Chapter 3

Decision tree analysis for predicting the best combination of predictor variables for increasing wastewater treatment capability and biomass production (generalized class)

Decision tree analysis for predicting the best combination of predictor variables for increasing wastewater treatment capability and biomass productio

were further categorized into four broad groups: (i) microalgae type; (ii) pre-cultivation stage deciding factors; (iii) nutrient content of wastewater; and (iv) operating variables. As a preliminary data preprocessing step, all missing values were removed and only valid values were used for creating the dataset. Units of all the predictor and response variables were also standardized (Table 3.1), and the hot encoding technique encoded categorical variables. Response/Output variables (nitrogen removal efficiency (NRE), phosphorus removal efficiency (PRE), and biomass production) were grouped into two categories: (i) high level, and (ii) low level, taking the median as the threshold value for preventing class imbalance during the classification process [417]. Values above the threshold level were grouped into high-level class and values below it were grouped into low-level class. The median value of NRE, PRE and biomass production was 83.93%, 88.70% and 1.22 g/L, respectively.

Table 3.1. Predictor variables used during analysis and their range.

S.No.	Predictor variables	Ranges for continuous variables or identities for categorical variables
1	Microalgae Class	*Chlorodendrophyceae, Chlorophyceae, Consortium, Cyanophycean, Eustigmatophyceae, Trebouxiophyceae Xanthophyceae*
2	Wastewater source	Domestic, Municipal, Agriculture, Industrial, Synthetic, Livestock
3	Cultivation type	Autotrophic, Heterotrophic, Mixotrophic
4	Culture type	Microalgae only, Microalgae-Bacteria Consortia
5	Pre-treatment method	Primary, Aerobic, Anaerobic, Autoclave
6	Reactor Type	Flasks, PBR
7	CO_2 Content	0.03-45% (v/v)
8	Temperature	9.3-35 °C
9	Cultivation Period	2-45 days
10	Initial Inoculum Level	0.01-2.3 g/L
11	pH	4.86-9.75
12	Light Intensity	0-4033 μmol m^{-2} s^{-1}

13	Photoperiod	0-24 h
14	Reactor Volume	100-500000 L
15	Initial Nitrogen Content	0.11-3410 mg/L
16	Initial Phosphorus Content	0.01-998.75 mg/L

PBR: Photobioreactor

3.2.2. Correlation Analysis

Pearson correlation coefficient was applied for detecting the linear correlation between the predictor variables and response variables as shown in Eq. (3.1):

$$\boldsymbol{r} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\sum_{i=1}^{n}(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \quad (3.1)$$

where, r denotes correlation coefficient, $\bar{x}$ and $\bar{y}$ are the corresponding mean values of x and y, and n is the total number of datasets. First, the correlation coefficient's significance level was calculated from Eq. (3.2), and then the p-value was calculated according to obtained two-tailed distribution.

$$\boldsymbol{t} = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \quad (3.2)$$

The correlation coefficient and t-test predicted the significant predictor variables affecting the response variables as a preliminary step. Predictor values showing no correlation with response variables were omitted during the training of classification models [46].

3.2.3. Decision tree classification and validation

Different combinations of predictor variables leading to high and low levels of response variables were determined by applying the decision tree classification algorithm. The decision tree algorithm works by interpolating the learned knowledge from the dataset in the form of a tree, where the tree's growth is governed by if-then rules [418]. MATLAB (version 9.2; R2017a) was used to perform decision tree classification. The "fitctree" function was used for building decision trees with the default CART algorithm. As a preliminary step for building

optimal tree structure, the hyperparameters, including split criterion, MaxNumSplits, MinLeafSize, and NumVariablesToSample, were optimized by using the "OptimizeHyperparameters" option keeping value 'all'. OptimizeHyperparameters option works by minimizing the cross-validation error.

In order to build a generalized tree that can classify any new data with high accuracy, training and testing procedures were performed by splitting the data randomly into two subsets: (i) training subsets (4/5 of the total data); (ii) testing subset (1/5 of the total data). Splitting of the dataset was performed by the 'cvpartition' function and 'HoldOut' validation procedure. At first, the training subset was used for constructing different tree structures. During the training step, the algorithm computes the Gini index of each attribute or input variable in order to make selection of root node. The attribute with lowest Gini index was then selected as root node. The process is continued for selection of further parent and lead nodes. Thereafter, the tree with the lowest cross-validation error was selected to test the data from the testing subset and predict the accuracy. No information from the testing subsets was used during the training step of the decision tree. The most influential predictor variables leading to different levels of response variables were determined by the "predictorImportance" function of MATLAB. After finding the optimal tree structure, the final validation of the models was done by testing them on experimental findings reported in recently published research papers [419]–[424].

3.3. Results and Discussion

3.3.1. Correlation Analysis

The correlation coefficient (R) values obtained between predictor and response variables after correlation analysis has been reported in Table 3.2. After correlation analysis, predictor variables showing no correlation with response variables were removed, including the cultivation period, reactor volume, type of light source, and wavelength.

Table 3.2. Correlation coefficients between predictor and response variables.

Predictor Variables	Cultivation Mode		
	Autotrophic	**Heterotrophic**	**Mixotrophic**
Response Variables			
NRE	0.1352	-0.23	-0.01189
PRE	0.0123	-0.13	0.06742
Biomass	-0.0555	-0.08	0.10667

Predictor Variables	Pre-treatment Method			
	Aerobic	**Anaerobic**	**Autoclave**	**Primary**
Response Variables				
NRE	0.06348	-0.1730	-0.0793	0.2266
PRE	0.22806	-0.0955	0.0714	0.1593
Biomass	0.01731	0.0351	-0.1000	0.0416

Predictor Variables	Wastewater Source					
	Agriculture	**Domestic**	**Industrial**	**Livestock**	**Municipal**	**Synthetic**
Response Variables						
NRE	0.0285	-0.1703	-0.1301	-0.0167	0.1020	0.1726
PRE	-0.0145	-0.0588	0.0347	0.1058	-0.2573	0.1976
Biomass	-0.0785	-0.1599	0.2163	0.0501	-0.0489	-0.0852

System	**Reactor**
-0.0327	0.0375
0.0505	0.0553
0.1042	0.3680

Predictor Variables	Other Variables	
	System	**Reactor**
Response Variables		
NRE	-0.0327	0.0375
PRE	0.0505	0.0553
Biomass	0.1042	0.3680

Predictor Variables	Microalgae Class						
	Chlorodendrophyceae	***Chlorophyceae***	**Consortium**	***Cyanophyceae***	***Eustigmatophyceae***	***Trebouxiophyceae***	***Xanthophyceae***
Response Variables							
NRE	0.0205	0.1391	-0.0939	0.0193	0.0485	-0.1251	0.0951
PRE	0.0095	0.0106	-0.0146	0.0230	0.0866	-0.0492	0.0386
Biomass	-0.0371	-0.0846	0.1516	-0.0370	0.0233	-0.1267	0.4073

Predictor Variables	**Nutrient Conc.**	
	Initial N Conc. (mg/L)	**Initial P Conc. (mg/L)**
Response Variables		
NRE	-0.1466	-0.1194
PRE	0.0219	-0.1174
Biomass	0.0456	-0.0026

Predictor Variables	**Operating Variables**							
	CO_2 Content %(v/v)	**Temp. (°C)**	**Cultivation Period (day)**	**Initial Inoculum Level (mg/L)**	**pH**	**Light Intensity (µmol m^{-2} $s^{-1)}$**	**Photo Period (h)**	**Reactor Volume (L)**
Response Variables								
NRE	0.0808	0.2162	-0.0323	0.0884	0.0764	-0.0096	-0.0306	0.0627
PRE	0.0354	0.2256	-0.1403	0.1419	0.0658	0.0584	-0.0996	0.0118
Biomass	-0.0528	0.1512	0.0333	0.1548	0.0668	0.1643	-0.0213	0.0951

InitialNConc_	**InitialPConc_**
-0.1466	-0.1194
0.0219	-0.1174
0.0456	-0.0026

3.3.2. Decision Tree Analysis

In the present work, three decision trees were constructed, each for NRE, PRE, and biomass production. The rules (combination of predictor variables) leading to high and low level of response variables were determined using these decision trees. The training dataset was used to construct the tree, and the testing dataset was applied to determine the trees' accuracy. Trees were grown by optimizing the hyperparameters using the 'OptimizeHyperparametrs' function, and the resulting hyperparameters' values are shown in Table 3.3.

Table 3.3. Accuracy of decision trees and optimized value of each hyperparameter obtained during the computational process.

Response Variable	Accuracy	Spilt Criterion	MaxNumSplit	MinLeafSize
NRE	76.08%	gdi	67	2
PRE	77.7%	gdi	67	1
Biomass	84.09%	gdi	80	5

3.3.2.1. Nitrogen Removal Efficiency

The optimal decision tree structure for NRE displaying rules leading to high and low value is shown in Figure 3.1. The tree first splits the data according to the CO_2 % (v/v) content, according to the gini index criteria.

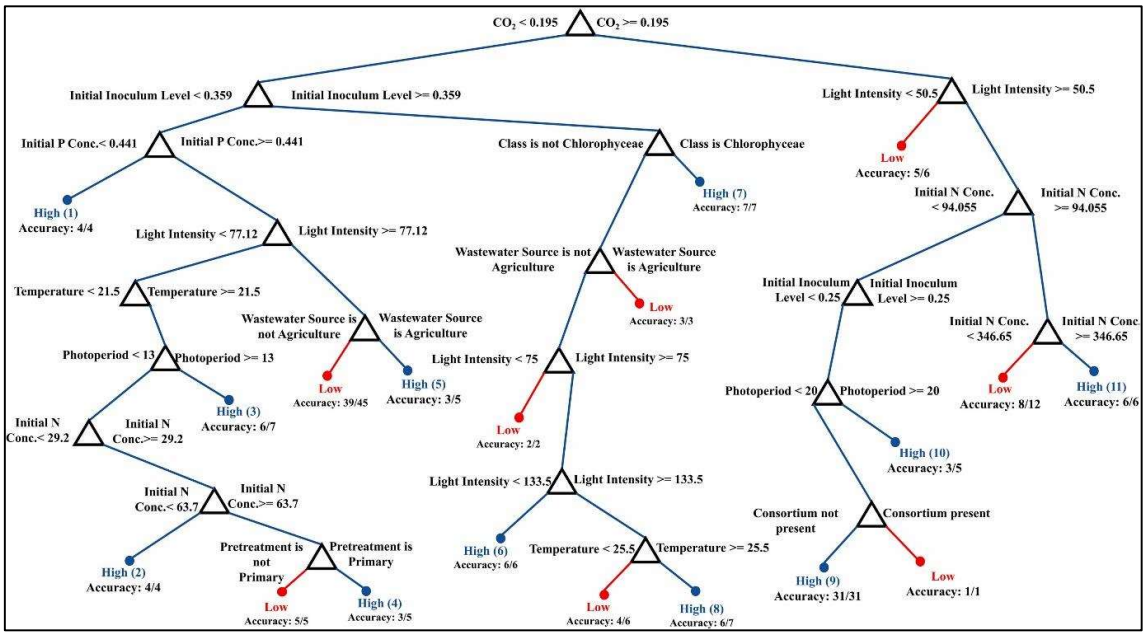

Figure 3.1. Decision tree classification for nitrogen removal efficiency (refer to Table 3.1 for units).

The confusion matrix depicting the classification accuracy of the decision tree model is shown in Figure 3.2.

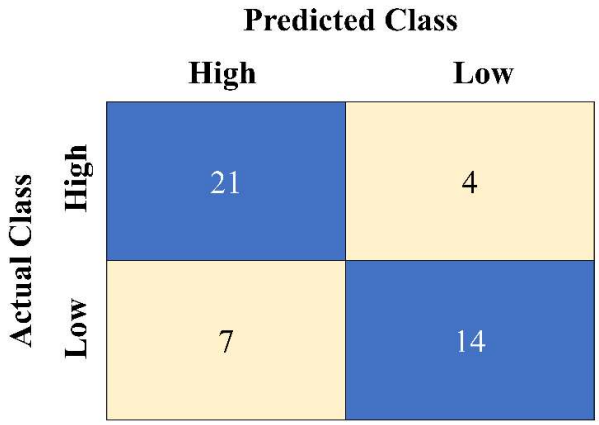

Figure 3.2. Confusion matrix for nitrogen removal efficiency.

On the left-hand side, when CO_2 content is less than 0.195 % (v/v), the initial inoculum level (IIL) provides the second branching. That means, IIL have lower Gini index than CO_2 content. Other factors that influenced the branching are initial N conc., initial P conc., temperature, light intensity, photoperiod, wastewater source, and pretreatment method. There are 8 different paths

or set of rules that leads to high NRE. Out of all paths, 2 paths provided the most generalized rules leading to high NRE. In the first path, conditions are CO_2 content < 0.195% (v/v), initial inoculum level < 0.359 g/L, initial P conc. ≥ 0.441 mg/L, light intensity < 77.12 μmol m^{-2} s^{-1}, temperature ≥ 21.5 °C, photoperiod ≥ 13 h. For the second path, conditions are CO_2 content < 0.195% (v/v), initial inoculum level ≥ 0.359 g/L, microalgae class not *Chlorophyceae*, wastewater source is other than agriculture, light intensity ≥ 133.5 μmol m^{-2} s^{-1}, and temperature ≥ 25.5 °C. Both paths provided the same accuracy of 85%, supported by 12 data points. On the right-hand side, the branch starts from CO_2 content > 0.195 %(v/v) and light intensity provides the second branching. Interestingly, continuous variables highly influenced the right-hand side's branching, and 3 paths led to high NRE.

The most generalized path leading to high NRE was provided by the right side, where the CO_2 ≥ 0.195 %(v/v), light intensity ≥ 50.5 μmol m^{-2} s^{-1}, initial N conc. < 95.05 mg/L, initial inoculum level < 0.25 g/L, photoperiod < 20 h was observed, and there was no presence of consortium. This path was supported by 31 data points with nearly 100% accuracy. Statistically ($p < 0.05$), this is the most reliable rule extracted from this tree. In order to determine the overall accuracy of the constructed tree models, rules extracted from them were simulated on the testing data, and the accuracy of the classifier models is described using a confusion matrix as shown in Figure 3.2. The model correctly classifies 21 high-class data points out of 25, and 14 low-class data points out of 21 with an overall accuracy of 76.08%.

Next, the importance of each predictor variable influencing NRE was determined by the 'predictor importance' function, and the obtained results are shown in Figure 3.3.

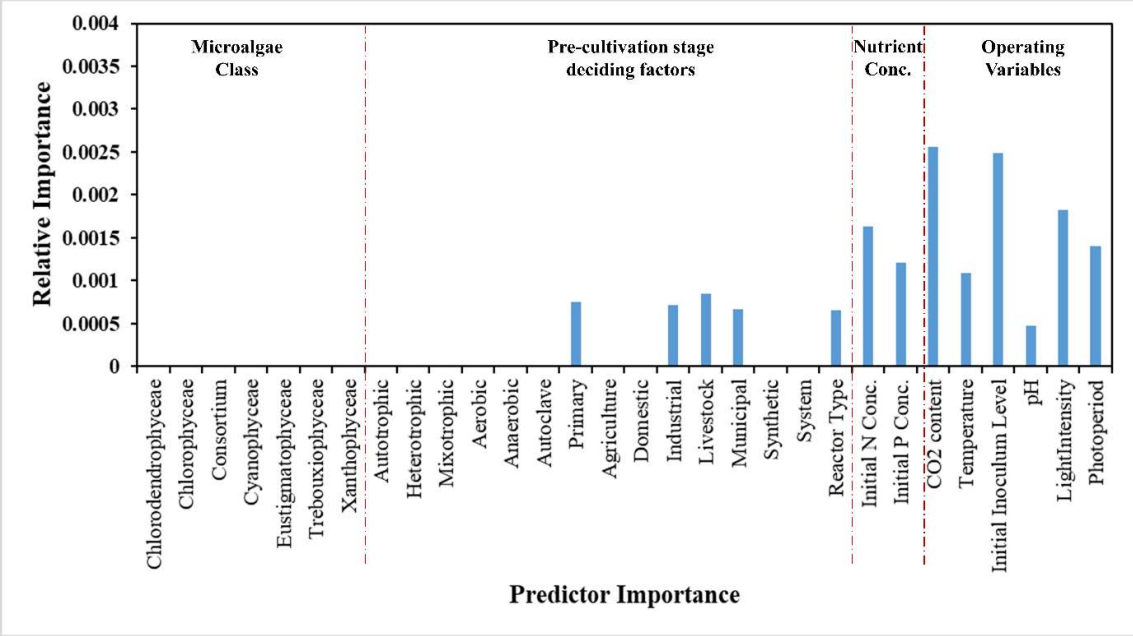

Figure 3.3. Relative importance of the predictor variables on nitrogen removal efficiency.

The operating variables such as light intensity, photoperiod, pH, initial inoculum level, CO_2 content highly influence NRE. Among nutrient concentration, initial N conc. is more influential variable than initial P conc. Among the categorical variables, wastewater sources including industrial, livestock, and municipal were more significant than domestic and synthetic sources, as they have more nutrient concentration and even more toxic compounds, which might affect the growth and nutrient removal. There are various reports in the literature that carried out the study on the effects of operating variables and nutrient concentration and tried to optimize them for achieving high NRE [29], [30]. Tran et al. (2021) cultivated *Chlorella variabilis* TH03 in both unsterile and sterile domestic wastewater at an optimized light intensity of 115 μmol m^{-2} s^{-1}, photoperiod of 16 h, initial inoculum level = 0.15 g/L, with the external supply of CO_2. The initial N. conc. in the unsterile and sterile domestic wastewater was 71.9 mg/L and 69.36 mg/L, respectively. These conditions were in accordance with the most generalized path extracted from the tree, and 96.3% and 93.3 NRE was achieved in unsterile and sterile wastewater, respectively [423]

3.3.2.2. Phosphorus Removal Efficiency

The optimal decision tree structure for PRE showing various rules leading to high and low value is shown in Figure 3.4. In this case, the tree first splits the data according to the initial N conc.

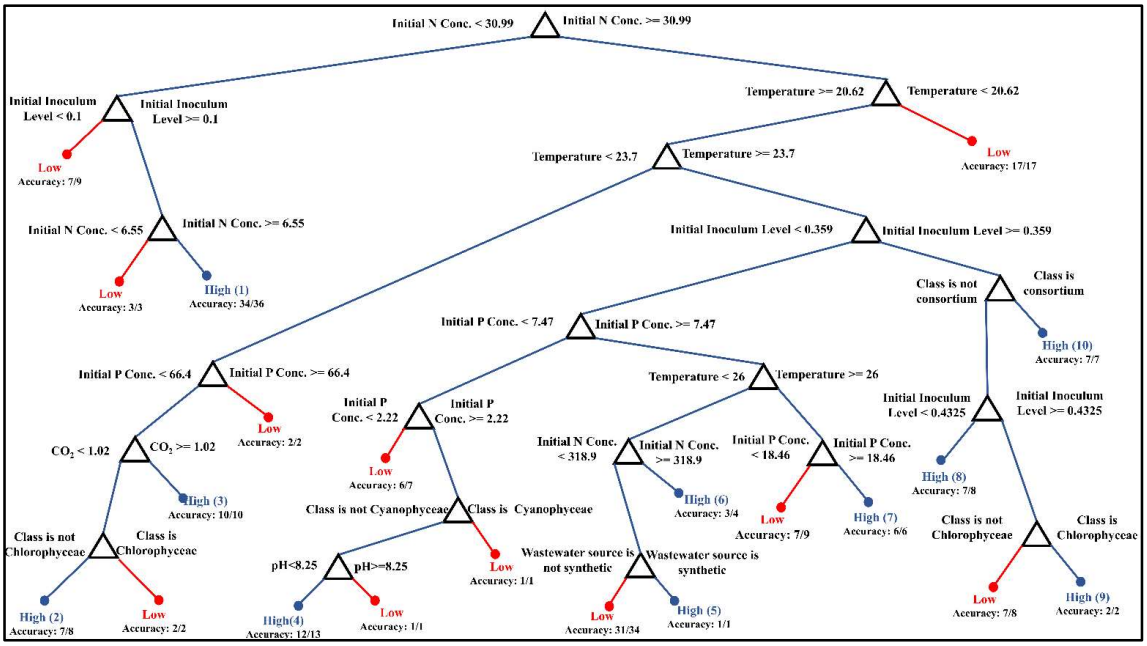

Figure 3.4. Decision tree classification for phosphorus removal efficiency (refer to Table 3.1 for units).

The confusion matrix depicting the classification accuracy of the decision tree models is shown in Figure 3.5.

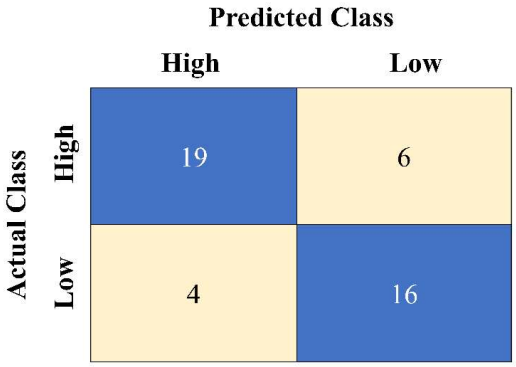

Figure 3.5. Confusion matrix for phosphorus removal efficiency.

Talking about the left-hand side, a very less level of branching occurred. The tree starts from the initial N conc. < 30.99 mg/L and continues to the initial inoculum level, which is the second important predictor. It has only one path leading to high PRE where initial inoculum level ≥ 0.1 g/L and initial N conc. ≥ 6.55 mg/L. On the right-hand side, where the tree starts from the initial N conc. ≥ 30.99 mg/L, the temperature is the second important factor, and 9 paths led to high PRE. Other factors, which influenced the branching are initial inoculum level, CO_2 content, initial P conc., pH, wastewater source, and microalgae class. The right-hand side provides the two most generalized paths leading to high PRE (path 3 and 4). The set of rules derived from path 3 leading to high PRE were initial N conc. ≥ 30.99 g/L, temperature between 20.62-23.7 °C, initial P conc. < 66.4 mg/L, CO_2 content ≥ 1.02% (v/v), supported by 10 data points with nearly 100% accuracy. For path 4, rules were initial N conc. ≥ 30.99 mg/L, temperature > 23.7 °C, initial inoculum level < 0.359 g/L, initial P conc. between 2.22-7.47 mg/L, class is other than *Cyanophyceae* and pH < 8.25, supported by 12 data points with an accuracy of 92%. Statistically, paths 3 and 4 are the most generalized set of rules obtained from the tree ($p>0.05$). The overall accuracy of the constructed tree models is described using a confusion matrix, as shown in Figure 3.6. The model correctly classifies 19 high-class data points out of 25 and 16 low-class data points out of 20 with an overall accuracy of 77.7%.

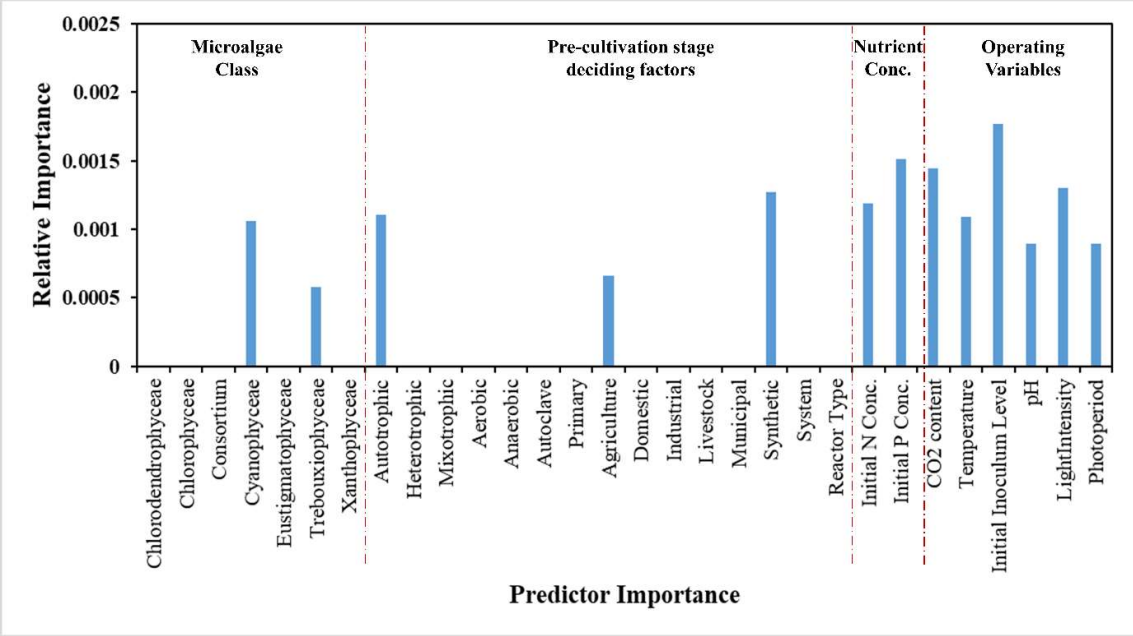

Figure 3.6. Relative importance of the predictor variables on phosphorus removal efficiency.

It is evident from Figure 3.6 that operating parameters including CO_2, temperature, initial inoculum level, pH, light intensity & photoperiod are more influential. In the nutrients, initial P conc. is more significant than initial N conc. Additionally, decision tree and predictor importance indicated that the PRE of microalgae class *Cyanophyceae* is less as compared to other classes. Song et al. (2020) cultivated four different strains of microalgae namely *Chlorella sp.* UTEX1602, *Chlorella sp.* L166, *Spirulina sp.* FACHB-439, and *Scenedesmus sp.* 336 in brewery wastewater. Their result indicated that *Spirulina sp.* belonging to class *Cyanophyceae* exhibited less PRE (92.36%) in comparison to other Chlorella *sp.* UTEX1602 (97.37%), *Chlorella sp.* (95.97%) and *Scenedesmus sp.* 336 (97.8%) [282].

3.3.2.3. Biomass Production

Figure 3.7 shows the optimal decision tree structure for biomass production. Initially, the tree splits the data according to the initial inoculum level.

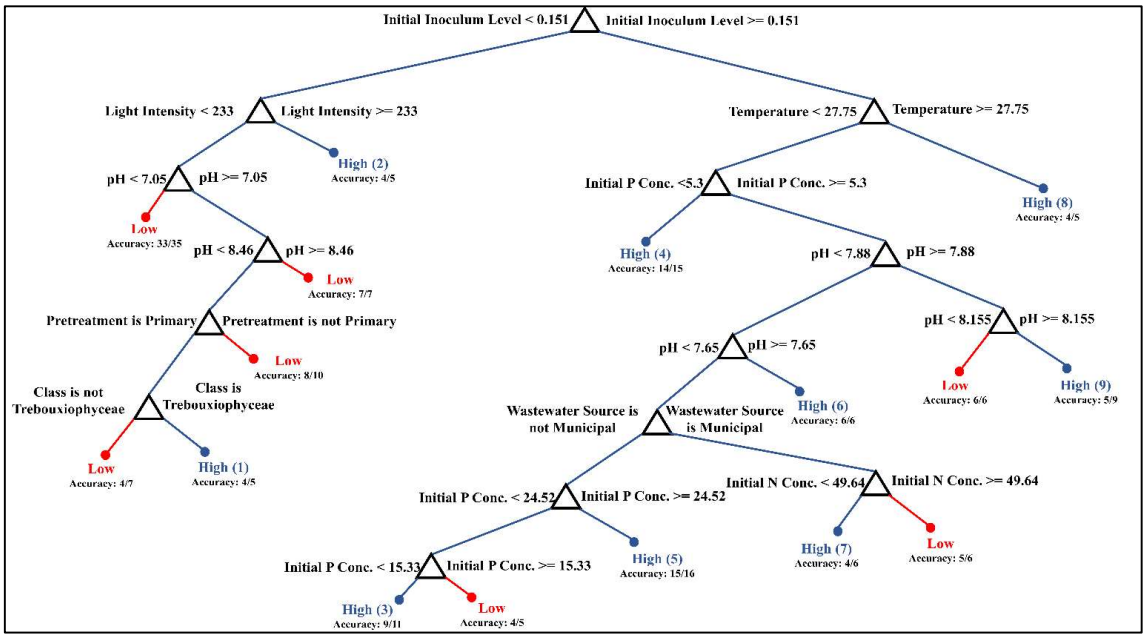

Figure 3.7. Decision tree classification for biomass production (refer to Table 3.1 for units).

The confusion matrix depicting the classification accuracy of the decision tree models is shown in Figure 3.8.

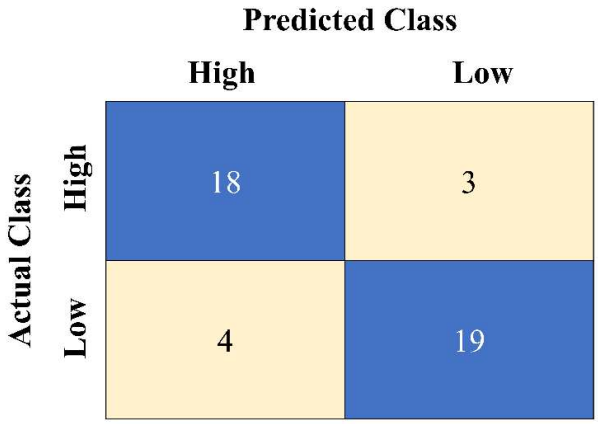

Figure 3.8. Confusion matrix for biomass production.

On the left-hand side, the tree starts from the initial inoculum level < 0.151, and the second important factor is light intensity. Other factors that influence the branching are pH, pretreatment, and microalgae class. There is only one path leading to high biomass production on the left-hand, whose rules are initial inoculum level < 0.151, light intensity < 233, pH

between 7.05-8.46, primary pretreatment method, and microalgae class *Trebouxiophyceae* used for cultivation. On the right-hand side, where the tree starts from the initial inoculum level ≥ 0.151 g/L, the temperature is the second important variable. Other variables influencing tree branching include initial nutrient concentration, pH, and wastewater source. On the right side, 8 different paths led to high biomass production; out of them, path 2,3,4 have a high level of accuracy. The most reliable rule from the tree is provided by path 4 in which initial inoculum level ≥ 0.151 g/L, temperature < 27.75 °C, initial P conc. ≥ 24.52 mg/L, pH between 7.65-7.88, and wastewater source is other than municipal. This path shows 93% accuracy level and is supported by 15 data. The overall accuracy of the constructed tree models is described using a confusion matrix, as shown in Figure 3.9. The model correctly classifies 18 high-class data points out of 21 and 19 low-class data points out of 23 with an overall accuracy of 84.09%.

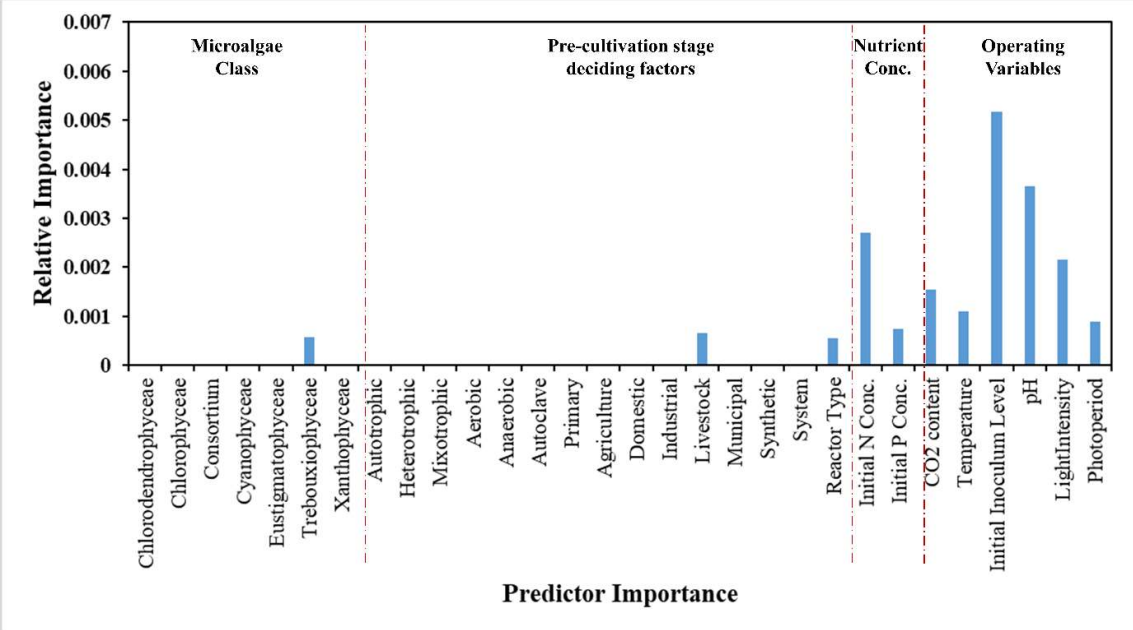

Figure 3.9. Relative importance of the predictor variables on biomass production.

The analysis of predictor importance (Figure 3.9) indicated that initial inoculum level highly influences the biomass production with other operating parameters such as pH, light intensity and photoperiod, CO_2 content, and temperature. Among the nutrient concentration, biomass production is more influenced by initial N conc. than initial P. conc, as nitrogen affects the

growth rate and varies the pH of the medium, whereas less than 1% (by weight) of P is present in the microalgae biomass [22]. Cosgun et al. (2021), through their machine learning analysis, also concluded that operating variables such as CO_2 content, photoperiod, light intensity and feed gas flow are more deterministic for biomass productivity, followed by nutrient concentration (N&P) and some heavy metals (Cu&Si) [425].

3.4. Conclusions

In the current study, the decision tree algorithm detected different patterns of predictor variables leading to high wastewater treatment capability and biomass productivity, among which most generalized paths or combinations are as follow:

- NRE: CO_2 >= 0.195%, Light Intensity >= 50.5 $\mu mol\ m^{-2}\ s^{-1}$, Initial N. conc. < 95.05 mg/L, initial inoculum level < 0.25 g/L, Photoperiod < 20 h.
- PRE: Initial N conc. >= 30.99 mg/L, Temperature between 20.62-23.7 °C, initial inoculum level < 0.359 g/L, initial P conc. between 2.22-7.47 mg/L, microalgae class is other than *Cyanophyceae* and pH < 8.25.
- Biomass: initial inoculum level >= 0.151 g/L, temperature < 27.75 °C, initial P. conc. >= 24.52 mg/L, pH between 7.65-7.88, and wastewater source is other than municipal.

It was also concluded that operating variables and nutrient concentration in wastewater highly influenced microalgae-based wastewater treatment capability.