitable algorithm is quite challenging for analyzing the data. The random forest regression algorithm is one of the supervised machine learning algorithm which have high accuracy and easy for the analyzing for the early stage of the analyzation of the pond data.



4.2. Data Collection:

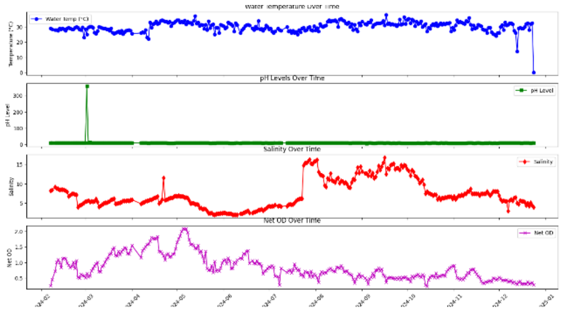
To do the analyzation the data for the analyzation process are collected from various sources like Kaggle, and the readings of the pond data are collected form the various sensors present in the pond. There are various parameters are present for the analysis of the data like ph, light intensity, salinity, water temperature etc.



		B	C	D	E	F	G	H	I	J	K	L
1		Time	Pond	Ambient Illuminan	Pond Dept	Salinity (â€°)	Water Temperature (Â°C)	pH	Culture OD	Filtrate OD	Hardness	Net OD
2	4	16:23:49	Pond 4	0	0	4	0	10.11	0.653	0.143	0	0.51
3	4	16:22:15	Pond 3	0	0	4	0	10.15	0.455	0.115	0	0.34
4	4	16:21:46	Pond 2	0	0	4	0	10.24	0.379	0.09	0	0.289
5	4	16:20:46	Pond 1	0	0	4	0	9	0.101	0.092	0	0.009
6	4	16:13:48	Pond 4	52800	22.7	4	37	10.56	0.762	0.078	0	0.684
7	4	16:12:11	Pond 3	52800	21.9	4	37	9.32	0.533	0.113	0	0.42
8	4	16:11:25	Pond 2	52800	23.4	3	37	9.29	0.32	0.046	0	0.274
9	4	16:10:48	Pond 1	52800	19.8	5	37	8.96	0.105	0.007	0	0.098
10	4	10:18:48	Pond 4	19300	22.8	5	28	9.87	0.827	0.165	0	0.662
11	4	10:17:50	Pond 3	19300	21.9	4	28	10.02	0.447	0.069	0	0.378
12	4	10:16:46	Pond 2	19300	23.6	5	28	8.87	0.381	0.062	0	0.319
13	4	10:15:30	Pond 1	19300	20.2	6	28	8.86	0.11	0.041	0	0.069
14	4	16:08:41	Pond 4	41900	23	5	36	0	0.684	0.159	0	0.525
15	4	16:07:40	Pond 3	41900	22	5	36	10.39	0.683	0.102	0	0.581
16	4	16:06:53	Pond 2	41900	24.7	5	36	10	0.208	0.06	0	0.148
17	4	16:05:15	Pond 1	41900	23.5	9.4	36	9.38	0.118	0.055	0	0.063
18	4	10:18:28	Pond 4	31700	25	4	28	9.42	0.785	0.148	0	0.637
19	4	10:17:47	Pond 3	31700	24	4	28	9.62	0.456	0.035	0	0.421

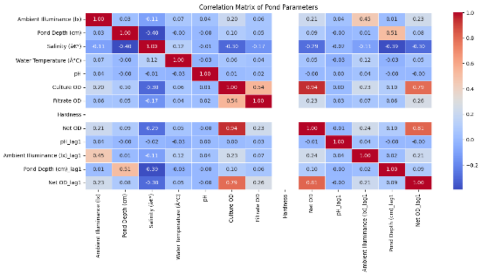
4.3. Data Preprocessing:

Standardizing column names, parsing dates into DateTime objects, handling missing values through imputation or removal, sorting chronologically, engineering features like rolling statistics and lagged variables, identifying and managing outliers, resampling to modify frequency, breaking down the series into trend, seasonality, and residuals, and guaranteeing stationarity for precise modeling are all part of preprocessing time-series data.



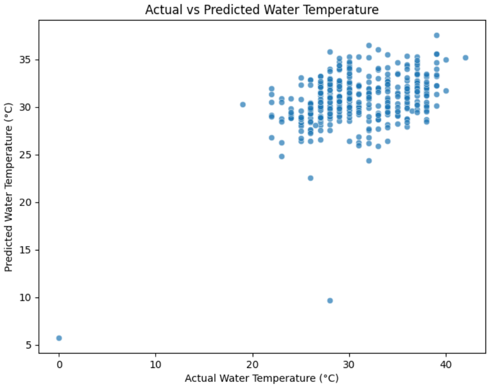
5. Model evaluation and Methods:

For the purpose of selecting important features such as pH, salinity, illuminance, depth, and OD, exploratory data analysis (EDA) started with the creation of a correlation matrix and heatmap to visualize feature interdependencies. This was followed by time-series analysis to find seasonal patterns influencing Spirulina growth. A Random Forest Regressor, selected for its robustness, was trained using this selection on an 80% training set, with the remaining 20% left aside for testing. Cross-validation techniques were used to verify the model's generalization abilities and reduce overfitting, and residual analysis was taken into account to make sure the model's assumptions were followed. The model evaluation process went beyond Mean Absolute Error (MAE) and R2 score. The model exhibited its ability to accurately predict temperature with a balance of low MAE and a satisfactory R2 score. Future monitoring is planned to evaluate performance in practical applications and take into consideration possible changes in environmental conditions.



6. Conclusion and Future Enhancement:

In order to anticipate pond water temperature—a crucial component of sustainable spiralina farming—this study showed how machine learning, more especially a Random Forest Regressor, may be applied effectively. The model's potential for real-world application is indicated by its respectable accuracy, as seen by its MAE and R2 score. However, future research will concentrate on a few crucial areas to further improve predictive skills and robustness. First off, adding a larger history dataset will provide the model a deeper comprehension of environmental variations and temporal patterns, which will enhance training and generalization. Second, investigating cutting-edge algorithms like XGBoost and Neural Networks, which are renowned for their capacity to capture intricate relationships, may result in more accurate predictions. Thirdly, improving feature engineering methods by developing more complex lagged variables and interaction features will try to glean more profound insights from the data that is now available. Fourth, by incorporating external environmental parameters like weather forecasts, seasonal fluctuations, and solar radiation statistics, a more robust and comprehensive model that can adjust to changing conditions will be produced. Together with the prediction model, the creation of a real-time monitoring system will also allow for proactive pond condition adjustments, maximizing Spirulina growth and yield. Last but not least, examining the model's sensitivity to different pond depths and geographic locations would guarantee that it can be applied to a variety of Spirulina farming configurations, thereby producing a more universal and widely accepted solution.



7. References:

1. Boyd, C. E. (2015). Water Quality: An Introduction. Springer.

2. Stumm, W., & Morgan, J. J. (2012). Aquatic Chemistry: Chemical Equilibria and Rates in Natural Waters. Wiley.

3. Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.

4. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32

5. Dou, X., & Weng, J. (2021). Machine Learning in Environmental Science: Applications and Trends. Elsevier

6. Li, J., Zhang, Q., & Guo, X. (2019). Predicting Water Quality using Machine Learning Models: A Case Study on Aquaculture Ponds