**+Bio-Crude oil Prediction with Microalgae by using Machine Learning**

Nalluri Rishi Chaitanya Sri Prasad1, Rohith Sai Rayapati1, Idupulapati Revanth1, Maddi Koushik1, Prabakaran Ganeshan2, Karthik Rajendran2 \*

1 Department of Computer Science and Engineering, School of Engineering and Sciences, SRM University-AP, Amaravati, Andhra Pradesh 522240, India.

2 Department of Environmental Science, School of Engineering and Sciences, SRM

University-AP, Amaravati, Andhra Pradesh 522240, India.

Corresponding author Email: [rajendran.k@srmap.edu.in](mailto:rajendran.k@srmap.edu.in)

# Highlights

# Abstract

# Keywords

# Introduction

# Methodology

## Data Collection and pre-processing

**Table 1.** Detailed Input features and the ranges (min-max) of each feature in different datasets.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Items |  | Dataset 1 | Dataset 2 | Dataset 3 | Dataset 4 |
| *Input features* | | | | | |
| Elemental  Composition | Carbon (%) | - | 32.3-76.2 | 32.3-76.2 | 32.3-76.2 |
| Hydrogen (%) | - | 3.80-11.4 | 3.80-11.4 | 3.80-11.4 |
| Nitrogen (%) | - | 0.80-8.70 | 0.80-8.70 | 0.80-8.70 |
| Oxygen (%) | - | 6.30-56.2 | 6.30-56.2 | 6.30-56.2 |
| Sulfur (%) | - | 0.00-3.50 | 0.00-3.50 | 0.00-3.50 |
| Biochemical composition | Protein (%) | - | - | 7.30-66.0 | 7.30-66.0 |
| Lipid (%) | - | - | 3.50-23.0 | 3.50-23.0 |
| Carbohydrate (%) | - | - | 11.00-51.9 | 11.00-51.9 |
| Operational conditions | Time (min) | 30.0-60.0 | 30.0-60.0 | 30.0-60.0 | 30.0-60.0 |
| Temperature (°C) | 150-420 | 150-420 | 150-420 | 150-420 |
| Pressure (bar) | 50.0-221 | 50.0-221 | 50.0-221 | 50.0-221 |
| Proximate Analysis | Moisture (%) | 3.70-13.50 | - | - | 3.70-13.50 |
| Volatile (%) | 51.40-81.40 | - | - | 51.40-81.40 |
| Ash (%) | 6.40-36.00 | - | - | 6.40-36.00 |
| *Output Response* | | | | | |
| Properties | Oil yield (%) | 11.60-65.00 | 11.60-65.00 | 11.60-65.00 | 11.60-65.00 |
| HHV (MJ/kg) | 10.90-38.20 | 10.90-38.20 | 10.90-38.20 | 10.90-38.20 |
| ERR (%) | 6.40-79.40 | 6.40-79.40 | 6.40-79.40 | 6.40-79.40 |

HHV -High Heating Value; ERR – Energy Recovery Ratio.

## Machine Learning Algorithms

## Hyperparameter optimization

**Table 2**. Hyperparameter Tuning

|  |  |  |
| --- | --- | --- |
| Algorithm | Hyperparameter Function | Optimized value |
| Random forest | Method | Bagging |
|  | n\_estimators | 25 |
|  | max\_depth | 10 |
| Decision Trees | max\_depth | 10 |
|  | min\_samples\_split | 60 |
|  | min\_samples\_leaf | 10 |
| Linear regression | Learning Rate | 0.001 |

## Feature Selection

### Pearson correlation

### Machine learning method

## Model evaluation

# Results and discussion

## Predictive analysis of ML model

### Actual vs predicted analysis

Calendar

Description automatically generated

**Fig 1**. Yield prediction plots(train and test) of optimal RF developed from dataset #1 (a-c), Dataset #2 (d-f) and dataset #3 (g-i) and dataset #4 (j-l).

### Best model selection procedure for prediction

* These algorithms are used for predicting the target value by taking all the input features. The best and efficient algorithm for predicting the Oil yield(%)
* In this diagram if the R2 is >0.6 then it is considered to the next step
* The RMSE values must be less then we can say that is efficient one, so the decision trees and random forest is filtered next step
* comparing with MAE values it also must be less then dataset #3 is more efficient the most efficient algorithm for this dataset is Random Forest.

Diagram

Description automatically generated

**Fig 2**. Algorithm comparison for prediction

### Hyperparameter tuning

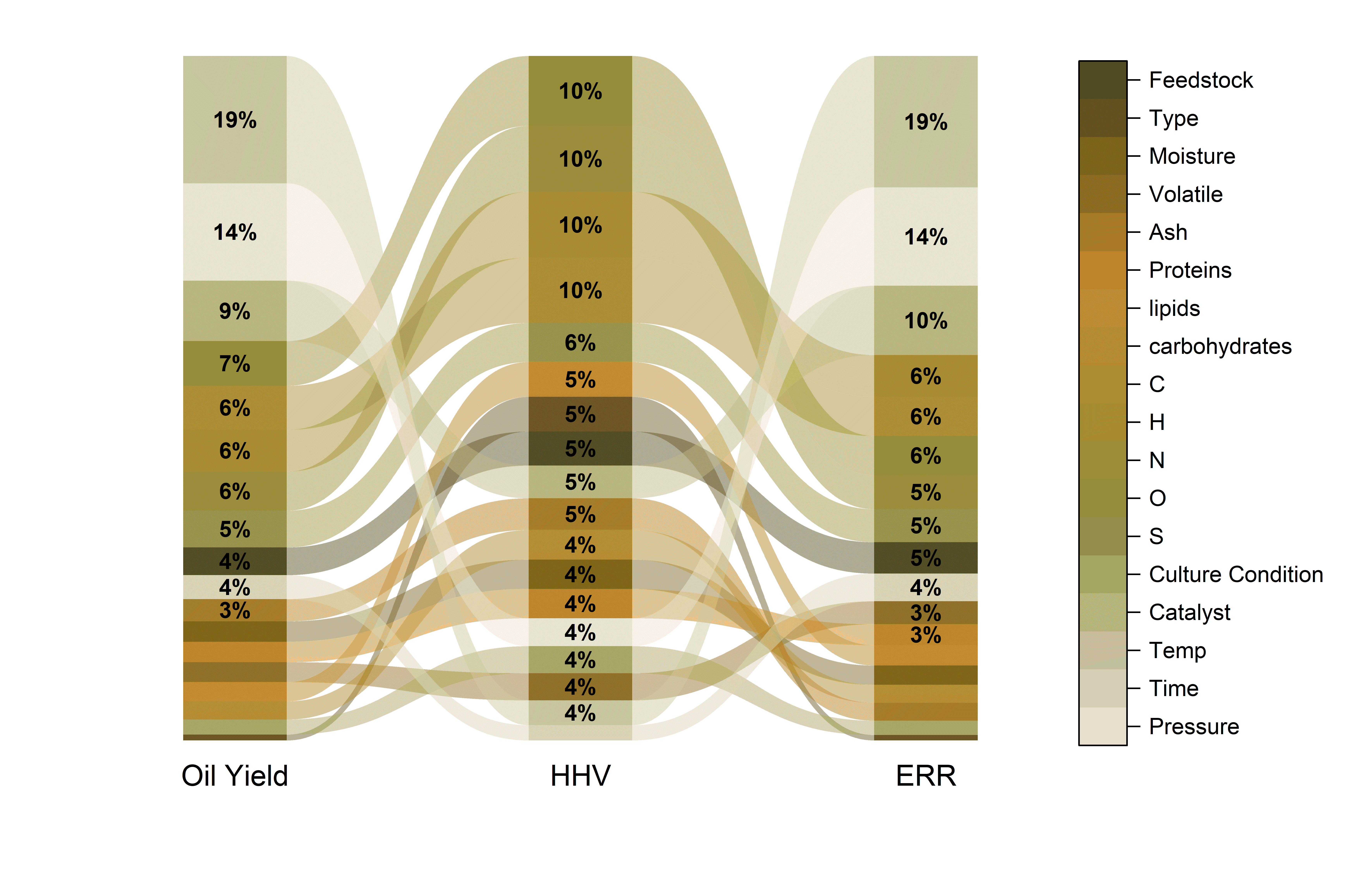
**Table 3.** Detailed values of Optimal model, R2, RMSE, MAE for all datasets

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data set** | **No. of features** | **Target** | **Algorithm** | **Train** | | | **Test** | | |
| **R2** | **RMSE** | **MAE** | **R2** | **RMSE** | **MAE** |
| 1 | 6 | Oil yield | RF | 0.90 | 0.004 | 2.21 | 0.84 | 0.65 | 2.89 |
| DT | 0.81 | 10.6 | 6.2 | 0.60 | 17.1 | 10.4 |
| LR | 0.13 | 8.33 | 6.14 | 0.10 | 0.51 | 4.71 |
| HHV | RF | 0.92 | 0.03 | 0.87 | 0.87 | 0.29 | 1.65 |
| DT | -0.33 | 146.0 | 12.1 | -0.55 | 146.0 | 12.2 |
| LR | -0.18 | 1.66 | 4.17 | -0.33 | 0.22 | 4.22 |
| ERR | RF | 0.83 | 0.51 | 2.81 | 0.75 | 0.41 | 3.23 |
| DT | 0.86 | 1.86 | 5.53 | 0.48 | 0.28 | 11.0 |
| LR | 0.35 | 2.72 | 5.73 | -0.52 | 2.01 | 7.53 |
| 2 | 10 | Oil yield | RF | 0.96 | 0.0003 | 1.86 | 0.82 | 3.30 | 2.57 |
| DT | 0.82 | 4.54 | 6.73 | 0.61 | 0.001 | 9.52 |
| LR | 0.40 | 3.75 | 5.47 | -0.28 | 0.07 | 7.81 |
| HHV | RF | 0.99 | 0.01 | 0.28 | 0.99 | 0.007 | 0.25 |
| DT | -0.12 | 136.1 | 12.07 | -0.45 | 104.9 | 12.1 |
| LR | 0.99 | 6.74 | 0.26 | 0.99 | 0.02 | 0.22 |
| ERR | RF | 0.88 | 0.05 | 2.37 | 0.79 | 1.30 | 3.72 |
| DT | 0.91 | 0.75 | 4.45 | 0.68 | 1.14 | 7.44 |
| LR | 0.30 | 4.26 | 6.21 | 0.24 | 3.11 | 7.18 |
| 3 | 13 | Oil yield | RF | 0.91 | 0.0001 | 1.53 | 0.78 | 1.55 | 2.92 |
| DT | 0.82 | 1.64 | 6.99 | 0.63 | 0.53 | 9.57 |
| LR | 0.30 | 1.98 | 5.77 | 0.29 | 0.72 | 5.95 |
| HHV | RF | 0.99 | 0.0001 | 0.21 | 0.98 | 2.16 | 0.45 |
| DT | -0.15 | 135.7 | 12.2 | -0.21 | 90.5 | 12.7 |
| LR | 0.99 | 1.38 | 0.09 | 0.97 | 0.06 | 0.38 |
| ERR | RF | 0.88 | 0.03 | 2.17 | 0.74 | 0.19 | 3.46 |
| DT | 0.88 | 0.04 | 5.10 | 0.79 | 0.14 | 6.25 |
| LR | 0.40 | 7.74 | 6.43 | 0.29 | 3.66 | 7.36 |
| 4 | 17 | Oil yield | RF | 0.94 | 0.01 | 1.66 | 0.80 | 0.24 | 2.77 |
| DT | 0.79 | 10.1 | 7.32 | 0.66 | 2.67 | 6.17 |
| LR | 0.56 | 1.17 | 4.34 | 0.47 | 0.50 | 6.11 |
| HHV | RF | 0.99 | 0.003 | 0.24 | 0.99 | 0.01 | 0.37 |
| DT | -0.17 | 131.1 | 11.7 | -0.63 | 190.7 | 14.1 |
| LR | 0.99 | 3.86 | 0.21 | 0.99 | 0.001 | 0.31 |
| ERR | RF | 0.91 | 0.22 | 2.23 | 0.79 | 1.22 | 2.22 |
| DT | 0.91 | 0.97 | 4.50 | 0.82 | 1.94 | 6.83 |
| LR | 0.62 | 4.06 | 5.13 | 0.32 | 6.66 | 7.48 |

HHV -High Heating Value; ERR – Energy Recovery Ratio; RF – Random Forest; DT – Decision Tree; LR – Linear Regression; RMSE – Root Mean Square Error; MAE – Mean Absolute Error.

## Feature determination for bio-yield prediction

### ML based feature importance analysis

****

**Fig 3.** Feature importance graphs

|  |  |  |
| --- | --- | --- |
| Oil yield | HHV | ERR |
| Temperature | O | Temperature |
| Pressure | N | Pressure |
| Catalyst | H | Catalyst |
| O | C | H |
| C | S | C |
| H | Lipids | O |
| N | Type | N |

### Pearson Correlation

A picture containing calendar

Description automatically generated

**Fig 4**. Data correlation heatmap

A picture containing window, open

Description automatically generated

Figure 5 Bio-oil yield VS All the input features

# Conclusion

**Table 4**. Predicting yield% with New Data

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset 4** | True Value | Predicted Value | Difference (%) |
| 28.0 | 31.0 | 9.7 |
| 27.1 | 30.4 | 10.9 |
| 22.0 | 21.8 | -0.9 |
| 34.0 | 33.1 | -2.7 |
| 38.5 | 34.0 | -13.2 |
| **Dataset 1** | 28.0 | 30.4 | 7.9 |
| 27.1 | 33.6 | 19.3 |
| 22.0 | 21.7 | -1.4 |
| 34.0 | 36.3 | 6.3 |
| 38.5 | 28.4 | -35.6 |

Chart, line chart

Description automatically generated

Figure 6 Probability plot of the different cases validated against the predicted result.

# References