Extra Tree Regression model for Atrazine Herbicide removal using Machine Learning strategies

Sai Venkata Surya punugoti1, Dhruv Kumar2, Kevin Mathew3 and Meena Vangalapati4*

12 Department of Artificial Intelligence and Machine Learning, Guru Gobind Singh Indraprastha university, Delhi 110032

3 Department of Computer science and systems engineering, AUCE, Andhra university, AP, India.

4*Professor, Department of Chemical Engineering, AUCE, Andhra university, AP, India.

Abstract:

In this study, Atrazine Herbicide was removed by nanofiltration Process using Graphene oxide nanoparticles. Experimental results were obtained by Optimation at different parameters like Time, Bed height, Flowrate,pH and concentration were modeled using Extra Tree Regressor of pesticides removal by Nano filtration 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 = 0.97) was selected as the best model for prediction of our input features (Time, bed height, flow rate) and an output which is Removal percentage of atrazine **using the five selected variables.**

The accounted for 97% contribution for removal efficiency by nano filtration. The accurate ability of the developed models' prediction could significantly reduce experimental screening efforts, such as predicting Atrazine Herbicide was removed by nanofiltration Process using Graphene oxide nanoparticles . The relative importance of variables could provide a right direction for better treatments of pesticides in the real wastewater.

Keywords machine learning; wastewater treatment; nano filtration ,; Extra Tree Regressor , pesticides removal.

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 [5-6].

The models developed in this study are used to predict dye adsorption efficiency in wastewater measured based on Atrazine Herbicide was removed by nanofiltration Process using Graphene oxide nanoparticles[7-8].

Experimental results were obtained by Optimation at different parameters like Time, Bed height, Flowrate,pH and concentration were modeled using Extra Tree Regressor of pesticides removal by Nano filtration[9-10].

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 pesticides

removal using Nano filtration[11-12].and proposed guidelines for the treatment of wastewater and contaminated water containing pesticides.

METHODOLOGY:

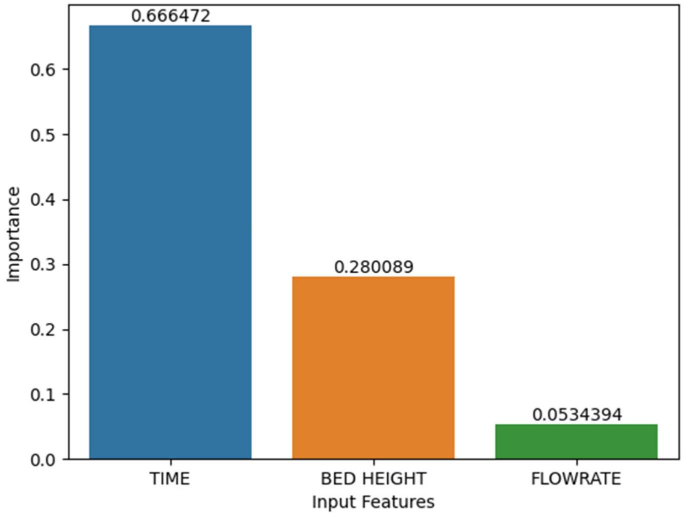
1. **Data Collection and description:**

   The dataset used in this study consists of four input features (Time, bed height, flow rate) and an output which is Removal percentage of atrazine.

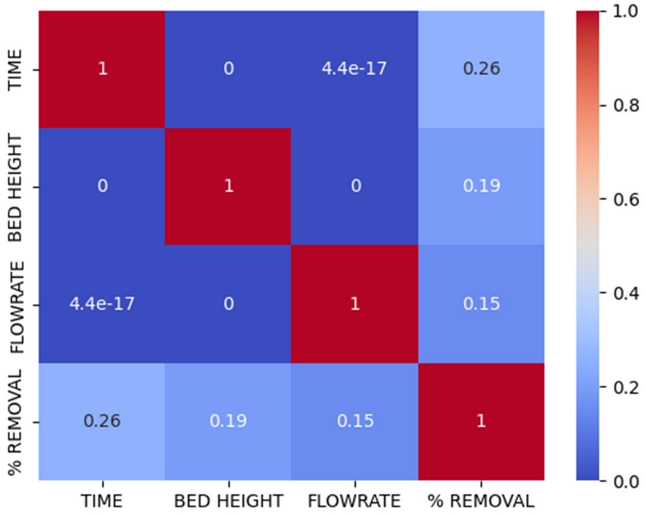
2. **Exploratory data analysis:**

   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.

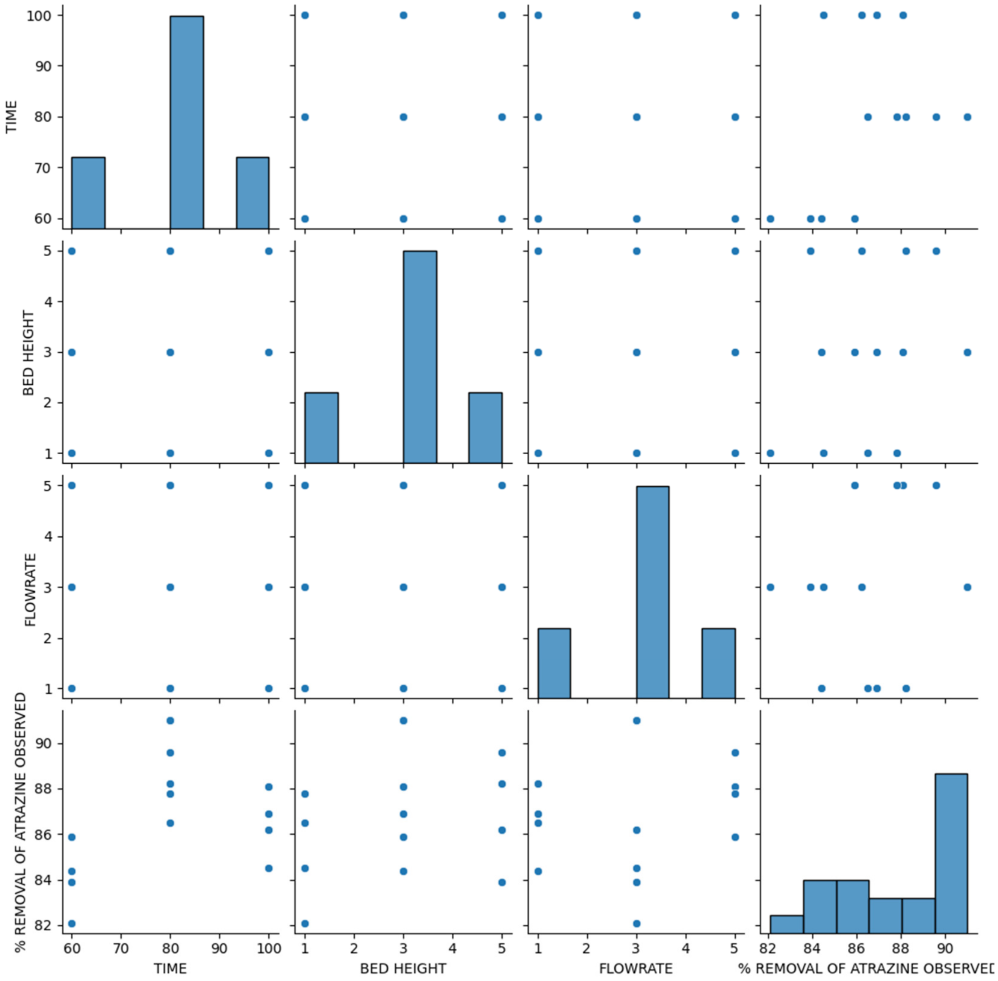
   2.1. **Feature significance**: Evaluating the significance or importance of each feature in predicting the goal variable. The technique used is Impurity-based Feature Significance (The higher, the more important the feature). The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. It is also known as **Gini importance**. [1] As We can see in the Bar plot below, the TIME feature holds the most importance in predicting the output.



   2.2. **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. [2] Plotting the heatmap of the correlation helped us to visualize the correlation between the variables as shown in the figure below. It tells us that time is highly correlated to the output as compared to the other input variables.



2.3. **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.[3]



2.4. **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.

3. Model estimation & 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.

3.1. 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.

3.1.1. Machine Learning algorithms evaluated:

10 different Machine learning algorithms are evaluated and compared in this study for the prediction of the Removal percentage of atrazine.

3.1.2. 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. [4]

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

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

Where:

$\hat{y}_i$ = Predicted value for the i^th^ data point

y_i= Actual value for the i^th^ data point

n = number of observations

3.1.2.2. **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.

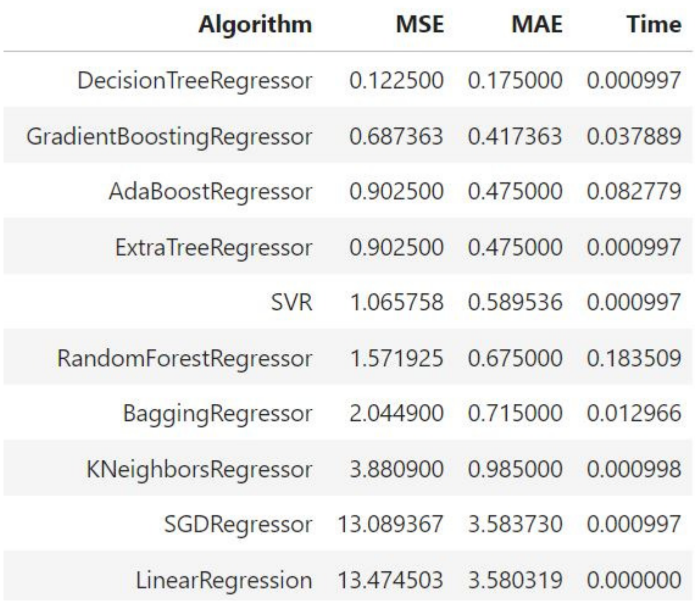
$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = mean squared error

n = number of data points

Y_i = observed values

$\hat{Y}_i$ = predicted values



Algorithm	MSE	MAE	Time
DecisionTreeRegressor	0.122500	0.175000	0.000997
GradientBoostingRegressor	0.687363	0.417363	0.037889
AdaBoostRegressor	0.902500	0.475000	0.082779
ExtraTreeRegressor	0.902500	0.475000	0.000997
SVR	1.065758	0.589536	0.000997
RandomForestRegressor	1.571925	0.675000	0.183509
BaggingRegressor	2.044900	0.715000	0.012966
KNeighborsRegressor	3.880900	0.985000	0.000998
SGDRegressor	13.089367	3.583730	0.000997
LinearRegression	13.474503	3.580319	0.000000

3.2. Best model:

Considering the data recorded in the table above, **Decision Tree Regressor** performs the best comparably to the other Algorithms.

Decision Tree Regressor: A Decision Tree Regressor is a machine learning algorithm used for regression tasks. It segments the feature space into smaller regions and fits a simple model within each segment. The algorithm splits the dataset into two subsets, minimizing the target variable's variance. It then applies recursive partitioning to each subset, creating a hierarchical tree-like structure. Each leaf in the tree holds a predicted value, and the algorithm predicts new data points based on the input's feature values. Decision trees are praised for their simplicity and ability to handle non-linear relationships. [6]

3.2.1. Hyperparameter tuning:

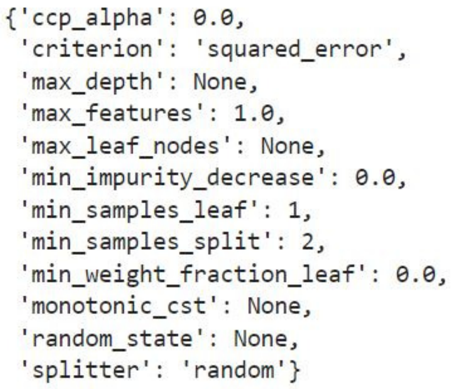
The number of trees in the ensemble, the maximum depth of the trees, and the number of elements considered at each split are some of the hyperparameters of Extra Trees Regression. These hyperparameters can be adjusted using techniques such as grid search or random search.

3.2.1.1. **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.[5]

3.2.1.2. 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'}
```

4. Results and discussion:

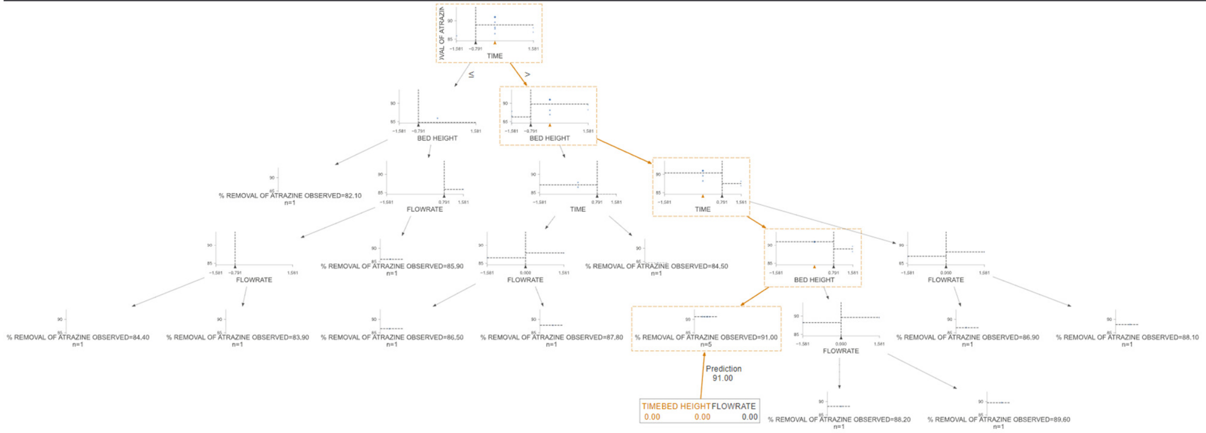
The **Decision Trees Regression** model was employed to predict the Removal percentage of atrazine in the research study. The model exhibited promising performance in capturing the complex relationship between the input features and the optimization percentage.

The model achieved a high coefficient of determination (**R-squared**) value of **0.97**, indicating that **97%** of the **variance** in the biosorption percentage could be explained by the model. This suggests that the Decision Tree Regressor model was able to effectively capture the underlying patterns in the data and make accurate predictions.

Furthermore, the feature importance analysis revealed that the time is the most influential factor in determining the removal percentage of atrazine. This information is valuable for understanding the key drivers of optimization 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 the model was found to be **0.175**, indicating the average squared difference between the predicted and actual biosorption percentages. This low error value further supports the model's accuracy in predicting optimization outcomes.

DTR VISUALISATION


OVAL OF ATRAZIN
TIME
≤
>
BED HEIGHT
BED HEIGHT
% REMOVAL OF ATRAZINE OBSERVED=82.10
n=1
FLOWRATE
TIME
TIME
FLOWRATE
% REMOVAL OF ATRAZINE OBSERVED=85.90
n=1
FLOWRATE
% REMOVAL OF ATRAZINE OBSERVED=84.50
n=1
BED HEIGHT
FLOWRATE
% REMOVAL OF ATRAZINE OBSERVED=84.40
n=1
% REMOVAL OF ATRAZINE OBSERVED=83.90
n=1
% REMOVAL OF ATRAZINE OBSERVED=86.50
n=1
% REMOVAL OF ATRAZINE OBSERVED=87.80
n=1
% REMOVAL OF ATRAZINE OBSERVED=91.00
n=5
FLOWRATE
% REMOVAL OF ATRAZINE OBSERVED=86.90
n=1
% REMOVAL OF ATRAZINE OBSERVED=88.10
n=1
Prediction
91.00
TIME BED HEIGHT FLOWRATE
0.00 0.00 0.00
% REMOVAL OF ATRAZINE OBSERVED=88.20
n=1
% REMOVAL OF ATRAZINE OBSERVED=89.60
n=1
90
85
−1.581
−0.791
0.000
0.791
1.581

5. Conclusion:

The research study used the Decision Tree Regression model to predict Removal percentage of atrazine. The model showed promising performance, explaining 97% of the variance in optimization percentage. The model's feature importance analysis revealed time as key driver of optimization. Cross-validation techniques confirmed the model's robustness and generalizability, with a mean squared error of 0.175, confirming its accuracy in predicting optimization outcomes.

References:

1. Menze, B.H., Kelm, B.M., Masuch, R. *et al.* A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics* **10**, 213 (2009). https://doi.org/10.1186/1471-2105-10-213
2. Othman W, Hamoud B, Kashevnik A, Shilov N, Ali A. A Machine Learning-Based Correlation Analysis between Driver Behaviour and Vital Signs: Approach and Case Study. Sensors. 2023; 23(17):7387. https://doi.org/10.3390/s23177387

3. Ahsan, Md Manjurul & MAHMUD, M. A. & Saha, Pritom & Gupta, Kishor Datta & Siddique, Zahed. (2021). Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance. Technologies. 9. 52. 10.3390/technologies9030052.
4. G. Mishra, D. Sehgal, D. Sehgal, and J. K. Valadi, "Quantitative structure activity relationship study of the anti-hepatitis peptides employing random forest and extra tree regressors," *Bioinformation*, vol. 13, no. 3, pp. 60–62, 2017. View at: Publisher Site | Google Scholar
5. Ranjan, G S K & Verma, Amar & Sudha, Radhika. (2019). K-Nearest Neighbours and Grid Search CV Based Real Time Fault Monitoring System for Industries. 1-5. 10.1109/I2CT45611.2019.9033691.
6. Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry. 2015 Apr 25;27(2):130-5. doi: 10.11919/j.issn.1002-0829.215044. PMID: 26120265; PMCID: PMC4466856.
7. Bandela Sowjanya, Pulipati King, Meena Vangalapati, "Skin of allium sativum (garlic) mediated green synthesis of ZnO nanoparticles and it's adsorption performance for congo red dye removal: kinetic, isotherm and thermodynamic studies", European Chemical Bulletin, 12(2023) 326-332.
8. U Sirisha, Bandela Sowjanya, H Rehana Anjum, Thanusha Punugoti, Ahmed Mohamed, Meena Vangalapati, "Synthesized TiO2 nanoparticles for the application of photocatalytic degradation of synthetic toxic dye acridine orange", Materials Today: Proceedings, 62(2023)3444-3449.
9. Basheera Hussain Khatoon, Avapati Surya Lokesh, Lavu Ramadevi, A Ajay Raj, Pulipati King, Meena Vangalapati, "Sodium Lauryl Sulphate removal using copper electrodes without and with perforations by electro coagulation process", Materials Today: Proceedings,59(2022)655-660.
10. M Subhashita, T Punugoti, B Sowjanya, VR Poiba, "M Vangalapati,Synthesis of Cu/ZnO nanoparticles and its exploitation as a catalyst for the removal of Cetrimonium Bromide", Advances in Materials and Processing Technologies ,8(2022)1880-1888.
11. D Parveen, VR Poiba, M Vangalapati, "Characterisation and performance of ZnO nano particles for the removal of sodium dodycl sulphate", Advances in Materials and Processing Technologies, 8(2022)913-921.
12. SI Vali, U Sirisha, VR Poiba, M Vangalapati, P King, "Synthesis and characterization of Titanium doped activated carbon nanoparticles and its application for the removal of dicofol", Materials Today: Proceedings, 44(2021)2290-2295.