Predicting Water Temperature in Ponds Using Machine Learning

Random Forest Approach

PROJECT REPORT

Submitted by

PRANESH M

23MCA033

Under the Guidance of

Dr. Nithyanandh Selvam

Assistant Professor

Department of Computer Applications (PG)

In partial fulfillment of the requirements for the award of the degree of

MASTER OF COMPUTER APPLICATIONS

of Bharathiar University

DEPARTMENT OF COMPUTER APPLICATIONS (PG)

PSG COLLEGE OF ARTS & SCIENCE

An Autonomous College, Affiliated to Bharathiar University

Accredited with ‘A++’ grade by NAAC (4th Cycle)

College with Potential for Excellence

(Status Awarded by the UGC)

Star College Status Awarded by DBT – MST

An ISO 9001:2015 Certified Institution

Civil Aerodrome Post, Coimbatore - 641 014

APRIL 2025

VIRIDIA

Viridia Biotech LLP
1/101, Near Raja Middle School,
Veduchipalayam PO,
Karur - 639 114,
Tamil Nadu, India

Project

Exploring Hidden Correlations in Multivariate Time-Series Data

Context:
Our company has collected 1 years worth of time-series data spanning 8 parameters across 4 ponds. While some correlations among these parameters are already known and some are derived, there may be underlying relationships or patterns that are not yet known to us. Uncovering these hidden correlations could provide deeper insights into the data, potentially leading to improved decision-making or process optimization.

Objective:
The goal of this project is to explore and analyze the time-series dataset to identify previously unknown correlations or patterns. This analysis will involve leveraging statistical and visualization techniques, and if possible, basic machine learning methods.

Scope:

- Perform exploratory data analysis to understand the distribution and trends of each parameter.
- Investigate potential correlations between parameters using statistical tools.
- Experiment with visualization techniques (e.g., heatmaps, time-series plots) to highlight any emerging patterns.
- Summarize findings, highlighting any newly discovered relationships.

Secondary Scope:
In addition to identifying static correlations among the parameters, an attempt can be made to explore temporal relationships, i.e., how changes in today's parameters influence tomorrow's parameters. This analysis could provide insights into causal or lagged effects within the dataset.

Expectations:
The project is designed as an internship learning opportunity. While groundbreaking discoveries are not anticipated, a structured and methodical approach to data exploration and documentation is expected. The results, regardless of the depth of new insights, will contribute to the company's knowledge base.

For Viridia Biotech LLP,

Designated Partner.

PSG COLLEGE OF ARTS & SCIENCE

An Autonomous College, Affiliated to Bharathiar University

Accredited with 'A++' grade by NAAC (4th Cycle)

College with Potential for Excellence

(Status Awarded by the UGC)

Star College Status Awarded by DBT – MST

An ISO 9001:2015 Certified Institution

Civil Aerodrome Post, Coimbatore - 641 014

DEPARTMENT OF COMPUTER APPLICATIONS (PG)

CERTIFICATE

This is to certify that this project work entitled **Predicting Water Temperature in Ponds Using Machine Learning Random Forest Approach** is a Bonafide record of work done by **PRANESH M (23MCA033)** in partial fulfillment of the requirements of the award of Degree of **Master of Computer Applications (PG)**, of Bharathiar University.

Faculty Guide   Head of the Department

Submitted for Viva-Voce Examination held on ____________________

Internal Examiner   External Examiner

PSG COLLEGE OF ARTS & SCIENCE

An Autonomous College, Affiliated to Bharathiar University
Accredited with ‘A++’ grade by NAAC (4th Cycle)
College with Potential for Excellence
(Status Awarded by the UGC)
Star College Status Awarded by DBT – MST
An ISO 9001:2015 Certified Institution
Civil Aerodrome Post, Coimbatore - 641 014

DEPARTMENT OF COMPUTER APPLICATIONS (PG)

DECLARATION

I, **PRANESH M (23MCA033),** hereby declare that this Project work entitled **Predicting Water Temperature in Ponds Using Machine Learning Random Forest Approach** is submitted to PSG College of Arts & Science (Autonomous), Coimbatore in partial fulfillment for the award of degree of Master of Computer Applications, is a record of original work done by me under the supervision and guidance of **Dr. Nithyanandh Selvam., MCA, M.Phil, Ph.D.,** Assistant Professor in Department of Computer Applications(PG), PSG College of Arts & Science, Coimbatore.

This project work has not been submitted by me for the award of any other Degree/ Diploma/ Associate ship/ Fellowship or any other similar degree to any other university.

PLACE : COIMBATORE **PRANESH M**

DATE : **(23MCA033)**

ACKNOWLEDGEMENT

My venture stands imperfect without dedicating my gratitude to a few people who have contributed a lot towards the victorious completion for my project work.

I would like to thank **Thiru. L. Gopalakrishnan, Managing Trustee, PSG & Sons Charities,** for providing me with the prospect and surroundings that made the work possible.

I express my deep sense of gratitude to **Dr.T.Kannaian, M.Sc., M.Tech., Ph.D.,** Secretary, PSG College of Arts & Science, Coimbatore for permitting and doing the needful towards the successful completion of this project.

I express my deep sense of gratitude and sincere thanks to our Principal In-charge **Dr. M. Senguttuvan, B.Sc., M.Sc., B.Ed., M.Phil., Ph.D.,** for his valuable advice and concern on students.

I am very thankful to **Dr.M.Umarani, M.B.A., MPhil., Ph.D., Vice-Principal, (Self Finance)** for her support and encouragement.

I kindly and sincerely thank **Dr.L.Thara, MCA., M.Phil., Ph.D., Associate Professor & HOD, Department of Computer Applications (PG)** for her whole-hearted help to complete this project successfully by giving valuable suggestions.

I convey my heartiest and passionate sense of thankfulness to my project guide, **Dr.Nithyanandh Selvam, Assistant Professor, Department of Computer Applications (PG)** for his timely suggestions which had enabled me to complete the project successfully.

This note of acknowledgement will be incomplete without paying my heartfelt devotion to my parents, my friends, and other people, for their blessings, encouragement, financial support and the patience, without which it would have been impossible for me to complete the job.

ABSTRACT

Monitoring water quality is vital for optimizing conditions for successful aquaculture, especially when it comes to the cultivation of spirulina. In keeping with Viridia Biotech's dedication to producing high-quality organic spirulina, this study uses machine learning techniques to forecast water temperature and monitor pond water quality metrics. The primary objective is to create a prediction algorithm that uses past data on water quality to predict temperature changes and provide the ideal conditions for spirulina development. To find significant associations, the methodology uses exploratory data analysis (EDA), feature construction with lagged variables, and thorough data preprocessing, which includes cleaning. To forecast water temperature based on key variables like pH, salinity, ambient illuminance, pond depth, and net oxygen dissolved (OD), a Random Forest Regressor model was trained on an 80-20 split dataset. Mean Absolute Error (MAE) and R2 score were used to assess the model's performance, and feature importance analysis revealed which parameters had the most influence.

The results show that the model can capture trends in water temperature variations with a low MAE and a relatively high R2 score, indicating reasonable accuracy. By improving its real-time monitoring and quality assurance procedures, Viridia Biotech can guarantee ideal pond conditions for the production of spirulina. The use of a comparatively limited dataset, the omission of outside variables like the weather, and the possible requirement for more complex feature engineering methods are some drawbacks. In order to further improve sustainability and prediction accuracy in spiralina farming, future improvements might include experimenting with sophisticated machine learning algorithms and adding more environmental variables.

TABLE OF CONTENT

1.INTRODUCTION

1.1 PROJECT OVERVIEW

The aim of project is to analyze the maintenance of the aquatic ecosystem by checking the water quality and it also helps us in the temperature prediction by using the random forest regression algorithm. The goal of Viridia Biotech is to provide ideal circumstances for the production of spirulina, and this research focuses on pond water quality measurement and temperature forecast. The temperature of the water has a significant effect on the growth of Spirulina and affects sustainability, quality, and production. Integrating machine learning-based predictive analytics can increase efficiency and proactive decision-making because traditional monitoring methods can be resource- and time-intensive.

In order to forecast temperature variations and examine important water characteristics, this study will use previous pond data gathered from Viridia Biotech's automated monitoring systems. The dataset comprises variables including oxygen dissolved (Net OD), pond depth, pH, salinity, and ambient illumination. By employing a Random Forest Regressor model and doing thorough exploratory data analysis (EDA), the project effectively forecasts trends in water temperature, which contributes to improved automated monitoring and sustainability. By utilizing data-driven decision-making, this project supports Viridia Biotech's dedication to quality and development. The model offers a foundation for future developments in predictive aquaculture management, and the results provide practical insights into the major variables affecting temperature variance.

1.2 COMPANY PROFILE

At Viridia Biotech, we specialize in cultivating premium organic Spirulina to meet the growing demands of the nutraceutical, food, and feed industries. Established in 2022 and backed by over 12 years of industry expertise, we have quickly earned a reputation for delivering products that combine superior quality, sustainability, and innovation.

Our fully automated production process ensures minimal handling, preserving the purity and freshness of our Spirulina. With a pond-to-powder conversion time of less than 15 minutes, rigorous quality monitoring of over 20 parameters, and state-of-the-art filtration and spray-drying systems, we guarantee consistent, top-tier products every time.

MISSION

To become a leading global supplier of Spirulina by offering unmatched quality, innovation, and sustainability while fostering trust and long-term partnerships.

VISION

To harness the power of Spirulina to enhance health and wellness worldwide while minimizing our environmental impact.

VALUES

We strive to produce the most premium Spirulina on the market, adhering to strict quality assurance practices. A research-driven approach ensures we are always finding new ways to improve our products and processes. We are committed to eco-friendly practices, including 100% water recycling and a low environmental footprint.

2.SYSTEM SPECIFICATION

2.1 HARDWARE SPECIFICATION

Development Environment

Processor (CPU)	: 11th Gen Intel® Core™
RAM	: 8 GB
Storage	: 512 GB(SSD)
Speed	: 2.40GHz

2.2 SOFTWARE SPECIFICATION

Operating System	: Windows 11 Home Single Language
Language	: Python 3.12.1
Develop platform	: Visual Studio Code
Packages	: Pandas, Seaborn, Matplotlib & Scikit-learn

2.3 SOFTWARE DESCRIPTION

Windows 11

Windows 11, a major update to the Microsoft Windows operating system, was painstakingly made for desktop and laptop computers and incorporates the touchscreen input used in contemporary Windows devices. This version provides a completely updated user interface that embodies the Fluent Design System's tenets. It features rounded corners, a centered taskbar and Start menu, and the sophisticated "Mica" material effect. The reduced design of the now-integrated Windows 10X components served as an inspiration for Windows 11, which offers a more sophisticated and unified user experience. This significant update places a high priority on a smooth and user-friendly computing environment that is designed to satisfy the changing needs of modern users.

Visual Studio Code

Microsoft carefully designed this source code editor, which is both lightweight and powerful, for developers working on a variety of platforms. By providing a rich and expandable environment through its extensive marketplace of extensions, this adaptable tool goes beyond conventional code editors. VS Code boasts a cutting-edge, user-friendly interface with integrated Git control, intelligent code completion, and powerful debugging features that are intended to simplify the coding process. This editor puts the productivity of developers first, facilitating smooth teamwork and effective workflow administration. Because of its open-source nature, which encourages a thriving community, it continues to evolve and adapt. The sophisticated and adaptable coding environment that Visual Studio Code offers is designed to satisfy the ever-changing needs of modern software development. Its performance-driven design, wide language support, and robust features are intended to enable developers to design, debug, and launch applications with previously unheard-of efficiency.

Packages

Pandas: It is a sophisticated and flexible Python framework for data manipulation and analysis. It makes it possible to manage structured data effectively by offering data structures like Data Frames and Series. It is a key component of data science processes since it makes tasks like data transformation, cleansing, and analysis easier.

Matplotlib: A complete Python package for making static, interactive, and animated visualizations is called matplotlib. With its extensive plotting features, users can create a variety of charts and graphs. It is an essential Python utility for data visualization.

Seaborn: It is a high-level Python visualization library. It provides a more intuitive and aesthetically beautiful interface for making statistical visuals. Seaborn makes it easier to explore and comprehend data relationships by streamlining the process of creating intricate visualizations.

Scikit-Learn: It offers a large selection of dimensionality reduction, clustering, regression, and classification algorithms. Because of its reliable and user-friendly interface, it is a well-liked option for machine learning jobs.

3. DATA COLLECTION

This project uses a dataset from Viridia Biotech's automated pond monitoring system that includes several water quality parameters that are essential for aquatic health analysis. Each data entry is timestamped (Date) to ensure accurate tracking of changes over time. The dataset records salinity (‰), which indicates the concentration of salt present, and pH, which measures the water's acidity or alkalinity; ambient illumination (lx) records the ambient light intensity, which can affect aquatic ecosystems; pond depth (cm) provides information on water levels, which may have an impact on other parameters; net OD (oxygen dissolved) is a crucial indicator of water quality and aquatic life sustainability; and the main target variable for prediction is water temperature (°C), a key factor influencing chemical and biological processes in the pond.

3.1 Data Acquisition

Key water quality parameters that are necessary for the best possible growth of Spirulina are regularly monitored by high-precision sensors placed across various pond locations. To maintain a steady atmosphere, pH sensors assess the concentration of hydrogen ions. To maintain appropriate conductivity levels, salinity sensors monitor the concentration of salt (‰). Photosynthesis is affected by ambient illuminance (lx), which is measured by light sensors (photometers/lux meters). While dissolved oxygen (DO) sensors evaluate oxygen availability, which is essential for spirulina metabolism, depth sensors give real-time water level readings. Temperature sensors assist control thermal conditions by tracking variations at various depths. Additionally, turbidity sensors can be used to monitor the quality of the water, which ensures optimal light penetration, and nutrient sensors can be used to evaluate important components such as phosphates and nitrates that encourage the growth of spirulina. Better management of water quality and increased spirulina output are made possible by these real-time data streams, which improve precision monitoring.

3.2 Data Logging and Transmission

Sensors in the monitoring system continually record data at predetermined intervals (e.g., every 15 minutes) and wirelessly send it to a centralized cloud-based storage system using an Internet of Things framework. Accurate temporal analysis is ensured by automatically time-stamped data entering. The cloud platform incorporates automatic quality control systems that manage missing values, identify anomalies, and indicate errors. By further analyzing trends, advanced data analytics and machine learning algorithms can anticipate any problems with water quality before they materialize. Remote monitoring and prompt decision-making are made possible by real-time dashboards, which offer visual information. In order to ensure prompt action and preserve the ideal circumstances for Spirulina growth, the system may also send operators alerts and notifications if crucial thresholds are surpassed.

3.3 Data Preprocessing

The raw data is preprocessed to ensure consistency, reliability, and analytical accuracy before analysis begins; column names are uniformly standardized by removing special characters; the Date column is transformed into a Date Time format to allow for precise time-based analysis; missing or corrupted values are identified and removed to preserve data integrity; the dataset is then sorted chronologically to preserve sequencing; feature engineering techniques, including rolling statistics, lagged variables, and moving averages, help capture temporal dependencies; outliers are identified using statistical methods and either corrected or removed to prevent skewed analysis; and the data is normalized or scaled where needed to ensure compatibility across various algorithms.

3.4 Seasonal Variability Considerations

The dataset captures seasonal variations in environmental conditions, which significantly influence Spirulina growth and water quality. Fluctuations in temperature between summer and winter impact metabolic activity, while seasonal shifts in light intensity affect photosynthesis. Changes in salinity levels due to evaporation or rainfall further alter water composition. Additionally, wind patterns can influence surface mixing, affecting nutrient distribution, while humidity variations may impact water loss through evaporation. By incorporating these seasonal effects into the analysis, Viridia Biotech can refine its pond management strategies, optimizing nutrient balance, aeration, and shading techniques. This proactive approach ensures stable growth conditions, enhances yield efficiency, and promotes long-term sustainability in Spirulina cultivation.

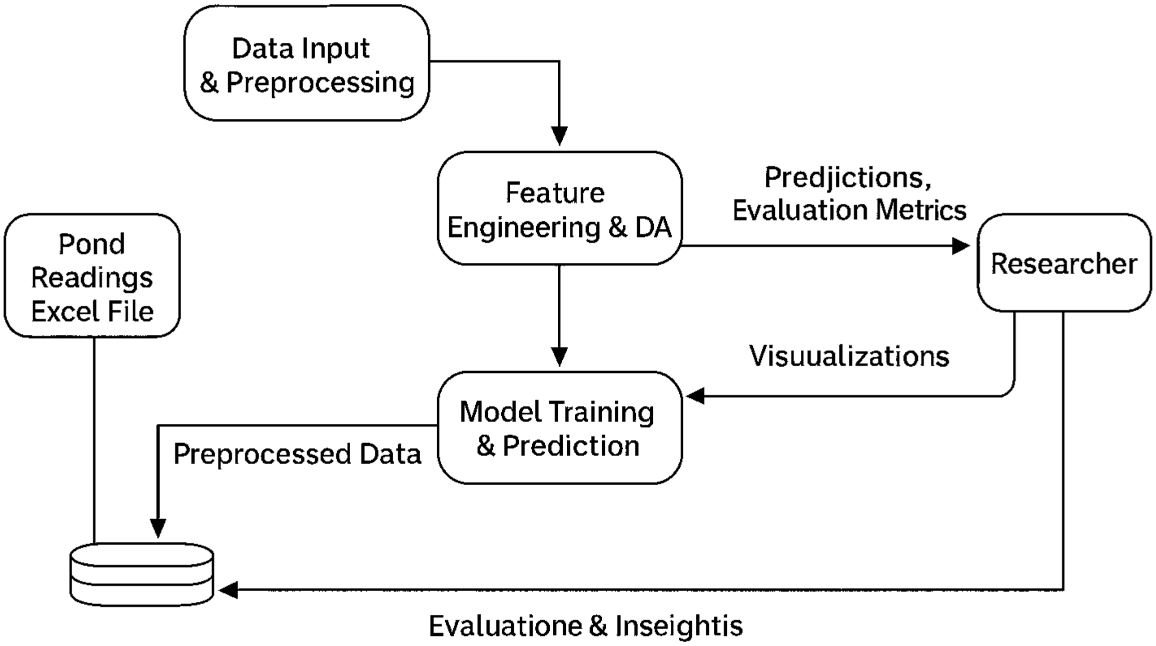
4. SYSTEM ARCHITECTURE

4.1 DATA FLOW DIAGRAM

DFD, or Data Flow Diagram, serves as a graphical representation of the flow of data within a system or process. It offers insights into the inputs, outputs, and interactions of each entity within the system, providing a high-level overview of data movement and transformations.

Unlike other diagrams, DFDs do not incorporate control flow, loops, or decision rules. Instead, they focus solely on illustrating the flow of data through various entities and processes. Specific operations and conditional logic dependent on the type of data can be better explained using flowcharts, which complement DFDs in depicting detailed process logic.

As a graphical tool, DFDs are invaluable for communication with stakeholders such as users, managers, and personnel involved in system development. They facilitate understanding and discussion of system requirements, processes, and interactions, making them an essential component of requirements analysis and system design.

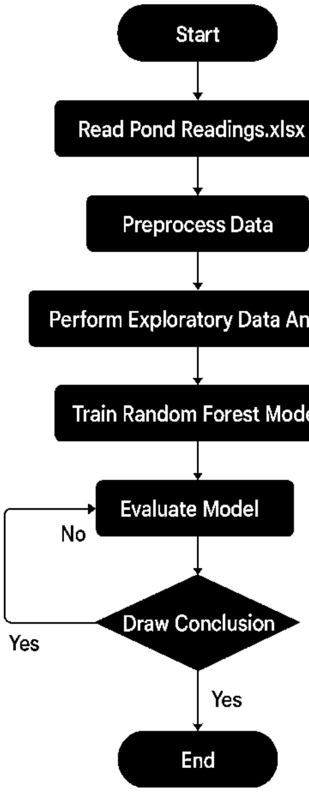


1. Overview of Data Flow: DFDs provide a comprehensive view of how data moves through the system, including its sources, transformations, and destinations.
2. Identification of Transformations: They highlight the various transformations performed on the data as it progresses through the system, aiding in understanding system functionality.
3. Storage of Data: DFDs illustrate where data is stored within the system, whether it's temporary storage during processing or permanent storage in databases or files.
4. Output Generation: They depict the results produced by the system, such as reports, notifications, or updates, giving stakeholders a clear understanding of system outputs.

DFDs can be represented in multiple ways, adhering to structured-analysis modelling principles. Their popularity stems from their ability to visually represent the major steps and data involved in software- system processes, making them invaluable tools for system analysis, design, and documentation.

4.2 SYSTEM FLOW DIAGRAM

An effective tool for process visualization is a System Flow Diagram (SFD), particularly in data-driven initiatives like pond water temperature prediction. It simplifies and clarifies complicated workflows by dividing them into organized steps. Being a universal language, it improves teamwork by facilitating more effective communication between technical and non-technical stakeholders. By pointing out bottlenecks, SFDs facilitate effective problem identification, troubleshooting, and optimization. Along with preserving process consistency and scalability, they also guarantee standardization and documentation, which project teams can use as a guide. SFDs also facilitate improved decision-making by making it simple to assess alternative approaches and assisting in the creation of upcoming improvements. They enable smooth upgrades without interfering with workflow by facilitating modular development, which makes system design and maintenance easier to handle. Furthermore, SFDs assist in risk assessment, ensuring that potential failures are identified early, and improve resource management by optimizing time and effort across different stages of the project.



1. By condensing intricate procedures into understandable visual representations, system flow diagrams improve stakeholder comprehension and communication.
2. They help streamline workflows and plan new procedures, function as useful teaching and documentation tools, and enable effective problem-solving by locating bottlenecks and inefficiencies.
3. They also encourage standardization by offering a common visual language that guarantees all participants understand the procedure in the same way.
4. Additionally, flowcharts make process analysis and optimization easier, allowing businesses to pinpoint problem areas and adjust boost productivity and cut expenses.
5. Additionally, by recording process stages and decision points—which are essential for regulatory compliance and quality control—they assist audit trails and compliance.

5. EXPLORATORY DATA ANALYSIS

The essential first step in analyzing and summarizing datasets to comprehend their features, trends, and potential problems is called exploratory data analysis (EDA). It uses a combination of graphical and non-graphical techniques, including univariate, bivariate, and multivariate analysis, as well as data visualization and descriptive statistics, to accomplish goals like comprehending data structure, spotting trends, identifying outliers, evaluating data quality, and directing additional analysis. In the end, this builds a deeper understanding of the data and enhances the precision and dependability of subsequent modelling and decision-making.

5.1 Correlation Analysis:

1. Correlation analysis is used to quantify and illustrate the linear correlations between pairs of variables in a dataset, especially when a correlation matrix is created and displayed as a heatmap.
2. The heatmap offers a visually intuitive depiction, with color intensity reflecting the extent of association, while the correlation matrix shows correlation coefficients, which range from -1 to +1, showing the strength and direction of these relationships.
3. In the heatmap, significant correlations are indicated by bright color changes. These correlations imply strong linear dependencies between features, which can be important for comprehending how variables interact and guiding further modelling or analysis.

5.2 Time-Series Analysis:

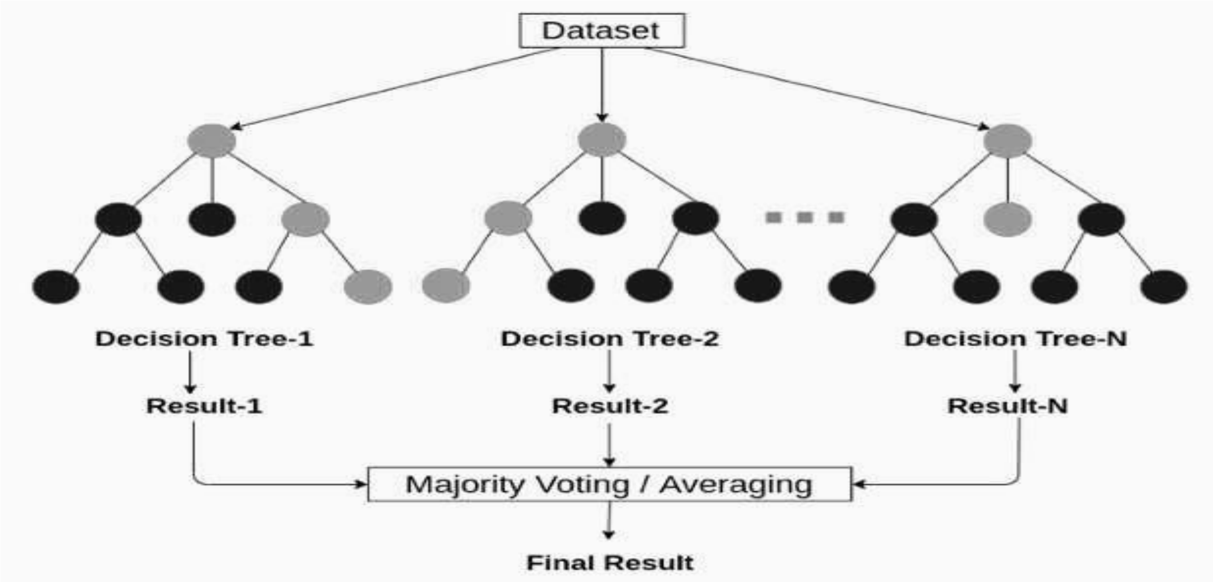
1. In order to observe and analyze trends of water temperature, pH, salinity, and net OD over time, time-series analysis was performed. This resulted in plots that show fluctuations and patterns that are essential for comprehending how seasonal and environmental variations affect the cultivation of spirulina.
2. Through the capture of the temporal history of variables and their interactions, this study seeks to discern temporal patterns, such as seasonality, cyclic variations, long-term trends, and possible outliers.
3. Additionally, it evaluates autocorrelation and stationarity, which eventually helps with forecasting to ensure initiative-taking reactions to sustain Spirulina development and yield by predicting future environmental changes and optimizing crop management.

6. METHODOLOGIES

6.1 Random Forest Regression

Random forest regression is a potent machine-learning technique that combines several decision trees into an ensemble model, allowing for precise predictions and the study of complicated datasets. The ultimate output is calculated by averaging or taking a weighted average of all the predictions made by each tree, which is trained on a distinct subset of data and makes separate predictions. This technique reduces overfitting by employing independent decision trees, which improves generalization, and it gives high accuracy by classifying data points based on several attributes. Real-world datasets with gaps or corrupted data can benefit from the effective handling of complicated interactions and non-linear relationships by random forests.

They also offer automatic feature selection, measuring each feature’s contribution to improve accuracy. The algorithm is highly robust, immune to the dimensionality curse, and supports parallelization, allowing for fast processing of large datasets. Unlike linear regression or support vector machines, random forests require fewer parameters and minimal tuning while still maintaining computational efficiency. Additionally, the ensemble approach improves stability, as the final outcome is based on multiple decision trees, reducing bias and variance. Their ability to work with high-dimensional data while providing reliable and automated results makes them an excellent choice for predictive modeling in various applications.



6.2 Model for Real-Time Water Temperature Forecasting

1. The dataset was divided into 80% training and 20% testing sets, and a Random Forest Regressor—selected for its capacity to manage noise and non-linear relationships—was trained to predict water temperature (°C) using lagged features of pH, salinity, ambient illuminance, pond depth, and Net OD.
2. Mean Absolute Error (MAE) and R2 score were used to assess the model's performance; it showed minimal prediction errors and a reasonably high capacity to capture temperature fluctuations.
3. A scatter plot comparing actual versus predicted temperatures validated the model's reliability, demonstrating close alignment and validating its potential for real-time automated monitoring systems.
4. A bar chart was used to visualize feature importance in order to understand variable influence and aid in monitoring protocol refinement.

7. FUTURE ENHANCEMENT

Several enhancements can be made to the current pond water quality prediction model in the areas of data preprocessing, modelling, and deployment. Initially, rolling statistics, several lag variables, and time-based features such as month, day of the week, and seasonality can be added to feature engineering in order to capture recurring patterns. Including wavelet and Fourier transformations can improve the modelling of seasonal fluctuations. Plotly or Streamlit interactive dashboards, dynamic correlation heatmaps across time, and anomaly identification using box plots, Z-score, or Isolation Forests are examples of advanced visualization approaches that can provide deeper insights into water parameters. While Bayesian Optimization, Grid Search, and Randomized Search can assist in optimizing hyperparameters, it is possible to enhance prediction accuracy in modelling by experimenting with different algorithms such as XGBoost, LightGBM, CatBoost, SVR, ARIMA, and LSTMs.

In order to successfully implement the concept, real-time user interaction can be facilitated by a web application that uses Flask, Fast API, or Streamlit. Automation through IoT-based sensors can also provide continuous monitoring, with data stored in cloud-based Firebase or NoSQL databases like MongoDB for real-time access. Incorporating environmental conditions into the model by integration with satellite-based remote sensing data and external meteorological APIs can improve its forecasting ability. Real-time monitoring can also be made possible by installing the model on edge devices or microcontrollers like the Raspberry Pi or NVIDIA Jetson, which eliminates the need for cloud processing entirely. By employing real-time drift detection methods such as Adaptive Windowing (ADWIN) or Kolmogorov-Smirnov tests, the model can be kept accurate even when new data patterns appear.

The system can become more scalable and accessible by deploying on cloud platforms such as AWS Lambda, Google Cloud AI, or Azure ML. Actionable insights for pond maintenance can also be obtained by integrating automatic alert systems via SMS, email, or IoT-driven actuators, as well as reinforcement learning-based optimisation for water quality management.

8. CONCLUSION

This study effectively illustrated the use of machine learning techniques for pond water quality analysis and temperature prediction. Using a Random Forest Regressor and historical data, the study offered important insights into the main variables affecting variations in water temperature. These results are particularly important for Viridia Biotech since they will improve real-time monitoring, maintain high-quality spirulina production, and optimise pond conditions.

The model's performance was good, as evidenced by its low Mean Absolute Error (MAE) and respectably high R2 score, which suggested that it was good at capturing changes in water temperature. Furthermore, feature importance analysis emphasised the major influence of factors including pond depth, ambient illumination, pH, and salinity on regulating pond temperature.

The study did have some limitations, though, including a relatively small dataset size and the exclusion of external environmental factors that could affect predictive accuracy, such as wind speed, solar radiation, and weather conditions. Other feature engineering techniques could be investigated to further improve model performance. Incorporating larger datasets for better model training and generalization; investigating cutting-edge machine learning algorithms like XGBoost, Support Vector Machines (SVM), or Deep Learning models; incorporating external factors like weather conditions and seasonal changes; and creating a real-time monitoring dashboard for pragmatic application in spirulina farming.

BIBLIOGRAPHY

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.
7. "Random forests for classification in ecology" by Andy Liaw and Matthew Wiener. Published in *Ecology*, 2002.
8. "A brief introduction to random forests for genomic data analysis" by Veronique G.B. Chevalier. Published in *Advances in Genomics and Genetics*, 2012.
9. "Variable selection using random forests" by Gökhan Kursa, Witold Rudnicki, and others (2010) Published in: Journal of Statistical Software, Volume 36, Issue 11, pages 1–13.
10. "Random forests for global sensitivity analysis: A selective review" by Bertrand Iooss and Philippe Lemaitre (2015) *Published in: Environmental Modelling & Software*, Volume 72, pages 219–239.

APPENDIX

This section contains the sample code for the data analysis process by using the random forest regressor algorithm and their results.

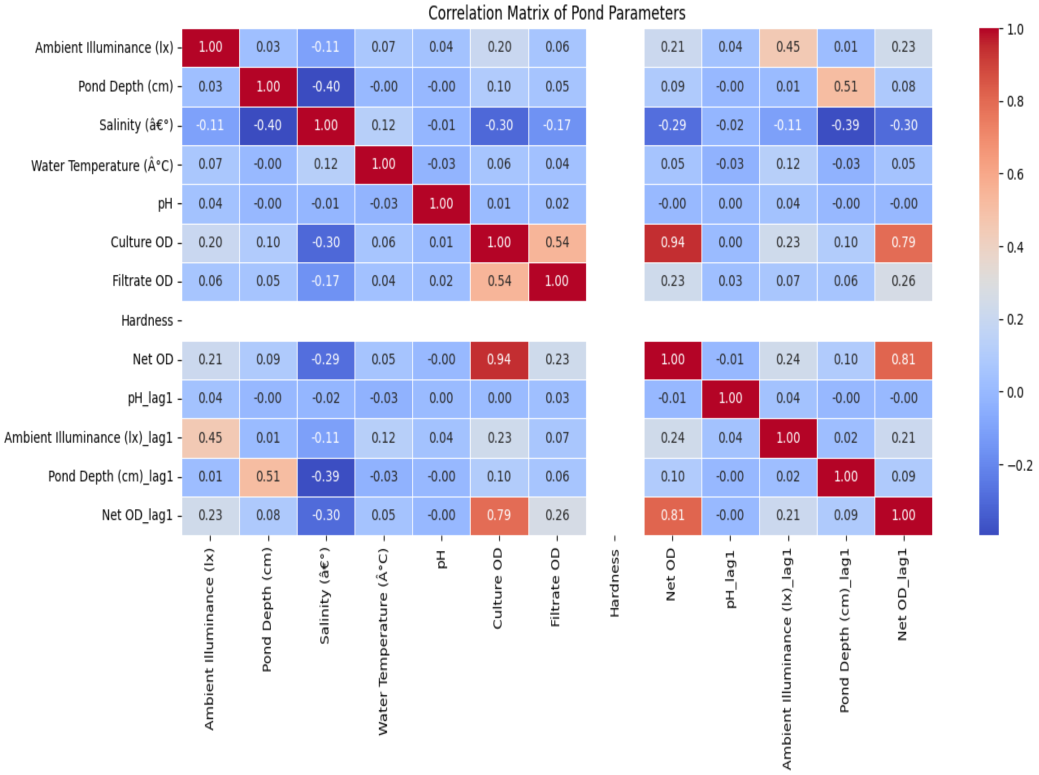
A.SOURCE CODE:

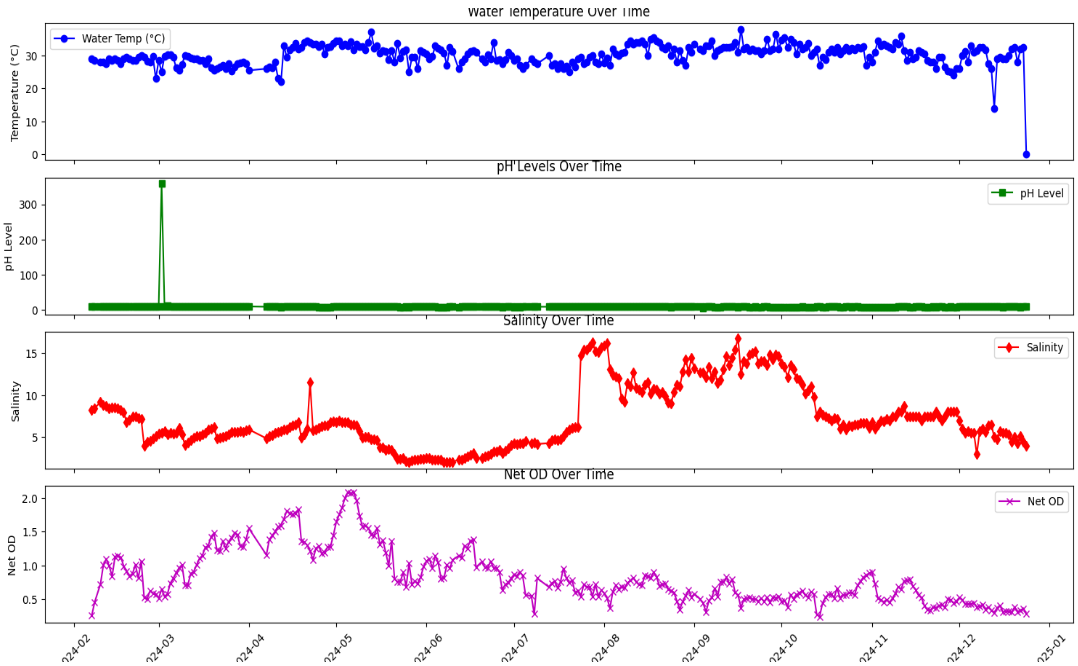
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error, r2_score

file_path = r"C:\Users\user\Documents\pranesh\OneDrive\Desktop\Pond Readings.xlsx"
df = pd.read_excel(file_path)

df.columns = df.columns.str.strip()

df = df.rename(columns={
    "Salinity (â€°)": "Salinity",
    "Water Temperature (Â°C)": "Water Temperature (°C)"
})

df["Date"] = pd.to_datetime(df["Date"])
df = df.dropna()
df = df.sort_values("Date")

lag_features = ["pH", "Salinity", "Ambient Illuminance (lx)",
                "Pond Depth (cm)", "Net OD", "Water Temperature (°C)"]

for feature in lag_features:
    if feature in df.columns:
        df[f"{feature}_lag1"] = df[feature].shift(1)
    else:
        print(f"Warning: {feature} not found in dataset")

df = df.dropna()

corr_matrix = df.select_dtypes(include=['number']).corr()
plt.figure(figsize=(10,6))
sns.heatmap(corr_matrix, annot=True, cmap="coolwarm", fmt=".2f", linewidths=0.5)
plt.title("Correlation Matrix of Pond Parameters")
plt.show()

df_daily = df.groupby("Date")[df.select_dtypes(include=["number"]).columns].mean()

fig, axes = plt.subplots(4, 1, figsize=(12, 16), sharex=True)
axes[0].plot(df_daily.index, df_daily["Water Temperature (°C)"], marker='o', linestyle='-', color='b', label="Water Temp (°C)")
axes[0].set_ylabel("Temperature (°C)")
axes[0].set_title("Water Temperature Over Time")
axes[0].legend()

axes[1].plot(df_daily.index, df_daily["pH"], marker='s', linestyle='-', color='g', label="pH Level")
axes[1].set_ylabel("pH Level")
axes[1].set_title("pH Levels Over Time")
axes[1].legend()

axes[2].plot(df_daily.index, df_daily["Salinity"], marker='d', linestyle='-', color='r', label="Salinity")
axes[2].set_ylabel("Salinity")
axes[2].set_title("Salinity Over Time")
axes[2].legend()

axes[3].plot(df_daily.index, df_daily["Net OD"], marker='x', linestyle='-', color='m', label="Net OD")
```

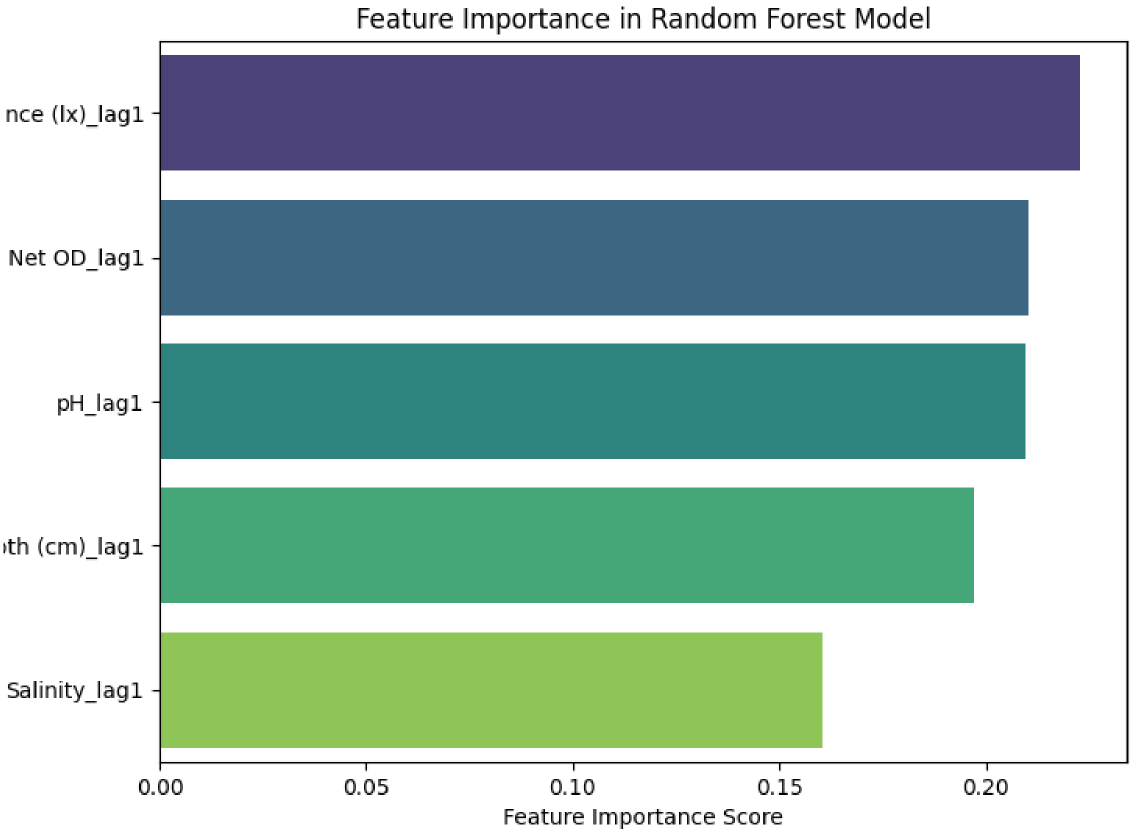
```
axes[3].set_ylabel("Net OD")
axes[3].set_title("Net OD Over Time")
axes[3].legend()

plt.xlabel("Date")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()

features = [f"{feature}_lag1" for feature in lag_features if feature != "Water Temperature (°C)"]
target = "Water Temperature (°C)"

missing_features = [f for f in features if f not in df.columns]
if missing_features:
    print(f"Warning: Missing Features -> {missing_features}")
    features = [f for f in features if f in df.columns]

X_train, X_test, y_train, y_test = train_test_split(df[features], df[target], test_size=0.2, random_state=42)

rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)

y_pred = rf_model.predict(X_test)

mae = mean_absolute_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print(f"Mean Absolute Error: {mae:.2f}")
print(f"R² Score: {r2:.2f}")

feature_importance = pd.Series(rf_model.feature_importances_,
index=features).sort_values(ascending=False)
plt.figure(figsize=(8,6))
```

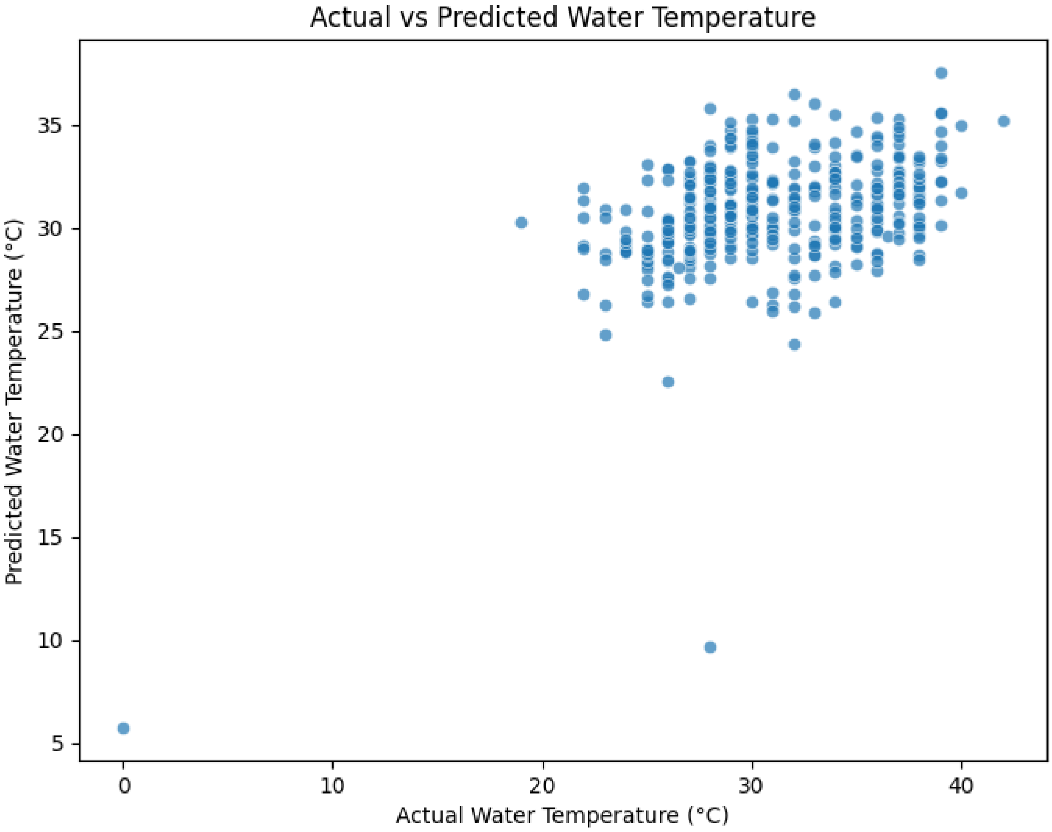
```
sns.barplot(x=feature_importance, y=feature_importance.index, palette="viridis")
plt.xlabel("Feature Importance Score")
plt.ylabel("Features")
plt.title("Feature Importance in Random Forest Model")
plt.show()

plt.figure(figsize=(8,6))
sns.scatterplot(x=y_test.values, y=y_pred, alpha=0.7)
plt.xlabel("Actual Water Temperature (°C)")
plt.ylabel("Predicted Water Temperature (°C)")
plt.title("Actual vs Predicted Water Temperature")
plt.show()
```

B.OUTPUT:




Water Temperature Over Time
Water Temp (°C)
Temperature (°C)
0
10
20
30
pH Levels Over Time
pH Level
pH Level
0
100
200
300
Salinity Over Time
Salinity
Salinity
5
10
15
Net OD Over Time
Net OD
Net OD
0.5
1.0
1.5
2.0
024-02
024-03
024-04
024-05
024-06
024-07
024-08
024-09
024-10
024-11
024-12
025-01


Feature Importance in Random Forest Model
nce (lx)_lag1
Net OD_lag1
pH_lag1
th (cm)_lag1
Salinity_lag1
0.00
0.05
0.10
0.15
0.20
Feature Importance Score


Actual vs Predicted Water Temperature
Predicted Water Temperature (°C)
35
30
25
20
15
10
5
0
10
20
30
40
Actual Water Temperature (°C)