Exploring Nitrogen Retention in Soil: Data Mining and Predictive Modeling with Biochar Data

Kao C Saelee†  
 Data Science  
 University of Colorado, Boulder  
 Bouder, Co, US  
 kasa2696@colorado.edu

ABSTRACT

This project aims to investigate the characteristics of biochar in nitrogen retention using data mining methods and machine learning. Biochar is a form of charcoal produced from the pyrolysis of organic materials, such as wood, crop residues, or manure. It has been shown to be an effective way to improve soil fertility, but more research is needed to fully understand its potential.

Methods such as data processing along with imputation techniques such k-nearest neighbors, non-negative matrix factorization and Singular Value Decomposition will be used to understand the effects of nitrogen retention from various types of biochar. To evaluate and compare the effectiveness of the imputation techniques, machine learning models such as adaboost and gradient boost for regression and classification are implemented. The results of this analysis could inform future policies and practices related to soil fertility.

By leveraging data mining methods, this project will shed light on the unique properties and characteristics of various biochar sources. It aims to identify more cost-effective biochar alternatives for improving soil fertility and potentially contribute to climate change mitigation efforts. The insights gained from this research can inform decision-making processes and promote sustainable agricultural practices.

CCS CONCEPTS

Data preprocessing, Visualization, Correlation, Clustering, Machine learning

KEYWORDS

Data mining, N20, Clustering, Machine Learning, random forest, non-negative matrix factorization, biochar, python

1. Introduction

In 2050, the world’s population is expected to reach approximately 9.8 million according to the United Nations “World Population Prospects: The 2017 Revision”.[1] To sustain the growing population, we need to be able to produce a large amount of food with the same amount of land and resources. Soil fertility is important for food production as the availability of nutrients in the soil have a direct impact on crop yield and quality.

Among the various methods available to enhance soil fertility, one sustainable approach is the utilization of biochar. Biochar is a form of charcoal that is produced from heating organic matter in the absence of oxygen, a process called pyrolysis. It is a highly porous material with a large surface area, which makes it useful for a variety of applications, including soil amendment, water filtration, and carbon sequestration.



Figure 1: An image of biochar [2]

One of the key benefits of biochar in agriculture is its ability to improve soil fertility and plant growth. Biochar can increase soil water-holding capacity, reduce nutrient leaching, and enhance soil nutrient availability. In addition, biochar can improve soil structure and microbial activity, which can further enhance soil fertility.

An important aspect of biochar's impact on nitrogen sources is its ability to enhance nitrogen retention in soils. Biochar can adsorb and stabilize nitrogen in the soil, reducing the loss of nitrogen through leaching and volatilization. This can help to reduce the environmental impact of agriculture by reducing nitrogen pollution in waterways and greenhouse gas emissions from nitrogen losses.

However, the impact of biochar on nitrogen sources can vary depending on the type of feedstock used to produce the biochar. For example, biochar produced from nitrogen-rich feedstocks such as manure may have a greater impact on nitrogen sources than biochar produced from nitrogen-poor feedstocks such as wood chips. In addition, the application rate of biochar can also affect its impact on nitrogen sources, with higher application rates potentially leading to increased nitrogen retention in the soil.

Overall, the use of biochar in agriculture has the potential to improve soil fertility and reduce the environmental impact of agriculture by enhancing nitrogen retention in soils. However, the choice of biochar material and application rate should be carefully considered to ensure that the desired benefits are achieved while minimizing any potential negative impacts.

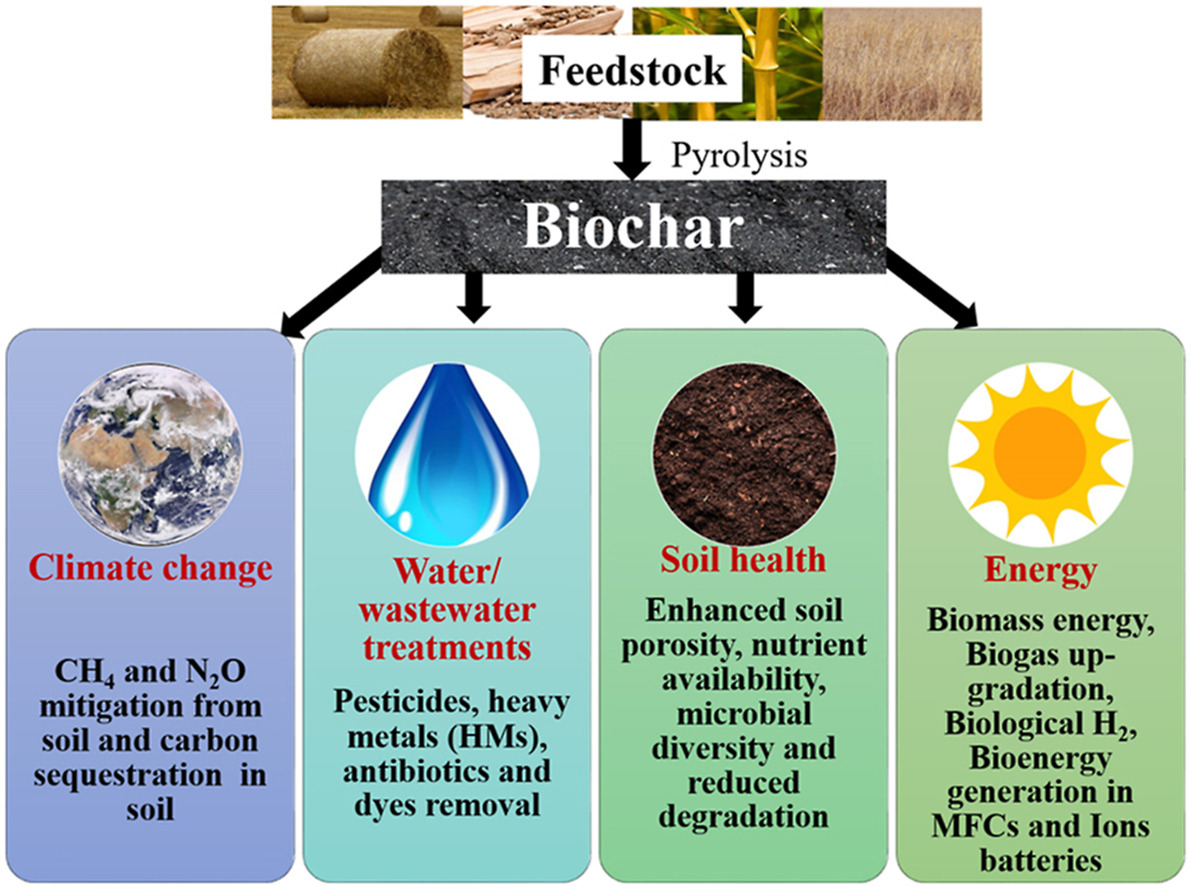


Figure 2: Popular applications of Biochar. [3]

We aim to build upon prior work by utilizing machine learning and data mining techniques to analyze three publicly available datasets to gain insight on biochar characteristics and their nitrogen retention potential. Additionally, we will use visualizations to present the data in a clear and intuitive way, allowing us to identify trends and patterns that may not be immediately apparent. By doing so, we hope to provide new insights into the effectiveness of biochar as a solution for soil fertility and contribute to the growing body of knowledge in this area.

2. Related Work

Biochar has been the subject of numerous studies in recent years, with researchers exploring its potential as a solution for carbon sequestration and climate change mitigation. A variety of techniques have been used to present the data and assess the effectiveness of biochar, including field trials, laboratory experiments, and modeling approaches.

Many studies have focused on the impact of biochar on soil health and crop productivity, with some demonstrating that biochar can improve soil fertility, nutrient retention, and water-holding capacity. Other studies have explored the potential of biochar for carbon sequestration, with some showing that it can help to reduce atmospheric CO2 levels.

Previous research has also used various visualization techniques to present the data, including graphs of basic statistics such as paired tests, charts, and maps. These visualizations have been used to show the distribution of biochar production and application around the world, the effectiveness of biochar in improving soil health, and the potential for biochar to mitigate climate change.

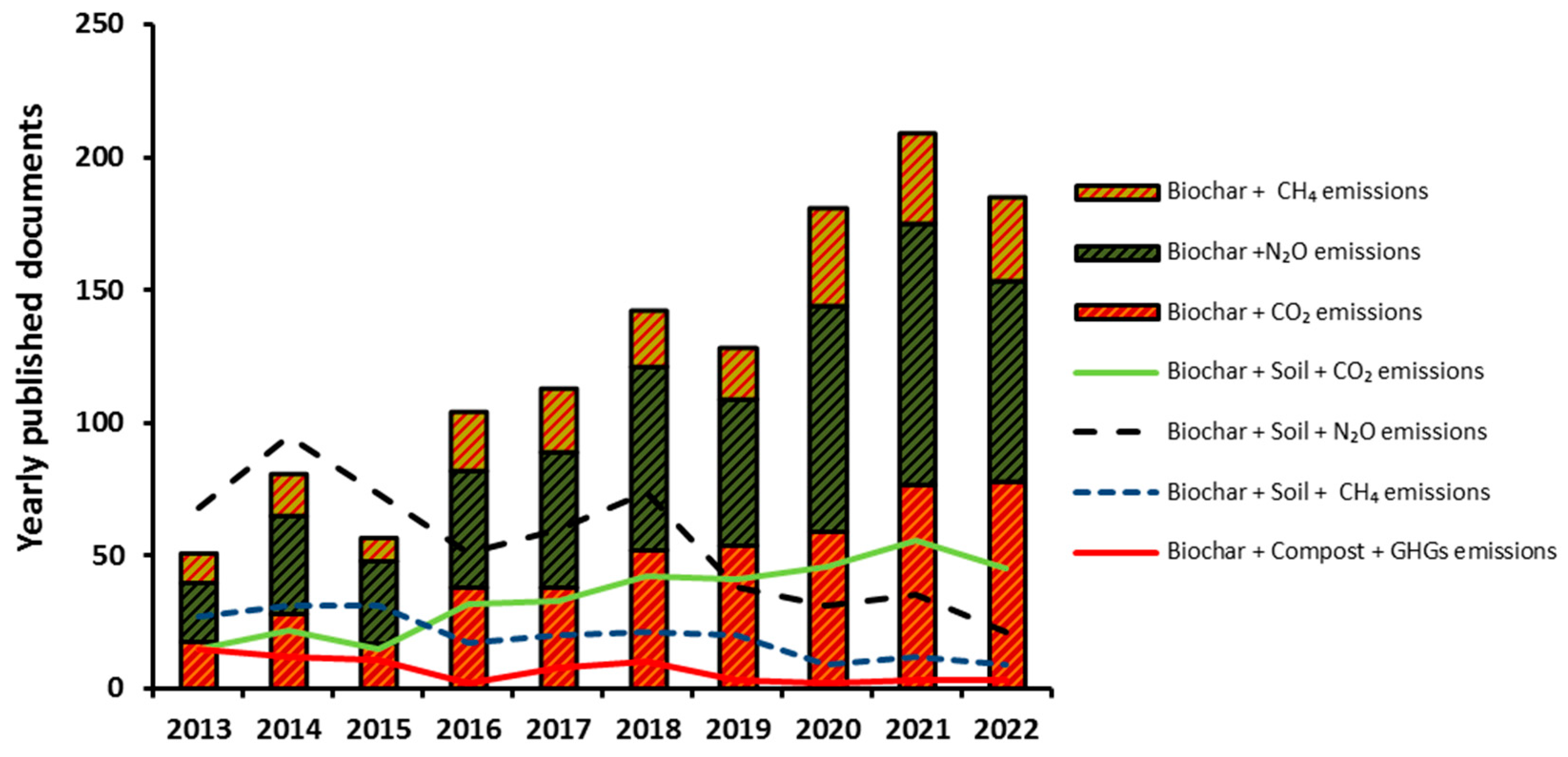


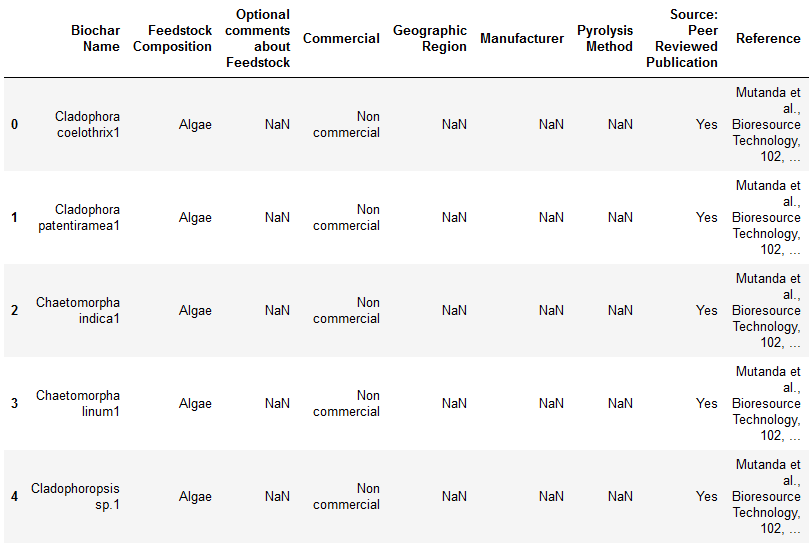
Figure 3: The number of published documents that include Biochar in the keyword by the year they were published. [4]

3. Data Collection and Tools

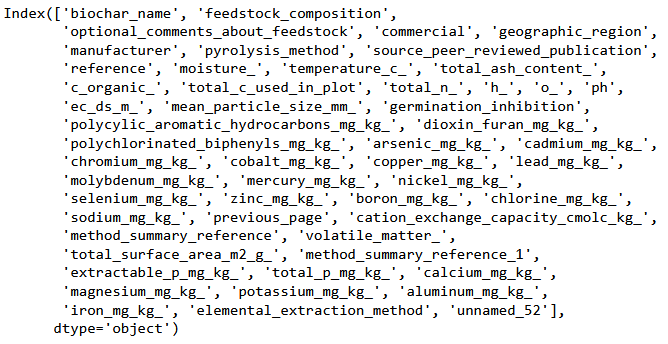
We begin the project by analyzing datasets from the public domain: one from the biochar UC Davis sorption website [2] and two from the USDA website [6]. The Sorption dataset contains data on the characteristics of various biochar types. One UDA dataset contains the characteristics of several biochar materials and the other contains experimental data with measure of dissolved nitrogen collected from the soil at two timepoints: 14 days and 28 days. We will be Python 3 on Jupyter and using modules such as pandas for data frame manipulation, seaborn and matplotlib and scikit learn for data mining and machine learning.

4. Exploratory Analysis and Visualization

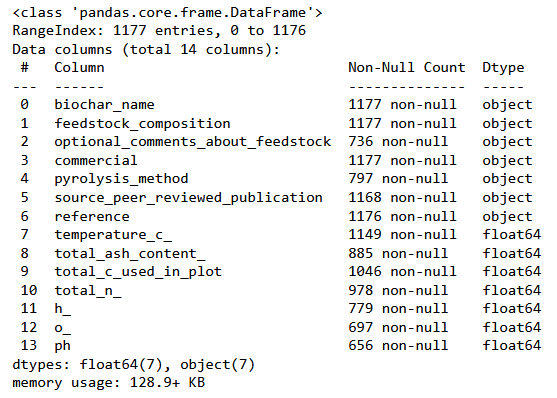
One important aspect of this project is to ensure the quality of the data used in the analysis. To accomplish this, we will implement data cleaning techniques to remove any errors, duplicates, or inconsistencies in the datasets. Here’s a snapshot of the Sorption dataset:



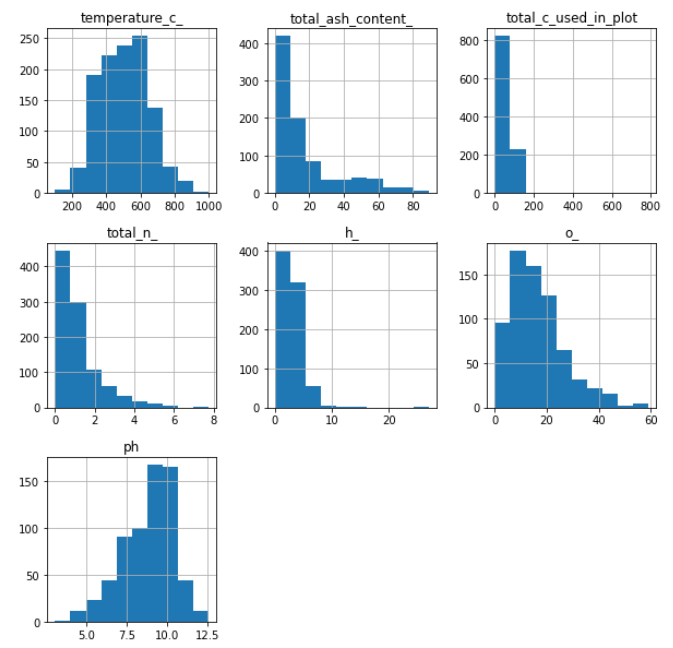
Here, we check and clean the names of the columns. The data has 1177 rows and 53 columns.



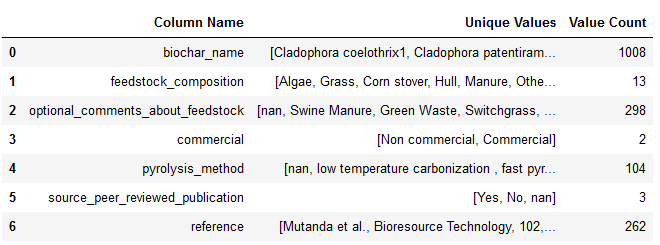
Now we can check that dataset for any null values and remove rows that contain null values that are at least half the rows of the dataset.



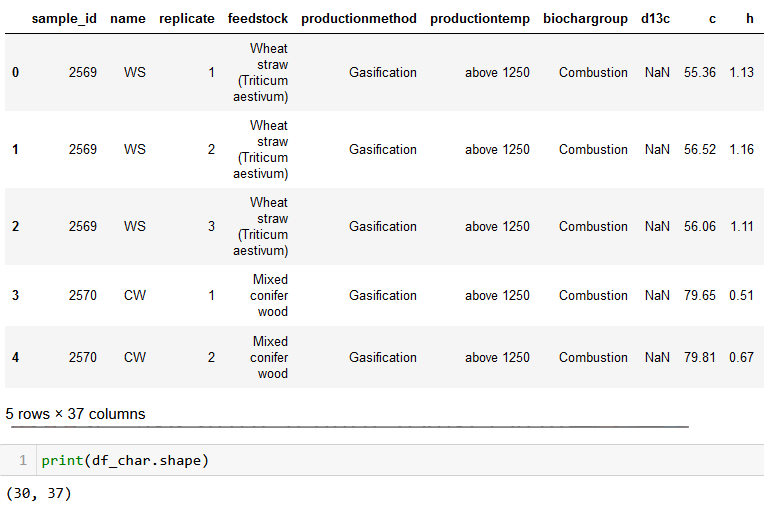
Let’s visualize the data distribution:

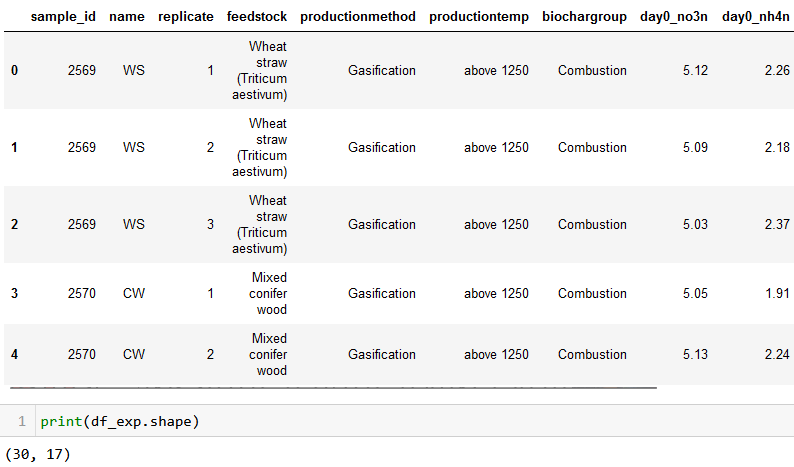


Check the unique values in each column

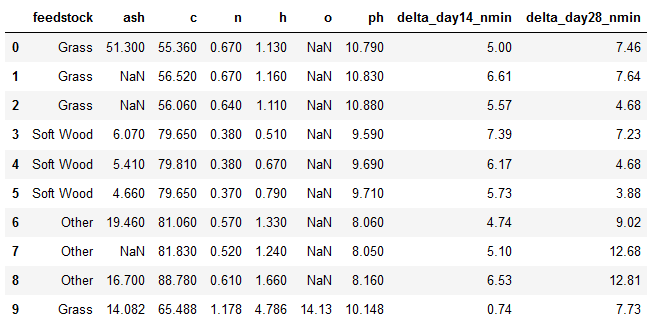


Now we can employ the same data preparation methods to the USDA biochar properties dataset and the USDA experimental timeseries datasets respectively:

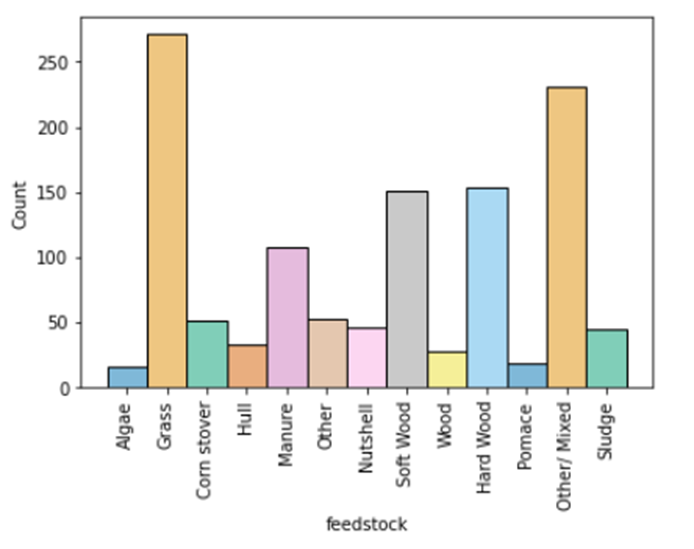




Lastly, we’ll combine the 3 datasets together and return this data frame:



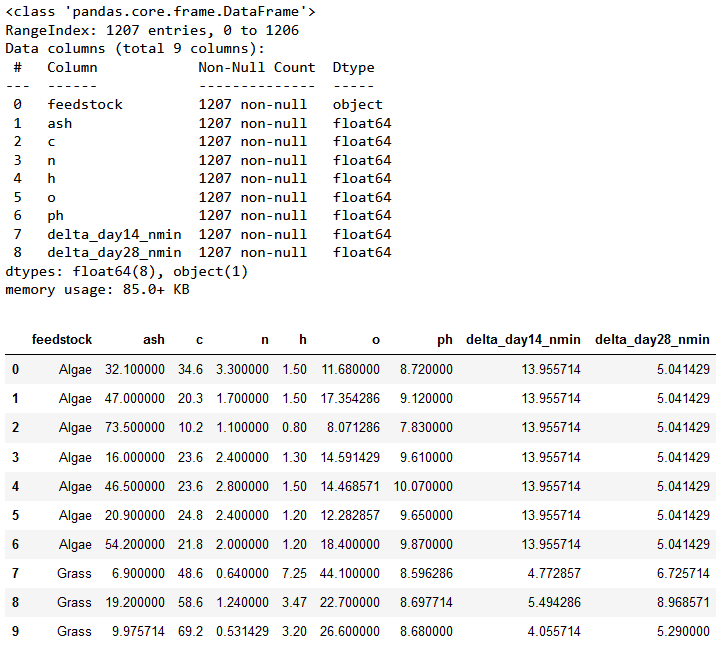
In addition, visualizations will be utilized to gain insights into the data, and to identify patterns and correlations that may not be immediately apparent. The below chart is the count of each biochar datapoint in all datasets combined.



5. Data Imputation and Preprocessing

To predict missing values in the biochar data, we will be using and comparing 3 techniques: k-nearest neighbors (KNN), non-negative matrix factorization (NMF), singular value decomposition (SVD).

K-nearest neighbors (KNN) is a non-parametric supervised machine learning algorithm that can be used to impute missing values based on the similarity of the input data. It works like this: First we need to identify the missing values, handle categorical variables or scale numeric variables. Second, we need to calculate similarity (distance) between the missing value point and the other points. Third, we select the number of neighbors used for imputation. Finally, we perform the imputation and do that for each missing variable. The choice of k or nearest neighbors is important because a smaller value may result in overfitting and sensitivity to noise while a higher value may underfit and may lose identification of certain patterns. The strength of KNN lies where it is used to predict missing values that have a random occurrence and not related to the pattern of the data. Here is the dataset where KNN was implemented:



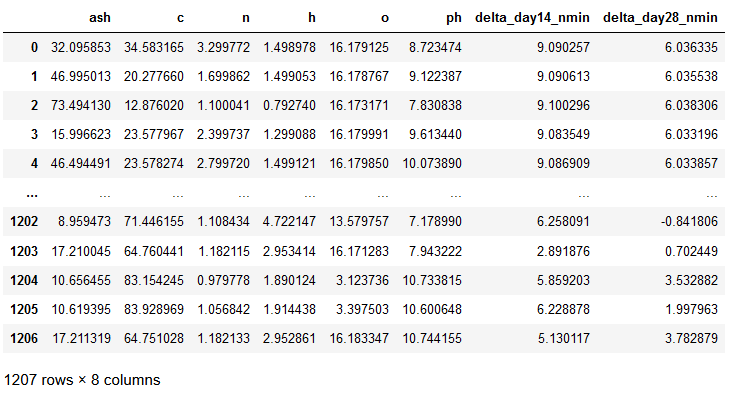
Non-negative matrix factorization (NMF) is a matrix factorization technique that aims to find two non-negative matrices, W and H, such that X ≈ WH, where W represents the basis matrix and H represents the coefficient matrix. The algorithm iteratively updates the values of W and H to minimize the difference between X and the reconstructed matrix WH. The equations are as follows::

W\_new = W \* ((X \* H.T) / (W \* H \* H.T))

H\_new = H \* ((W.T \* X) / (W.T \* W \* H))

The goal of NMF is to find the optimal values for W and H such that the product approximation WH closely matches the original matrix. This is achieved by minimizing the distance between the matrix and WH, usually measured using metrics like Euclidean distance.

Here is the output using NMF below:



Singular Value Decomposition (SVD) is a matrix factorization method that factorizes a given m x n matrix A into 3 matrices.

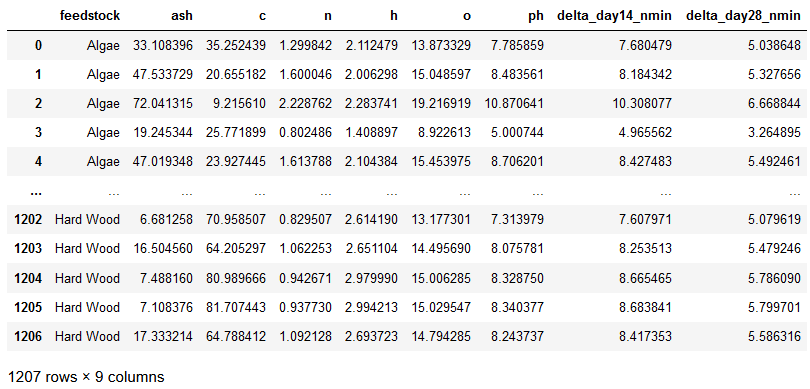
The formula is: A = U\*Σ\*V^T where

1. U is the orthogonal matrix of m x m

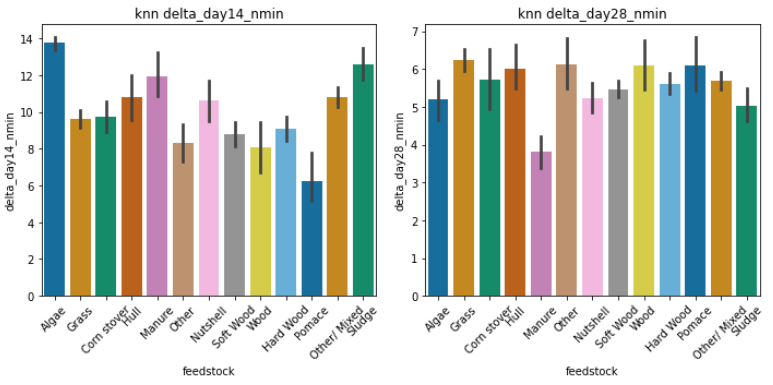
2. Σ is a diagonal matrix with non-negative singular values in descending order of m x n and

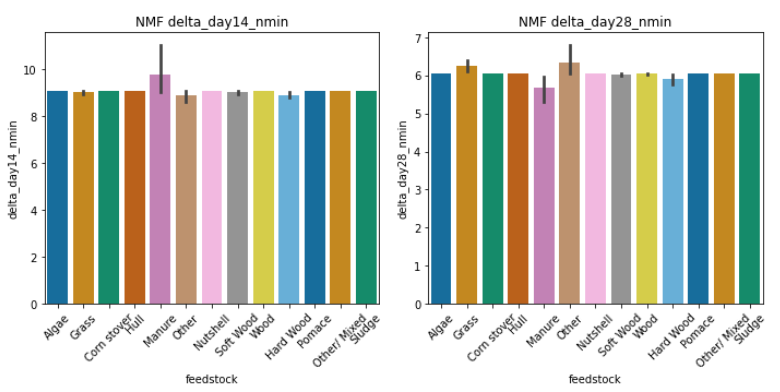
3. V^T is a transpose of an n xn orthogonal matrix

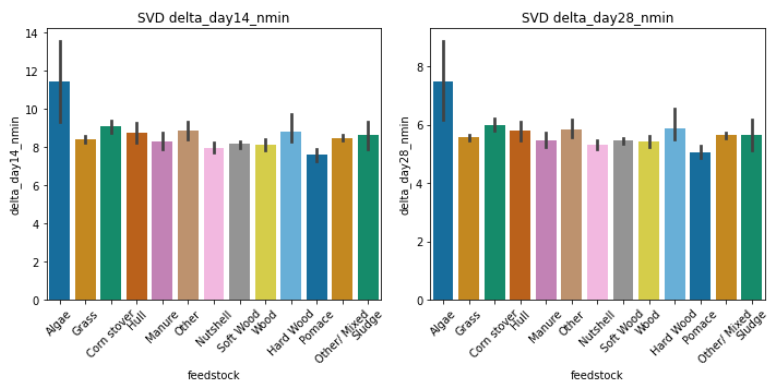
A snapshot of SVD output used to impute the missing values is shown below:



The figures below shows the barplots of each of the dissolved oxygen measurements at 14 and 28 days as well as the biochar type with that measure.



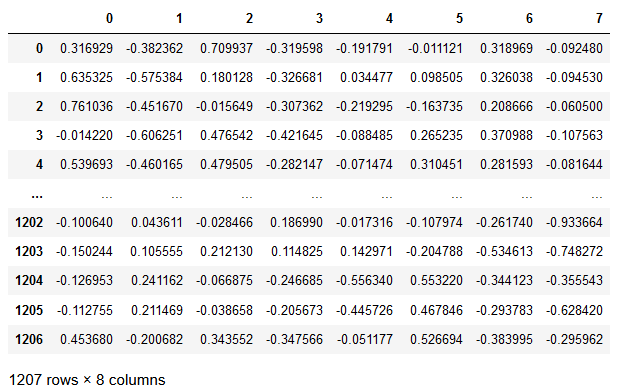




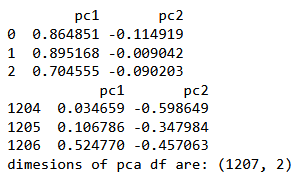
The KNN method resulted in datapoints that are more heterogenous while the matrix factorization methods, NMF and SVD, are closer together.

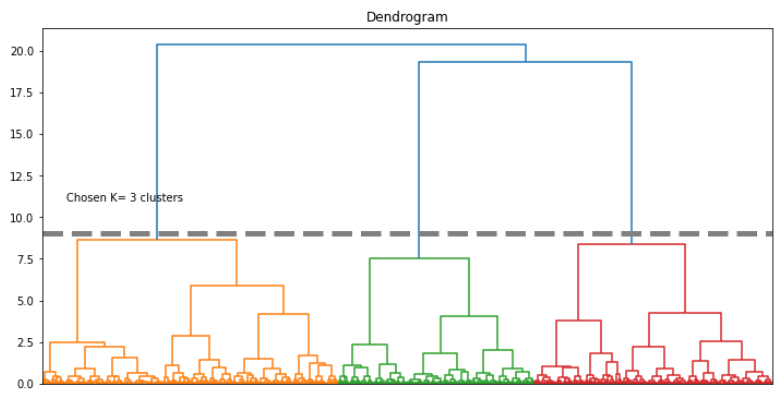
6. Clustering

Now that we have our three imputed datasets, we can do some clustering analysis to visualize the clusters to see whether there are similarities between biochar materials. First, we will use the dataset imputed by KNN and fit the scaler to the data and transform it. Then, we will normalize the data so the data is in a common scale from -1 to 1. Here is the output:

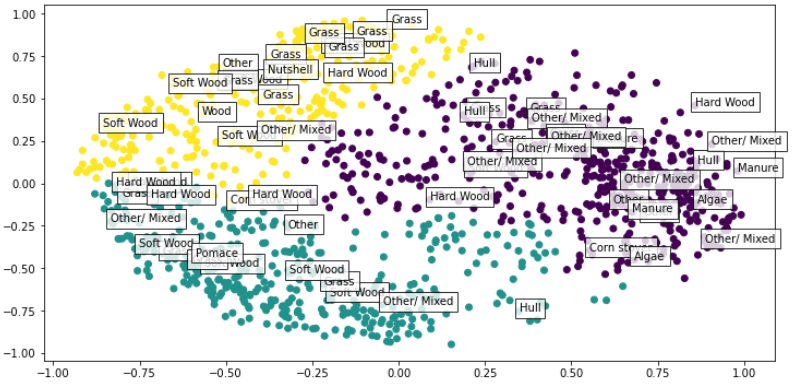


Now we’ll do a principal components analysis to select for two components and visualize in a dendrogram to get the optimal number of clusters. In this case, we get 3 clusters:



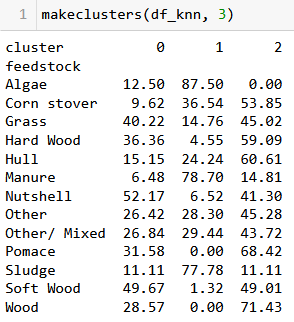


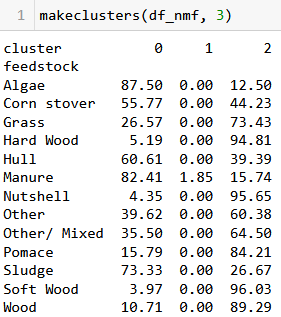
Let’s visualize the individual points to see if we can see a pattern:

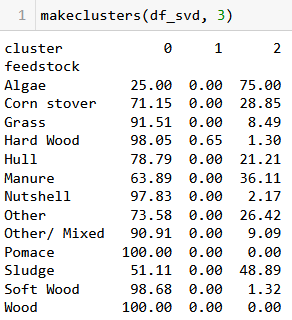


It appears that we have Algae, sludge and Manure in the same cluster, while grass, hard and soft wood are in another cluster. This indicates there are some general similarities between those particular biochar groups in the dataset.

Here we take a look to see all clusters from the 3 imputed datasets and determine if there are strong similarities between certain biochar group:







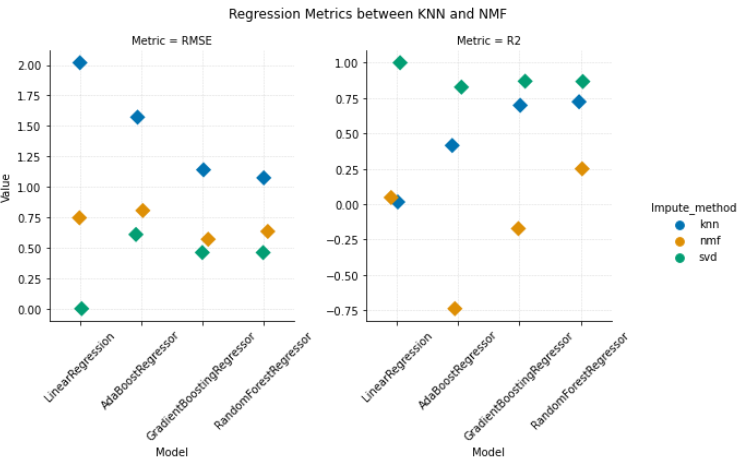
Viewing the data this way, the KNN-imputed dataset has more biochars of the same group appearing in all three clusters making it hard to discern if there are truly predictable similarities between treatment groups. NMF and SVD on the other hand, has most points on the same cluster despite the biochar group. They all are in cluster 0 while almost none are in cluster 1.

7. Evaluation Using Machine Learning and Cross-Validation

We’ve imputed the missing variables using three different methods but the results from these methods should be further evaluated for validity. To assess how well soil nitrogen retention is predicted on new data, we will use the following machine learning models: linear regression, adaboost regression, gradient boosting regression and random forest regression. We will run a 10-kfold cross validation on these regressors and get the root mean squared error (RMSE) and coefficient of determination (R^2).

Linear regression is a statical method for modeling the relationship between the dependent variable and independent variables. Adaboost is a machine learning algorithm that combines multiple weak regression models into a strong predictive model. Gradient Boosting is an ensemble learning technique that building the model in an iterative manner by sequentially adding weak models to minimize the loss function. Finally, random forest is another tree ensemble method that combines multiple decision trees to create a robust regression model. All the above models can be used for regression or classification tasks.

RMSE is a measure of the average deviation from the predicted values of a model and the actual observed values. Lower RMSE indicate better model performance as it signifies smaller errors. R^2 is a statistical measure that represents the proportion of variance in the dependent variable that can be explained by the independent variables in a regression model. The value of R^2 ranges from 0 to 1 with high values representing a better fit.

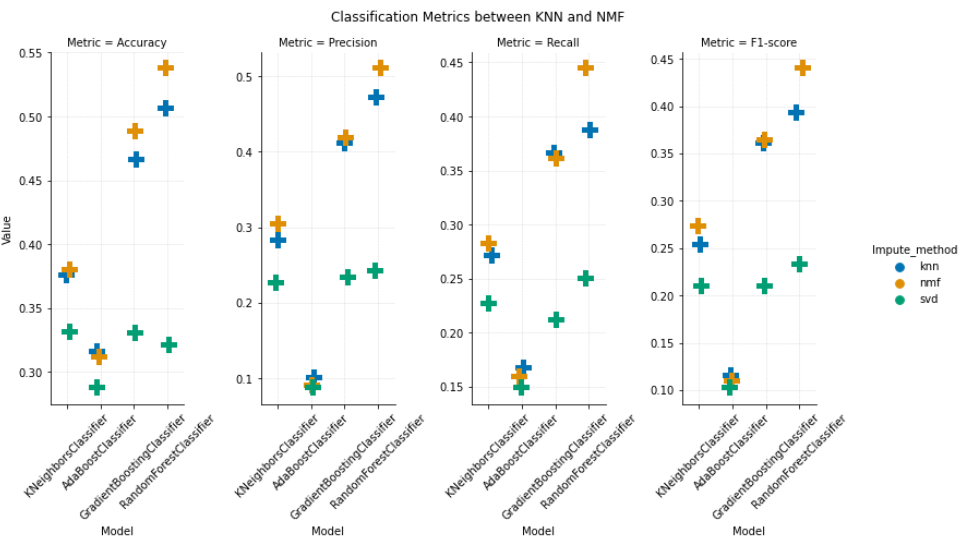


The above plots show the RMSE and R^2 values respectively for each machine learning model and the imputation methods. The KNN method has the highest RMSE across all ML models and above average R^2 especially for gradient boost and random forest. SVD has the lowest RMSE and highest correlating coefficient indicating a very high fit.

Similarly, we can use classification models and evaluate their performance via measures for accuracy, precision, recall and F1-score.

Accuracy is measuring the overall correctness of a model’s prediction. Precision focuses on the positive predictions made by the model and calculates the ration of true positives to the sum of the true positives and false positives. Recall, also known as sensitivity, measures the proportion of true positives predicted by the model out of all the actual positive instances in the dataset. F1-score ranges from 0-1 and is the mean of both precision and recall and balances the trade-off between the two. A score of 1 indicates perfect precision and recall

The following models were used for classification: K-nearest neighbors, adaboot, gradient boot and random forest classifier. We will run a 10-kfold cross validation with these classifiers predicting on the feedstock label. We produce the results below:



We can see that although the SVD method had the lowest RMSE and highest R^2 in the regressor models, it had the lowest performance measures across all ML classifier models. This is because the model may be overfitting. The KNN and NMF methods on the other hand had fairly decent performance measures overall. It had high accuracy, precision, recall and F1-score.

8. Discussion

In this project, we can see that data mining is an iterative and important process in helping us finds trends, patterns and relationships in the data. It is important to allocate time to find the correct data to perform data wrangling and visualizations. This is necessary to help us answer any related research questions or hypotheses.

In the dataset that was combined with data from the UC Davis sorption database and USDA nitrogen retention data, we can see that there are several missing values. Many of these missing values are the amount of dissolved nitrogen in the soil for various biochar types because they simply did not have experimental data collected for them. Imputation methods such as KNN, NMF and SVD helped to predict these missing values which gave us additional insight on the degree of similarities between certain biochar types.

Using PCA and clustering methods, the KNN shows that biochars Algae, Manure, and Sludge are mostly in the same cluster at 87.50%, 78.70% and 77.78% respectively. Interestingly NMF also shows that Algae, Manure and Sludge are in the same cluster with points within the cluster at 87.50%, 82.41%, and 73.33%. Also, in KNN, Corn Stover, Hard Wood, Hull, Pomace and Wood have more than 50% of their datapoints in the same cluster whereas in NMF, Grass, Hard Wood, Nutshell, Pomace, Soft Wood and Wood have more than 50% of their data points in the same cluster indicating some similarity in clusters using two distinct imputation methods. SVD however shoes most of the data points in the same cluster, indicating poor differentiation between the biochar materials.

Upon evaluating the data for predictive performance using regression models, SVD overall had the lowest RMSE across all machine learning models used and highest R^2. However, this is likely the case for overfitting since in the classification cross-validation models, SVD showed the lowest predictive performance metrics with all metrics including accuracy, precision, recall and F1-score below 25%. KNN and NMF data were similar using the classification methods, with highest predictive performance being for the ensemble methods, specifically gradient boosting and random forest.

Although the results look promising, the methods used were executed based on several underlying assumptions such that the properties of the biochar groups follow a predictable pattern and there are no interactions between particular biochar traits. Additionally, the models used could be tuned further to attain better predictive performance and reduce the risk of overfitting. Also, the quality of the data itself can be further examined before attempting to perform appropriate statistical analyses.

9. Conclusion

This project highlights the importance data mining and machine learning to understand the properties of biochar and its soil nitrogen retention potential. With techniques such as k-nearest neighbors, non-negative matrix factorization, and singular value decomposition to impute missing values in the dataset, we can predict the nitrogen retention of various new types of biochar sources. This is useful because it helps utilize and gather additional insights from already available data and helps us save time, resources, and others costs of running actual experiments in the lab or in the field. Furthermore, dimensionality reduction and clustering methods helps us visualize and identify possible similarities in biochar materials which has potential real-world uses such as identifying similar but more cost effective-or more sustainable biochar alternatives.

This project not only showcases the valuable application of data mining in unraveling the properties of biochar but also emphasizes its indispensable role in comprehending the limitless possibilities of biochar utilization.

ACKNOWLEDGMENTS

I would like to thank the contributors of the UC Davis Sorption database as well as USDA for making these data used in the analysis public available.

REFERENCES

[1] United Nations Department of Economic and Social Affairs. "World

Population Projected to Reach 9.8 Billion in 2050 and 11.2 Billion in 2100." United Nations, www.un.org/en/desa/world-population-projected-reach-98-billion-2050-and-112-billion-2100. Accessed 6 May 2023.

[2] Biochar Adsorption Reference Database, Aqueous Solutions [www.aqsolutions.org ], 2016.

[3] Malyan et al. “Biochar for environmental sustainability in the energy-water-agroecosystem nexus”, Renewable and Sustainable Energy Reviews, Volume 149, 2021, 111379, ISSN 1364-0321, https://doi.org/10.1016/j.rser.2021.111379.

[4] Mosa, Ahmed, et al. "Biochar as a Soil Amendment for Restraining Greenhouse Gases Emission and Improving Soil Carbon Sink: Current Situation and Ways Forward." Sustainability, vol. 15, no. 2, 2023, article 1206. DOI: 10.3390/su15021206.

[5] (original article) Phillips, C. L., Meyer, K. M., Garcia-Jaramillo, M., Weidman, C. S., Stewart, C. E., Wanzek, T., Grusak, M. A., Watts, D. W., Novak, J., & Trippe, K. M. (2022). Towards predicting biochar impacts on plant-available soil nitrogen content. Biochar, 4(1). https://doi.org/10.1007/s42773-022-00137-2

[6] (dataset) Phillips, Claire L.; Meyer, Kylie M.; Garcia-Jaramillo, Manuel; Weidman, Clara S.; Stewart, Catherine E.; Wanzek, Thomas; Grusak, Michael A.; Watts, Donald W.; Novak, Jeff M.; Trippe, Kristin M. (2021). Data from: Towards predicting biochar impacts on plant-available soil nitrogen content. Ag Data Commons. https://doi.org/10.15482/USDA.ADC/1523372. Accessed 2023-06-18.

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00