PREDICTING THE RATE OF ADSORPTION USING MACHINE LEARNING ALGORITHMS

1.ABSTRACT:

The research focused on utilizing zinc oxide for the adsorption of hydrogen sulfide from aqueous solutions. The study gathered a wide range of data points, including adsorption capacities under varying conditions of temperature, solution pH, and time (measured in minutes). The observations were meticulously recorded during practical experiments, which involved different volumes (5, 10, 15, and 20 ml).The primary objective was to predict the adsorption capacity for a larger volume, specifically 25 ml, without the need for additional physical experiments. The decision tree regression and web-based model has been developed Subsequently, the expected values were compared to the actual experimental results, and deviations were quantified using machine learning algorithms. This approach allowed researchers to efficiently forecast future values without the requirement for additional practical experiments. This method not only saved valuable time and resources but also utilized a web application to construct the predictive model. Through the web application interface, users could input data, and the corresponding outputs would be generated and displayed, streamlining the entire process. .the development of this model has been compared with other regressors and after comparing with random forest and linear regression this comparison result in selecting the decision tree regressor for developing the model. The selection of this decision tree regressor gives better results where accuracy is taken for classification.

Key words: machine learning, regressor, random forest, decision tree, linear regression,web application, html, css.

2.INTRODUCTION:

Supervised machine learning involves training algorithms with labeled data sets to predict output values. The model is guided and supervised during training using this labeled data. Supervised machine learning can be categorized into two main types: classification and regression. Regression includes linear and multiple regressions. Among the various classification techniques such as decision tree, SVM, logistic regression, and random forest, the problem at hand can be effectively solved using the decision tree ID tree algorithm.

The paper is organized into five sections, each following a logical flow of work. The first section presents a basic introduction to the problem addressed in the research article, followed by a literature survey, which enhances our understanding of the chosen topic.

The third section, "Proposed Model and Discussions," plays a crucial role as it covers the key factors influencing the entire article. This section details the methodology, including the implementation of the algorithm and the development of the model, presented through a flow diagram. Additionally, the data preprocessing steps are clearly explained.

Section four is the most intriguing part of the paper as it reveals the results of the model, providing a deeper understanding of its performance. Finally, sections five and six serve as the conclusion and discuss potential future research avenues for the project.The proposed model and discussion are as fallows

3. PROPOSED MODEL AND DISCUSSION:

Initially while developing a model the work has categories into four main segmentations fallows raw data set segmentation, preprocessing segmentation , model development and evolution in details all this stages are explained below.

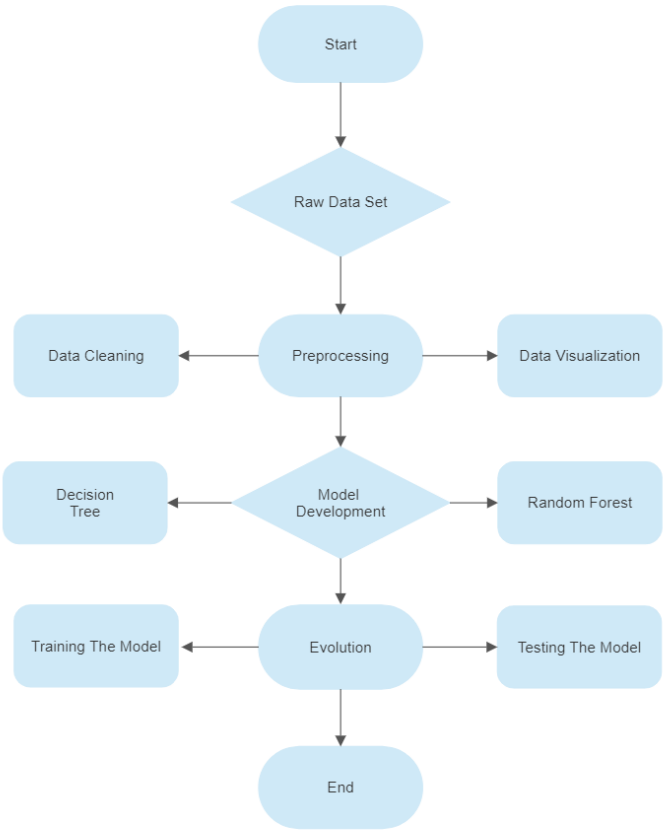


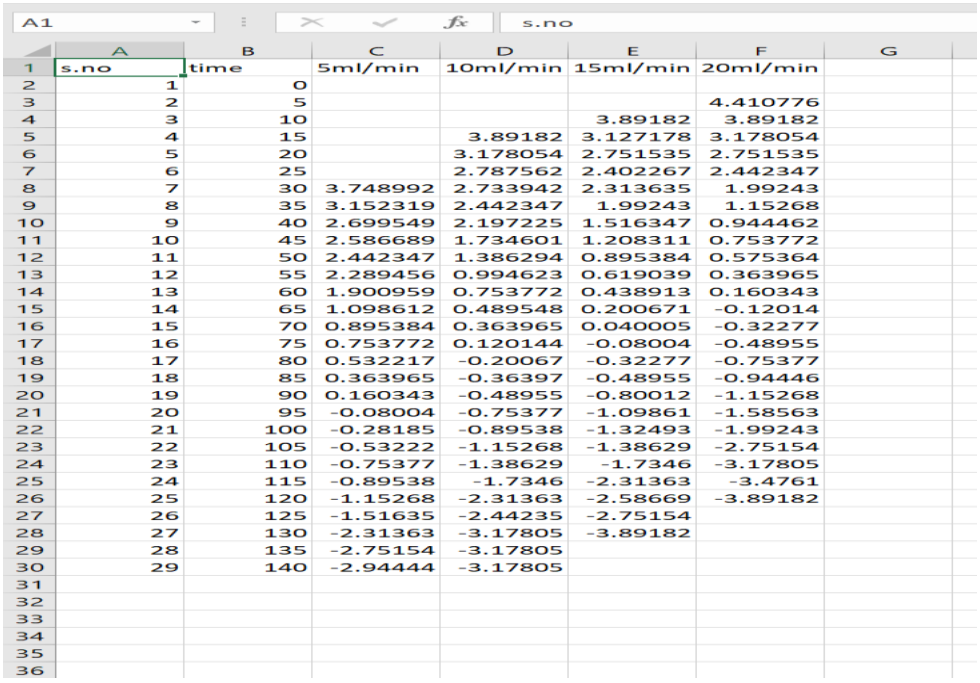
Fig (1): Flow Chart Represents The Entire Flow of Work

This work begins with selecting a raw data set in CSV format. Initial model building and preprocessing techniques are applied, including data cleaning and visualization. During data cleaning, null values are addressed using methods like backfill and forward fill, and Python commands are utilized to handle these null values. Subsequently, three data sets are created, each with the null values filled. For model building, three algorithms—Decision Tree Regressor, Random Forest, and Linear Regression—are employed. These algorithms aid in predicting the

next rate of adsorption based on the provided data. The process involves the creation of a training and testing data set for model evaluation and to ensure its ability to generalize to new data. Finally, after implementing all the techniques, model validation techniques such as accuracy and error measurement are used to assess the trained model's performance and reliability. Through this workflow, researchers can make informed predictions about future adsorption rates based on the experimental data, facilitating valuable assumptions about the upcoming behavior of the adsorption process. The discussion on data set is as fallows

3.1 Dataset description:

The data has been collected while conducting this experiment the data has been catogerised into 5ml,10ml,15ml,20ml as solution flows the data has been recorded upto 140 mins so this application development is going to predict the values or rate of adsorption after 140mins



	A	B	C	D	E	F	G
1	s.no	time	5ml/min	10ml/min	15ml/min	20ml/min	
2	1	0					
3	2	5				4.410776	
4	3	10			3.89182	3.89182	
5	4	15		3.89182	3.127178	3.178054	
6	5	20		3.178054	2.751535	2.751535	
7	6	25		2.787562	2.402267	2.442347	
8	7	30	3.748992	2.733942	2.313635	1.99243	
9	8	35	3.152319	2.442347	1.99243	1.15268	
10	9	40	2.699549	2.197225	1.516347	0.944462	
11	10	45	2.586689	1.734601	1.208311	0.753772	
12	11	50	2.442347	1.386294	0.895384	0.575364	
13	12	55	2.289456	0.994623	0.619039	0.363965	
14	13	60	1.900959	0.753772	0.438913	0.160343	
15	14	65	1.098612	0.489548	0.200671	-0.12014	
16	15	70	0.895384	0.363965	0.040005	-0.32277	
17	16	75	0.753772	0.120144	-0.08004	-0.48955	
18	17	80	0.532217	-0.20067	-0.32277	-0.75377	
19	18	85	0.363965	-0.36397	-0.48955	-0.94446	
20	19	90	0.160343	-0.48955	-0.80012	-1.15268	
21	20	95	-0.08004	-0.75377	-1.09861	-1.58563	
22	21	100	-0.28185	-0.89538	-1.32493	-1.99243	
23	22	105	-0.53222	-1.15268	-1.38629	-2.75154	
24	23	110	-0.75377	-1.38629	-1.7346	-3.17805	
25	24	115	-0.89538	-1.7346	-2.31363	-3.4761	
26	25	120	-1.15268	-2.31363	-2.58669	-3.89182	
27	26	125	-1.51635	-2.44235	-2.75154		
28	27	130	-2.31363	-3.17805	-3.89182		
29	28	135	-2.75154	-3.17805			
30	29	140	-2.94444	-3.17805			
31							
32							
33							
34							
35							
36							

Fig(2): The above figure represents the experiment values which are used for training our model

The dataset has been created by conducting experiments to measure the adsorption rate. Within this dataset, there is an record of the adsorption rates up to 140 minutes.

3.2 Preprocessing techniques:

Data preprocessing is a critical step in data mining, involving the transformation of raw data into a more accessible and usable format. It plays a crucial role as machine learning and data mining algorithms cannot directly operate on raw data. Before applying these algorithms, it is essential to ensure the quality of the data through preprocessing.

Data Cleaning: This technique involves cleaning up the raw data set, which includes dealing with missing values and removing irrelevant rows or columns that may hinder the analysis.

By applying data cleaning as one of the preprocessing techniques to the considered data set, missing values are handled, and irrelevant rows or columns are removed, making the data ready for training and testing. This step ensures the data is in a suitable state for further analysis and modeling with machine learning or data mining algorithms.

2.3 Model development:

Machine learning is broadly classified into two types supervised machine learning and unsupervised machine learning. In supervised machine learning, labeled datasets are used to train algorithms that can accurately classify data or predict outcomes. Some common algorithms used in supervised learning include logistic regression, linear regression, naive Bayes classifier, artificial neural networks, random forest, and predictive modeling. This type of learning relies on labeled input and output data for training. On the other hand, unsupervised learning deals with unlabeled or raw data. It involves tasks such as clustering, association, and dimensionality reduction, where the model identifies patterns and relationships within the data without predefined labels.

In this proposed model, two supervised machine learning algorithms, namely multiple linear regression and random forest, are utilized. These algorithms are selected for their widely acknowledged better performance in prediction tasks. Specifically, for predicting adsorption, the supervised algorithms chosen are random forest, decision tree, and linear regression. These algorithms will use the labeled data to establish the relationship between input features and the adsorption outcomes, enabling accurate predictions without the need for manual labeling or classification.

3.3.1 Decision tree:

The algorithm described here is a non-parametric supervised learning algorithm capable of both classification and regression tasks. It is commonly categorized into three types: ID3, CART, and MARs. Decision trees, represented in a tree-like structure, are used to solve data problems.

In this specific paper, we have utilized the ID3 algorithm for training the model. ID3 was invented by Ross Quinlan and follows a top-down greedy approach to construct the decision tree. Starting from the top, the algorithm iteratively selects the best feature at each step and creates new nodes to build the tree.

To measure the entropy, the mathematical representation provided above is employed. Entropy quantifies the amount of uncertainty present in the dataset.

$$H(S) = \sum_{c \in C} -p(c) log_2 p(c)$$

The variables used in the entropy calculation are as follows:

• S: The current dataset for which entropy is being calculated.

• C: The set of classes in S.

• P(c): The proportion of the number of elements in class c to the total number of elements in the dataset.

By measuring entropy, the ID3 algorithm determines the information gain at each node, enabling effective decision-making in building the decision tree. This approach efficiently classifies and predicts outcomes for new data points based on the tree's structure.

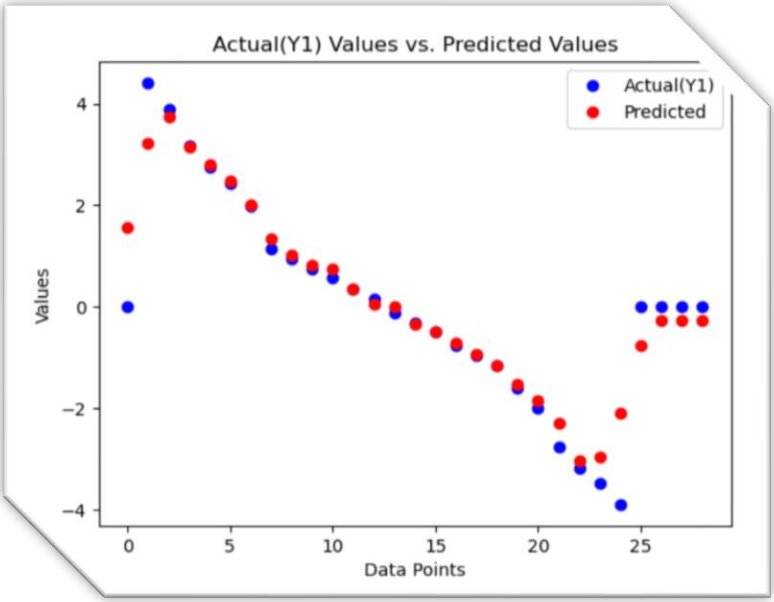
During the model development process, the decision tree algorithm played a pivotal role in enhancing the efficiency of the model. The functioning and analysis of this model improved its ability to predict the adsorption rate with greater accuracy and efficiency.

We selected two columns and labeled one as 'X' and the other as 'Y.' Our next step involved training the model using this dataset and assessing its performance. This dataset, commonly referred to as 'training data,' is also synonymous with 'predicted data' in this context. We applied the decision tree algorithm to the chosen X and Y variables, which aided in the model's training process and generated real-time predictions as output.

4. RESULTS AND DISCUSSION:

4.1 Model prediction:

The graph presented below illustrates the comparison between the actual data and the predicted data for all the regression models used. After testing and training the data set, the predicted values have been plotted in a bar graph for visual representation.



Fig(3): The above graph represents the plotting of the test and train data set.

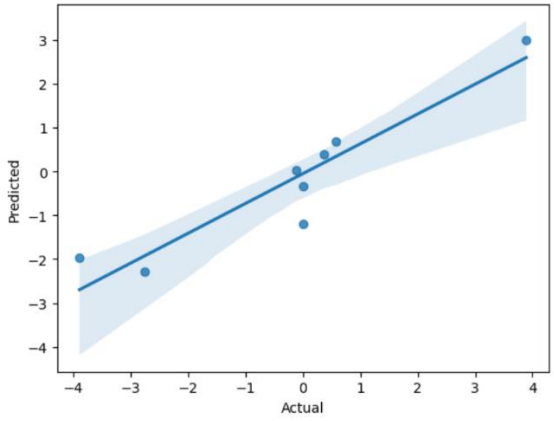
We chose two columns, designating one as 'X' and the other as 'Y.' We proceeded to train the model with this dataset and evaluated its performance. The resulting dataset, often referred to as 'train data,' is also known as 'predicted data.'

4.2 Performance measure:

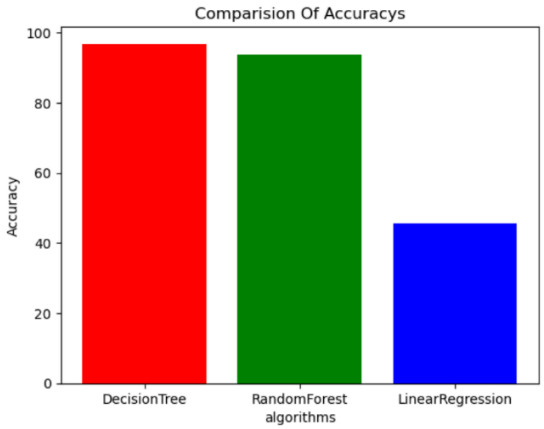
4.2.1 Accuracy:

The model's performance is evaluated using the testing dataset, where the actual output is compared with the predicted output to calculate the prediction accuracy. The outputs generated by linear regression, random forest, and decision tree are compared to determine their effectiveness.

Upon analyzing the performance and efficiency of the models, it is observed that the decision tree regressor yields the best outcomes among the three.



Fig(4): The above graph shows the plotting of actual and predicted values.



Fig(5): Graphical representation of accuracy of different algorithms.

Following the implementation of all the algorithms, the decision tree exhibited the highest accuracy compared to the other two. (Tell me the reason for why it have high accuracy) The graph displayed above provides a clear visualization of the accuracy results comparison among the different algorithms, enabling a better understanding of their performance.

4.2.2 Calculating errors:

The development of model has been successfully implemented every model is not an accurate model In order to rectify the error in this stage we are going to calculate the error. Finally, three types of errors were calculated to evaluate the model's performance: mean absolute error, mean square error, and root mean square error. These error metrics provide valuable insights into the accuracy of the model's predictions. A lower error value indicates a more accurate model, as it means the predicted values are closer to the actual values in the dataset.

4.2.2.1 mean absolute error:

The mean absolute error (MAE) is a metric used in statistics and machine learning to measure the average absolute differences between predicted values and actual values in a dataset. It provides a straightforward way to assess the accuracy of a predictive model, with lower MAE values indicating better predictive performance. MAE is particularly useful when the magnitude of errors is essential to consider in a modeling context, as it treats all errors equally and provides a clear measure of how far off the predictions are from the actual values.

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

MAE=mean absolute error

Yi = prediction

Xi = true value

n = total number of data points

4.2.2.2 mean square error:

Mean Square Error (MSE) is a widely used metric in statistics and machine learning to measure the average squared difference between the predicted values and the actual values in a dataset. It quantifies the overall accuracy of a predictive model, with lower MSE values indicating better model performance. MSE is a valuable tool for assessing and comparing the quality of different models or tuning model parameters.

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

Yi = prediction

Yi^ = actual

n = total number of data points

4.2.2.3 root mean square error:

Root Mean Square (RMS) is a mathematical measure used to calculate the square root of the average of squared values in a dataset. It is commonly employed to find the magnitude or effective value of a set of numbers, making it particularly useful in fields like statistics and signal processing. RMS helps provide a representative value that accounts for both positive and negative deviations from the mean, making it valuable for various applications, including in analyzing alternating currents and evaluating data variability.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}},$$

Yi = prediction

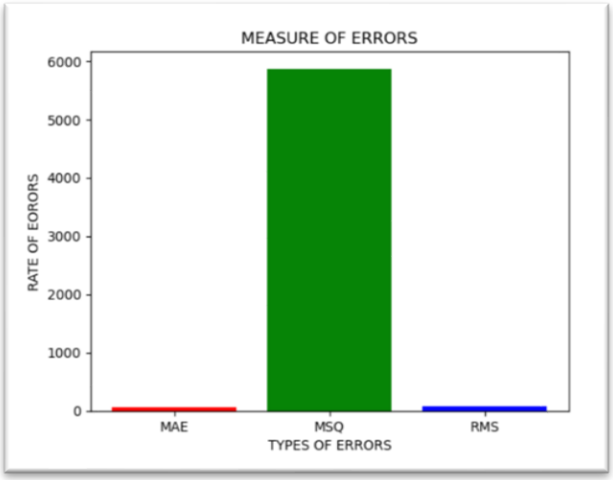
Yi^ = actual

n = total number of data points

4.2.2.4 MAE,MSE,RMSE :

The Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) have been calculated for the model, and the model's errors have been obtained.

```
Mean Absolute Error: 0.2945374815
Mean Square Error: 0.1462930743019257
Root Mean Square Error: 0.3824827764774849
```



Fig(6): Graphical representation of rate of errors

4.3 Predicted data or predicted outcome:

The central focus of the study is on forecasting upcoming outcomes or the forthcoming adsorption rates. The accompanying figure provides a visual representation that contrasts the real results with the projected ones.

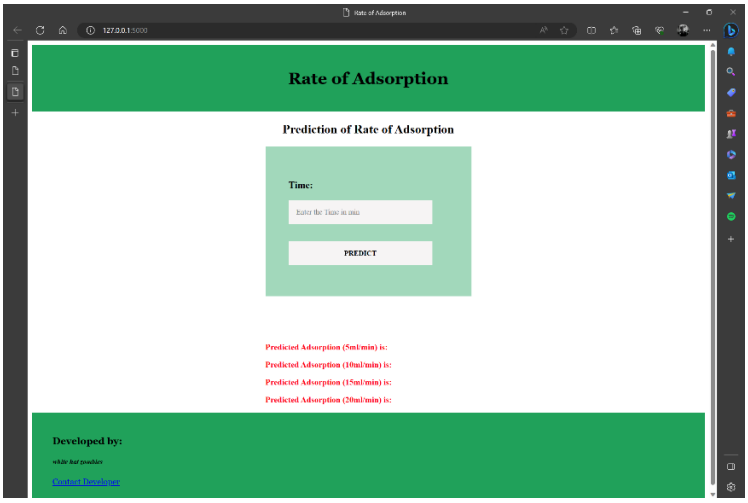
	Actual	Predicted
0	0.363965	0.389542
1	0.000000	-0.345548
2	-3.891820	-1.961499
3	-2.751535	-2.292328
4	0.000000	-1.203078
5	0.575364	0.676691
6	3.891820	3.005780
7	-0.120144	0.022389

Fig(7): The predicted values has been obtained after the implementation of algorithm.

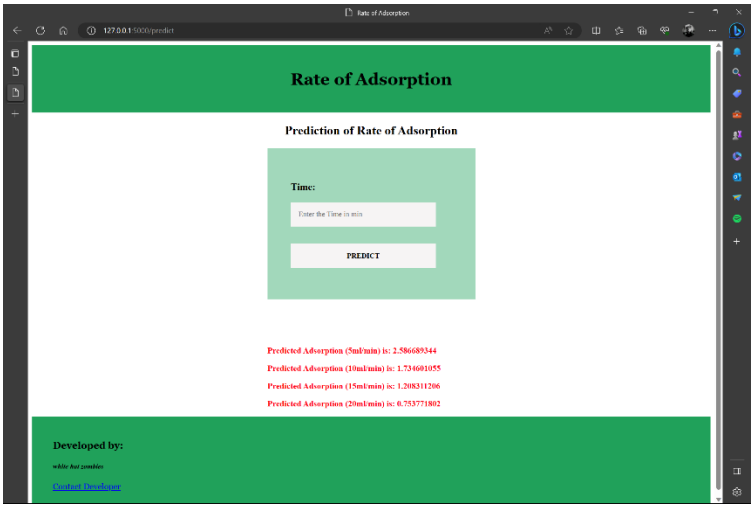
We selected two columns, labeling one as 'X' and the other as 'Y.' Subsequently, we trained the model using this data and tested its performance. The resulting trained dataset has been documented as 'train data.' The train data also known as predicted data.

4.4 web application:

As part of the deployment process, we integrated our machine learning model into a web application. The web application was developed using HTML and CSS. To connect the machine learning model with the web application or web page, we utilized Python libraries such as Pickle and Flask. Flask facilitated the seamless integration of our machine learning model into the web application, enabling users to interact with the model through the web interface.



fig(8): this picture represents our web application



fig(9): this picture shows the output of predicted value

This application has been designed to predict adsorption values using a dataset comprising only 30 values. We will predict the values by inputting the predicted output for the columns, as displayed below. The outcome has been obtained from the trained model.

5.Conclusion:

In conclusion, the implementation of machine learning techniques has proven invaluable in estimating upcoming values and rates of adsorption. During the model development phase, we employed three supervised machine learning algorithms: decision tree regressor, random forest, and linear regression. The results indicated that the decision tree regressor achieved an impressive accuracy of 95%, outperforming the other two algorithms (random forest with 91% accuracy and linear regression with 46%).By training the data and utilizing the existing values, we were able to predict outcomes with greater precision. The decision tree regressor emerged as the most accurate algorithm among the three, making it an effective tool for predicting unknown outcomes in random chemical experiments.

Overall, the application of machine learning algorithms has significantly enhanced our ability to anticipate and understand adsorption rates, offering valuable insights and potential optimizations for chemical experiments in the future.

6.future scope of a projection:

The initial development of this model has successfully provided valuable insights into predicting upcoming rates of adsorption. Our future plan involves taking this model to the next level, where it will be further improved and integrated into the development of a chatbot. This advancement will expand the model's capabilities, making it versatile and applicable beyond just adsorption prediction.

In the future, this model will not be limited to adsorption prediction alone. It can be adapted and utilized for various other concepts where the final outcomes are uncertain or unknown. The power

of machine learning algorithms, such as decision tree regressor, random forest, and linear regression, will enable easy identification and prediction of outcomes in diverse scenarios. This opens up new possibilities and opportunities for harnessing the potential of machine learning in addressing a wide range of complex problems and challenges.

7.references:

1.Weidong, L. I., et al. "Implementation of AdaBoost and genetic algorithm machine learning models in prediction of adsorption capacity of nanocomposite materials." *Journal of Molecular Liquids* 350 (2022): 118527.

2. Toyao, Takashi, et al. "Toward effective utilization of methane: machine learning prediction of adsorption energies on metal alloys." *The Journal of Physical Chemistry C* 122.15 (2018): 8315-8326.

3. Yang, Hongrui, et al. "Predicting heavy metal adsorption on soil with machine learning and mapping global distribution of soil adsorption capacities." *Environmental Science & Technology* 55.20 (2021): 14316-14328.

4. Pardakhti, Maryam, et al. "Machine learning using combined structural and chemical descriptors for prediction of methane adsorption performance of metal organic frameworks (MOFs)." *ACS combinatorial science* 19.10 (2017): 640-645.

5. El Bilali, Ali, et al. "Prediction of sodium adsorption ratio and chloride concentration in a coastal aquifer under seawater intrusion using machine learning models." *Environmental Technology & Innovation* 23 (2021): 101641.