Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 5, May 2024

Machine Learning Methods: Application for Amido Black Dye Adsorption Prediction on to Green Algae Powder-Activated Carbon

Sai Venkata Surya Punugoti, Dhruv Kumar
Department of Artificial Intelligence and Machine Learning,
Guru Gobind Singh Indraprastha university,
Delhi 110032

Meena Vangalapati
Professor, Department of Chemical Engineering,
AUCE, Andhra university, AP, India.

***Abstract*—The adsorption of amido black dye using green algae powder- activated carbon adsorbent were modeled using Extra Tree Regressor of adsorption experimental data. In addition, the correlation between variables and their importance was applied. After comprehensive feature selection analysis, three important variables were selected from six variables. The RF with the highest accuracy (R2 = 0.92) was selected as the best model for prediction of adsorption capacity of amido black dye using green algae powder- activated carbon using the five selected variables. The results suggested that amido black dye using green algae powder-activated carbon characteristics (time, dosage of amido black dye using green algae powder- activated carbon pH, and particle size) accounted for 50.7% contribution for adsorption efficiency. The accurate ability of the developed models' prediction could significantly reduce experimental screening efforts, such as predicting the dye removal efficiency of green algae powder activated carbon. The relative importance of variables could provide a right direction for better treatments of dyes in the real wastewater.**

***Keywords*—machine learning; wastewater treatment; dye adsorption; Green algae powder; activated carbon**

I. INTRODUCTION

Water, being an essential and dynamic asset, undergoes damage due to the release of waste materials containing biologically resilient and unclean components into the natural environment [1]. According to the ''United Nations World Water Development Report" published in March 2012, approximately 80 % of wastewater is directly discharged into the environment without undergoing any treatment, leading to the pollution of both surface and groundwater [2]. The majority of researchers in various fields such as chemistry, geology, agronomy, plant physiology, and medicine within the environmental sciences are focused on developing innovative methods to decrease the presence of persistent pollutants in wastewater [3]. It is worth noting that wastewater treatment is not only crucial for maintaining good health but also for preserving the environment [4]. Moreover, a healthy population can contribute to enhancing the socio-economic development of their country [5].

The models developed in this study are used to predict dye adsorption efficiency in wastewater based on measurable green algae powder AC characteristics such as time, dosage, pH, particle size, temperature, and initial concentration of the solution. This study with the aid of machine learning, which would be valuable for future applications with the increasing accumulation of big data in the scientific literature, while detecting the relative importance of each factor in improving adsorption efficiency; it provides a comprehensive understanding of dye removal using green algae powder and proposed guidelines for the treatment of wastewater and contaminated water containing dyes.

II. METHODOLOGY

A. Data Collection and description

The dataset used in this study consists of four input features (w, Co, pH, T) and an output which is Biosorption Percentage of Amido Black.

B. Exploratory data analysis

Data exploratory analysis (EDA) is a method of analyzing a data set to identify its key features, often using charts and other data visualization methods. The main purpose of EDA is to understand patterns in data, reveal relationships between variables, identify anomalies and flaws, and develop hypotheses for investigation.

1) Feature significance*:* Evaluating the significance or importance of each feature in predicting the goal variable. The technique used is Impurity-based Feature Significance (The higher, the more important the feature). The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. It is also known as Gini importance. [6]

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 5, May 2024

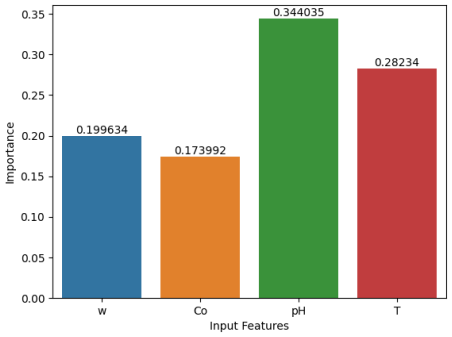

Fig.1. Feature Importance Graph

2) Correlation evaluation*:* Correlation is a statistical method used to assess a possible linear association between two continuous variables. It is measured by a statistic called the correlation coefficient, which represents the strength of the putative linear association between the variables in question. The correlation coefficient is a dimensionless quantity that takes a value in the range −1 to +1. A correlation coefficient of zero indicates that no linear relationship exists between two continuous variables, and a correlation coefficient of −1 or +1 indicates a perfect linear relationship. [7]

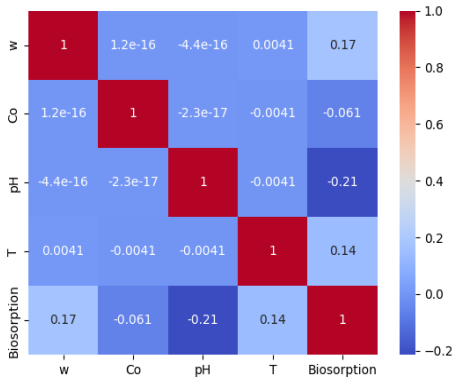

Fig.2. Heatmap

3) Univariate feature evaluation*:* It involves analyzing individual features independently to understand their distribution and characteristics. This process helps in identifying patterns, outliers, and the overall behavior of each feature in the dataset. It is crucial for gaining insights into the data before proceeding with more complex analyses. By combining univariate feature evaluation with Pair plot Visualization (It displays the relationship for each combination of variables as a matrix of plots, with diagonal plots showing univariate distributions), data analysts can gain a comprehensive understanding of the dataset's individual features and their relationships with each other.[8]

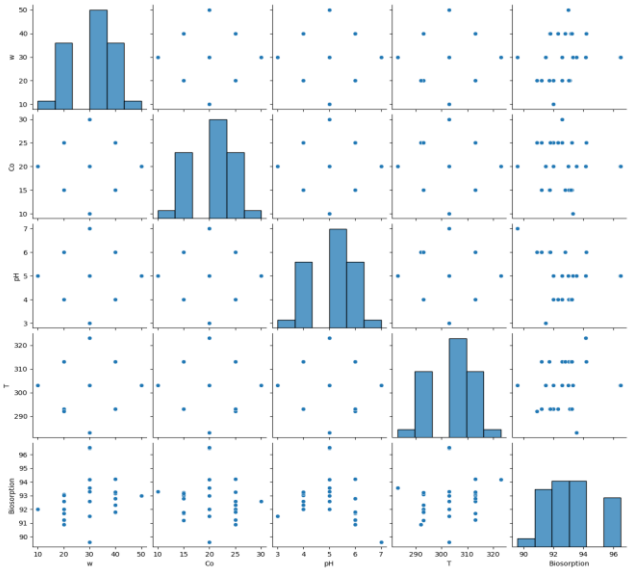

Fig.3. Pairplot

4) Feature Scaling*:* It is a technique used in data preprocessing to standardize independent features in a dataset to a fixed range. It is performed during the data preprocessing stage to handle highly varying magnitudes, values, or units. Feature scaling is essential because it ensures that all features are on a comparable scale and have comparable ranges, preventing larger scale features from dominating the learning process and producing skewed outcomes. This method scales features to have a mean of 0 and a standard deviation of 1, by subtracting the mean value of each feature and dividing the result by the standard deviation of that feature.

C. Model Estimation and Training

Model estimation, also known as version fitting or schooling, is the method of use of a system getting to know set of rules to analyse patterns from the training information and create a predictive version. The goal is to construct a Machine Learning Model that can generalize well to unseen data and make correct predictions.

5) Model evaluation & comparison*:* After training various models on the training data, Evaluating and Comparing their performance on the validation set using the chosen metrics to select the best-performing Algorithm for the desired dataset.

a) Machine Learning algorithms evaluated*:* 10 different Machine learning algorithms are evaluated and compared in this study for the prediction of the percentage of biosorption of Amido Black.

b) Comparison metrics: Evaluation metrics are used to assess the performance of machine learning models. They are crucial in determining the quality of a model's predictions and its ability to generalize to new data. [9]

- Mean Absolute Error (MAE): MAE represents the average absolute error between actual and predicted values. It sums up the absolute differences between predictions and actual and then averages them. It is

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 5, May 2024

easily understood because it's in the same scale as the target variable you're predicting for.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y_i}|$$

where,

n: *number of observation*

y_i: *the actual value of the* i^{th} *observation*

$\widehat{y_i}$: *the predicted value of the* i^{th} *observation*

- *Mean Squared Error (MSE):* MSE calculates the average squared error between actual and predicted values. It squares the differences between predictions and actuals, then computes the average. It is advantageous because it penalizes larger errors more severely than smaller ones due to the squaring operation. However, since MSE involves squaring the errors, its value is not in the same unit as the target variable, making interpretation less intuitive compared to metrics like MAE.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2$$

Where:
$\widehat{y_i}$ = Predicted value for the i^{th} data point
y_i= Actual value for the i^{th} data point
n = number of observations

Table I. Model Comparison

Algorithm	Evaluation Metrics		
	MSE	***MAE***	***Time***
ExtraTreeRegressor	0.362627	0.385000	0.000967
RandomForestRegressor	1 .004008	0.788662	0.324965
GradientBoostingRegressor	1.213315	0.862767	0.152565
AdaBoostRegressor	1.236813	1.023889	0.082982
BaggingRegressor	1.374911	0.913291	0.054762
SVR	1.741491	1.244259	0.009927
DecisionTreeRegressor	2.925786	1.057917	0.009137
KNeighborsRegressor	3.023693	1.451	0.008724
LinearRegression	4.871803	1.825286	0.006687
SGDRegressor	5.244485	1.881152	0.007599

6) *Best Model:* Considering the data recorded in the table above, Extra Tree Regressor performs the best comparably to the other Algorithms. Extra Tree Regressor also known as Extremely Randomized Trees Regressor, is an ensemble learning method for regression tasks that belongs to the family of decision tree algorithms. It is an extension of the Random Forest algorithm and shares many similarities with it. Extra Trees builds more than one decision tree during training. They differ in the way each tree is constructed. Extra Trees randomly selects split points without looking for the best one, instead of choosing the best split point based on an optimal criterion. This causes a higher degree of randomness in the decision-making process. Each decision tree in the ensemble predicts a target variable during regression voting. The final prediction can be obtained by averaging the predictions of the trees in the ensemble. [10].

7) *Hyperparameter Tuning:* The number of trees in the ensemble, the maximum depth of the trees, and the number of elements considered at each split are some of the hyperparameters of Extra Trees Regression. These hyperparameters can be adjusted using techniques such as grid search or random search.

c) *Grid search CV:* It is a hyperparameter tuning technique in scikit-learn that systematically explores all possible combinations of hyperparameters from a specified parameter grid. It fits and evaluates the model for each combination using cross-validation and selects the combination that yields the best cross-validation score.[11]

d) *Best parameters:* The Parameters which are best suited for the generalization of the data are:

```
{'ccp_alpha': 0.0,
 'criterion': 'squared_error',
 'max_depth': None,
 'max_features': 1.0,
 'max_leaf_nodes': None,
 'min_impurity_decrease': 0.0,
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'min_weight_fraction_leaf': 0.0,
 'monotonic_cst': None,
 'random_state': None,
 'splitter': 'random'}
```

III. RESULTS AND DISCUSSION

The Extra Trees Regression model was employed to predict the percentage of biosorption of Amido Black in the research study. The model exhibited promising performance in capturing the complex relationship between the input features and the biosorption percentage.

The model achieved a high coefficient of determination (R-squared) value of 0.92, indicating that 92% of the variance in the biosorption percentage could be explained by the model. This suggests that the Extra Trees Regression model was able to effectively capture the underlying patterns in the data and make accurate predictions.

Furthermore, the feature importance analysis revealed that the pH of the solution, and temperature were the most influential factors in determining the biosorption percentage. This information is valuable for understanding the key drivers of biosorption and can guide future experimental design and optimization strategies.

The model's performance was validated using cross-validation techniques, ensuring its robustness and generalizability to new data. The mean squared error (MSE) of

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 5, May 2024

the model was found to be 0.385, indicating the average squared difference between the predicted and actual biosorption percentages. This low error value further supports the model's accuracy in predicting biosorption outcomes

IV. CONCLUSION

The research study used the Extra Trees Regression model to predict Amido Black biosorption percentage. The model showed promising performance, explaining 92% of the variance in biosorption percentage. The model's feature importance analysis revealed pH and temperature as key drivers of biosorption. Cross-validation techniques confirmed the model's robustness and generalizability, with a mean squared error of 0.385, confirming its accuracy in predicting biosorption outcomes as shown in the ETR Visualisation.

ETR VISUALISATION

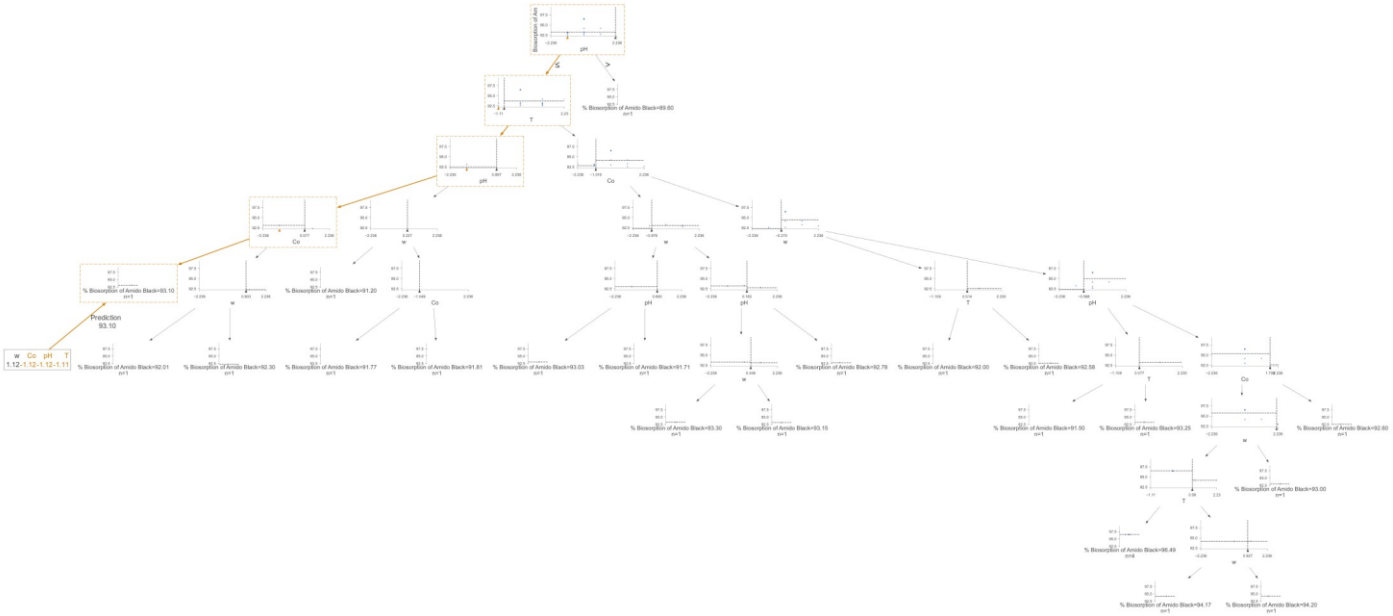

Fig.4. ETR Prediction Flow Diagram

REFERENCES

[1] G. Eason, B. Noble, and I.N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955. (*references*)

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.

[3] I.S. Jacobs and C.P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.