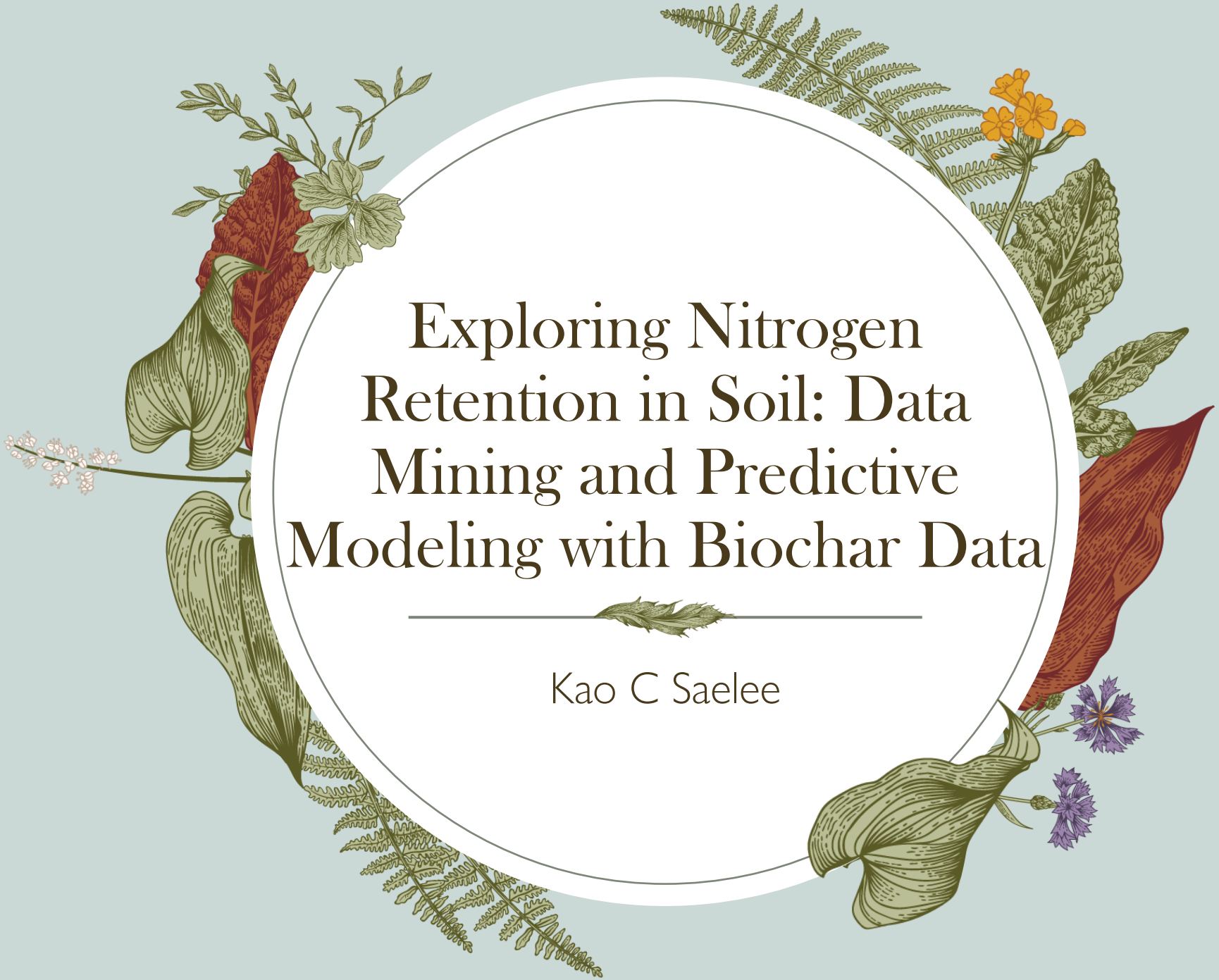


# Exploring Nitrogen Retention in Soil: Data Mining and Predictive Modeling with Biochar Data

Kao C Saelee

# Agenda

Executive Summary

Problem Statement

Related work

Proposed work

Evaluation

Discussion

Conclusion

# Executive Summary

# Executive Summary

- 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.

# Problem Statement

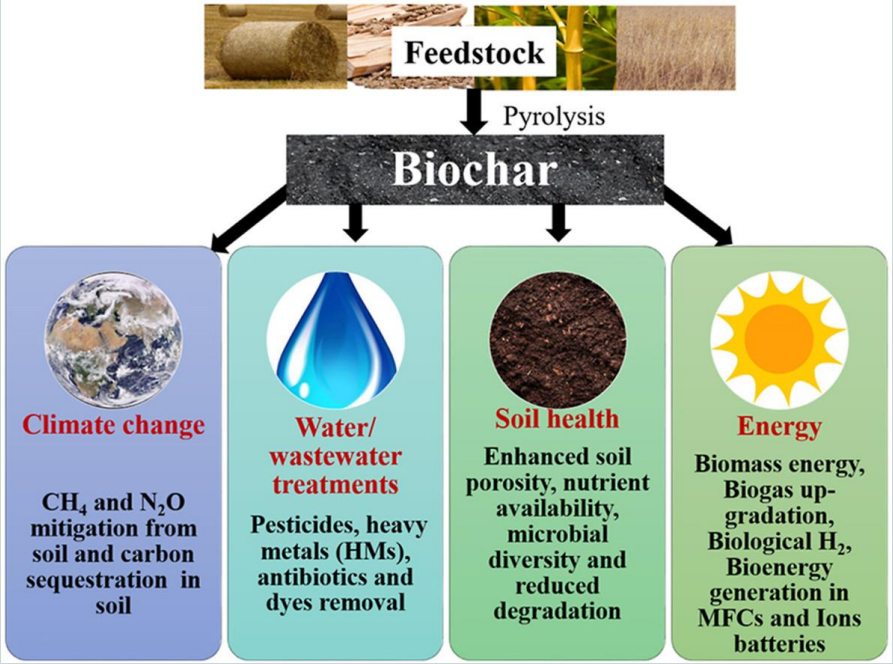
# Problem Statement

- What is the problem?
  - There needs to be more sustainable ways to increase soil fertility to produce food for a growing population. There has been growing interest for the use of biochar as a sustainable solution. Although there are numerous literatures on biochar's impact on soil fertility, there is however is limited data that is publicly available on soil nitrogen retention of some biochar sources.
- Why is it important?
  - 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. The use of biochar sources is a sustainable way to improve soil fertility by increasing nitrogen retention. Being able to understand the characteristics of various biochar would help us utilize these sources most effectively in agriculture.
- Limitations to existing solutions
  - Precision agriculture: costly and not financially feasible for all farmers
  - Cover cropping: requires significant land area to be effective
  - Chemical fertilizers: contribute to nutrient runoff and water pollution
- Potential solution/ contribution
  - One such approach that has gained significant attention in recent years is biochar.
  - One important aspect of biochar's impact on soil fertility 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 Not only is the environmental impact positive by reducing nitrogen pollution in waterways and reducing N20 greenhouse gas emissions, the enhanced soil fertility with biochar will increase the efficiency of nitrogen use by plants and improve crop yields and quality.

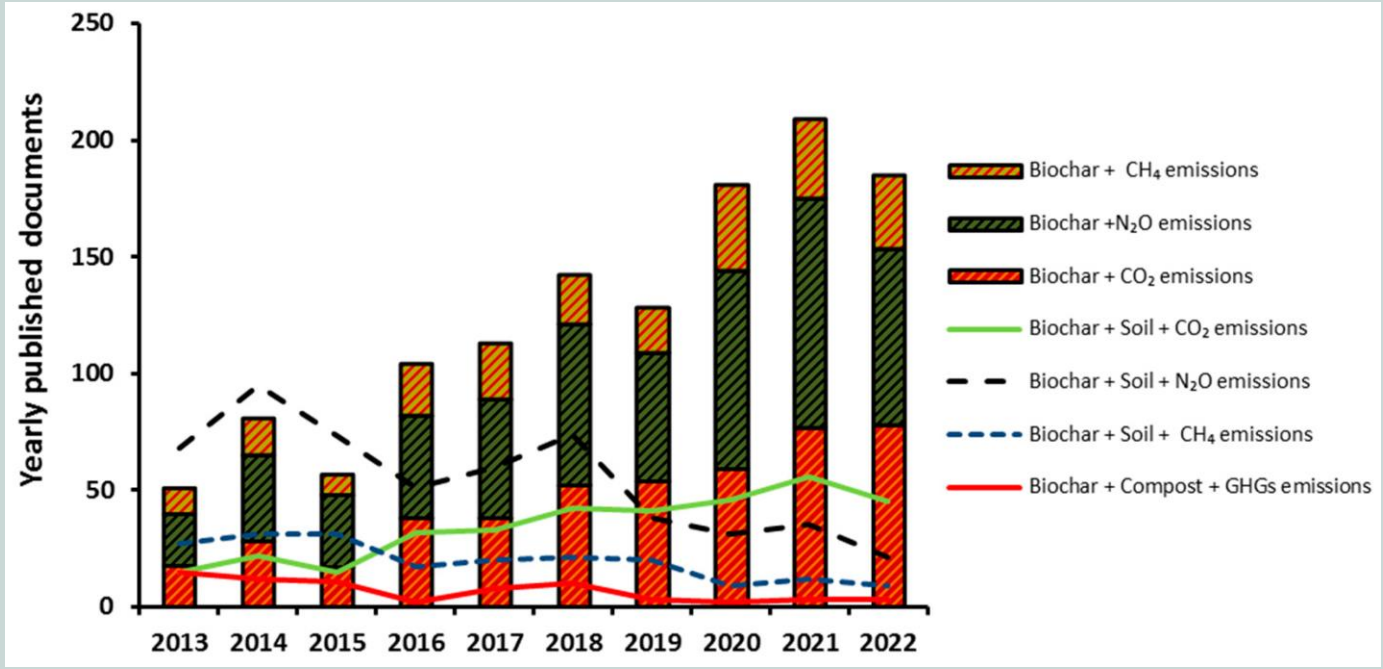
# Related Work

# Related Work

- Previous research has also used various visualization techniques to present the data, including graphs, 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.




Source: 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.



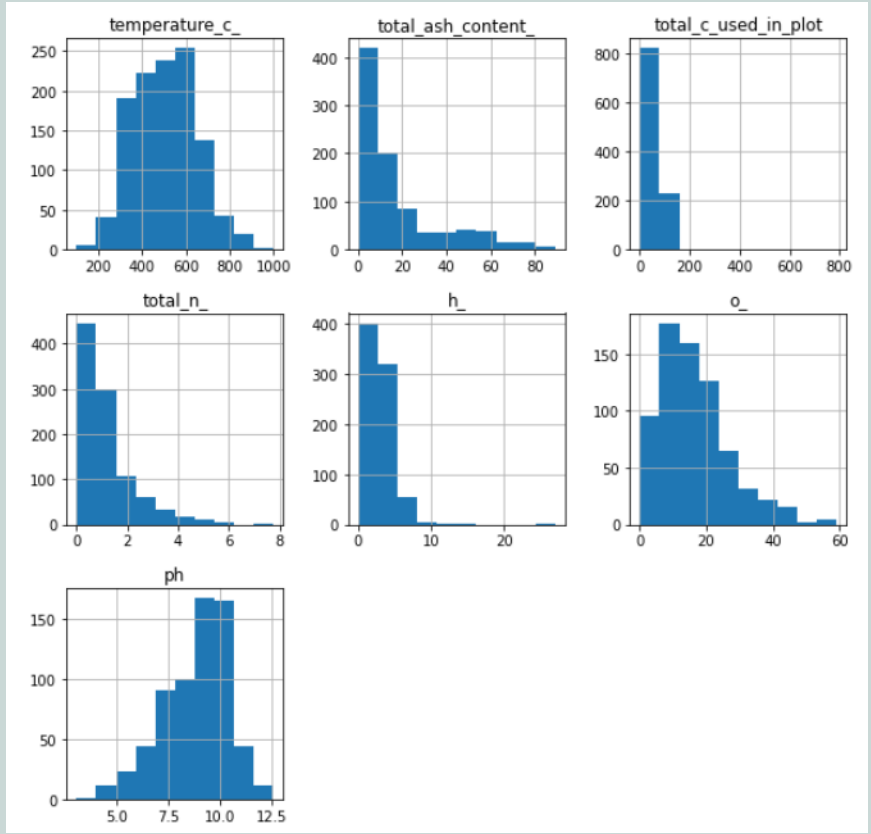

Source: Mosa, Ahmed, et al. "Biochar as a Soil Amendment for Restraining Greenhouse Gases Emission and Improving Soil Carbon Sink: Current Situation and Ways Forward." Sustainability, vol. 15, no. 2, 2023, article 1206. DOI: 10.3390/su15021206.

# Proposed Work

# Data Collection and Tools

- We began with the data collection. The datasets were analyzed by two sources The UC Davis Sorption Database and the USDA website. The data was downloaded from the cite and analyzed.

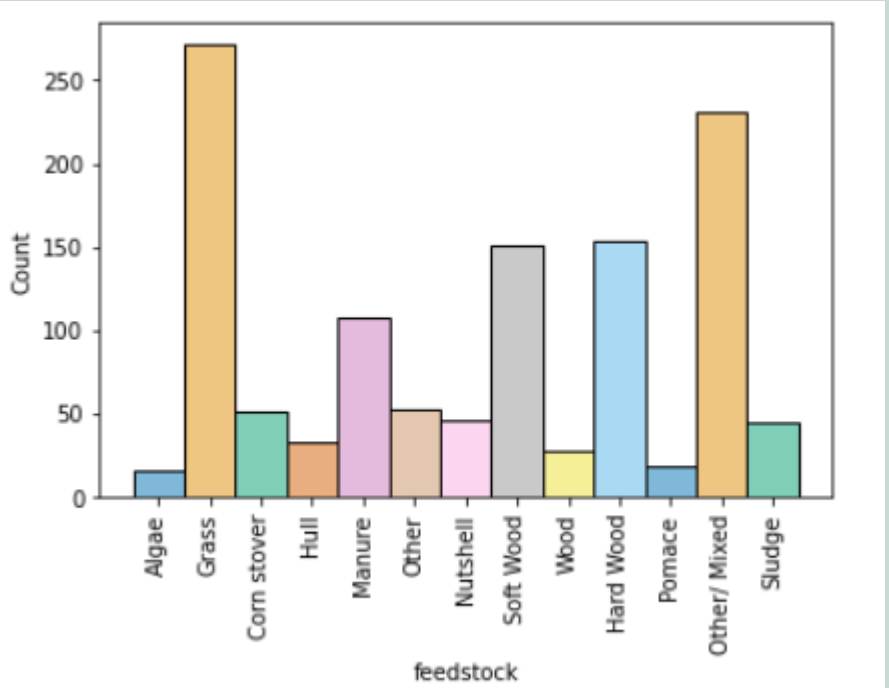
| | Biochar Name | Feedstock Composition | Optional comments about Feedstock | Commercial | Geographic Region | Manufacturer | Pyrolysis Method | Source: Peer Reviewed Publication | Reference |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Cladophora coelothrix1 | Algae | NaN | Non commercial | NaN | NaN | NaN | Yes | Mutanda et al., Bioresource Technology, 102, ... |
| 1 | Cladophora patentiramea1 | Algae | NaN | Non commercial | NaN | NaN | NaN | Yes | Mutanda et al., Bioresource Technology, 102, ... |
| 2 | Chaetomorpha indica1 | Algae | NaN | Non commercial | NaN | NaN | NaN | Yes | Mutanda et al., Bioresource Technology, 102, ... |
| 3 | Chaetomorpha linum1 | Algae | NaN | Non commercial | NaN | NaN | NaN | Yes | Mutanda et al., Bioresource Technology, 102, ... |
| 4 | Cladophoropsis sp.1 | Algae | NaN | Non commercial | NaN | NaN | NaN | Yes | Mutanda et al., Bioresource Technology, 102, ... |



# Exploratory Analysis and Visualization

- The datasets were analyzed including checking the shape and features and handling missing values. We merged the datasets and made graphs to visualize the data to check for patterns and increase our understanding of the data.

| | feedstock | ash | c | n | h | o | ph | delta_day14_nmin | delta_day28_nmin |
|---|---|---|---|---|---|---|---|---|---|
| 0 | Grass | 51.300 | 55.360 | 0.670 | 1.130 | NaN | 10.790 | 5.00 | 7.46 |
| 1 | Grass | NaN | 56.520 | 0.670 | 1.160 | NaN | 10.830 | 6.61 | 7.64 |
| 2 | Grass | NaN | 56.060 | 0.640 | 1.110 | NaN | 10.880 | 5.57 | 4.68 |
| 3 | Soft Wood | 6.070 | 79.650 | 0.380 | 0.510 | NaN | 9.590 | 7.39 | 7.23 |
| 4 | Soft Wood | 5.410 | 79.810 | 0.380 | 0.670 | NaN | 9.690 | 6.17 | 4.68 |
| 5 | Soft Wood | 4.660 | 79.650 | 0.370 | 0.790 | NaN | 9.710 | 5.73 | 3.88 |
| 6 | Other | 19.460 | 81.060 | 0.570 | 1.330 | NaN | 8.060 | 4.74 | 9.02 |
| 7 | Other | NaN | 81.830 | 0.520 | 1.240 | NaN | 8.050 | 5.10 | 12.68 |
| 8 | Other | 16.700 | 88.780 | 0.610 | 1.660 | NaN | 8.160 | 6.53 | 12.81 |
| 9 | Grass | 14.082 | 65.488 | 1.178 | 4.786 | 14.13 | 10.148 | 0.74 | 7.73 |

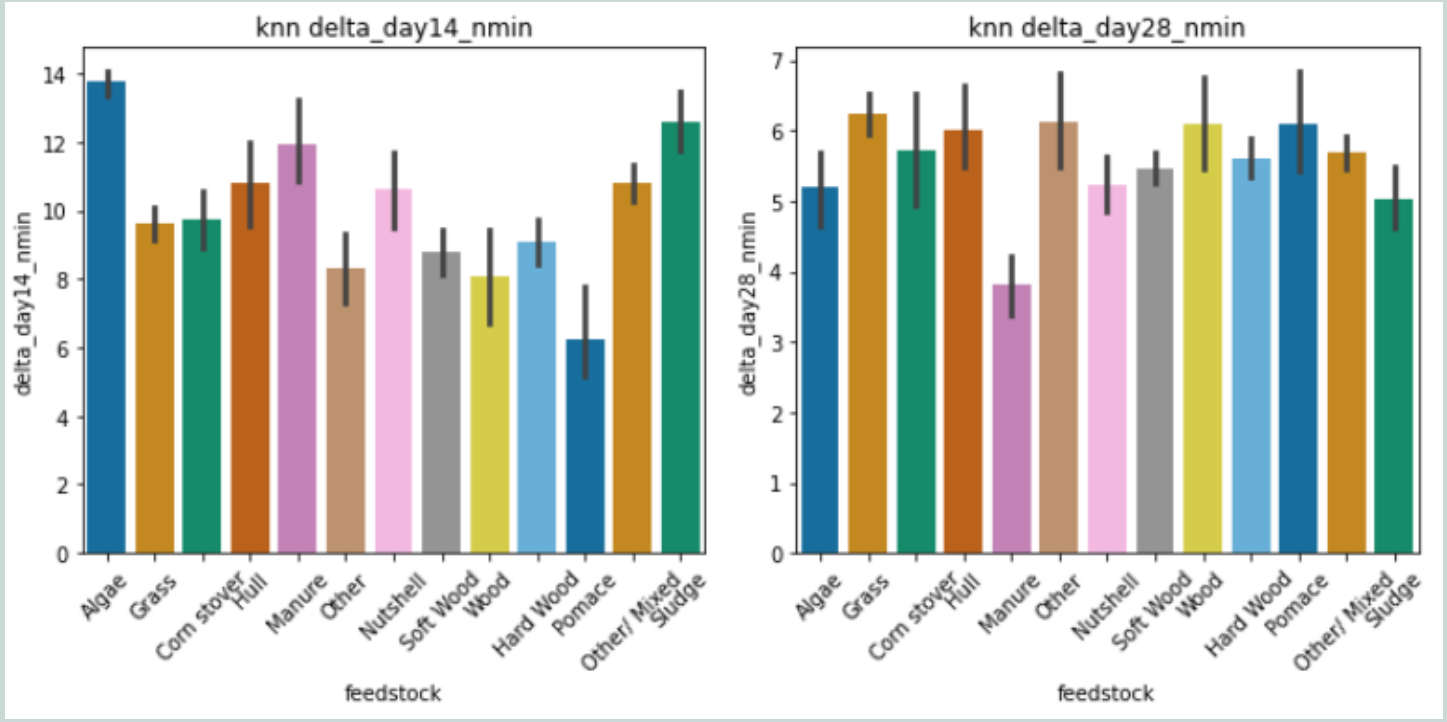


# Data Imputation and Preprocessing

- Imputations methods such as K-nearest neighbors (KNN), Non-Negative matrix factorization (NMF) and Singular Value Decomposition (SVD) were used to impute missing values. An example of the KNN imputation results are displayed below:

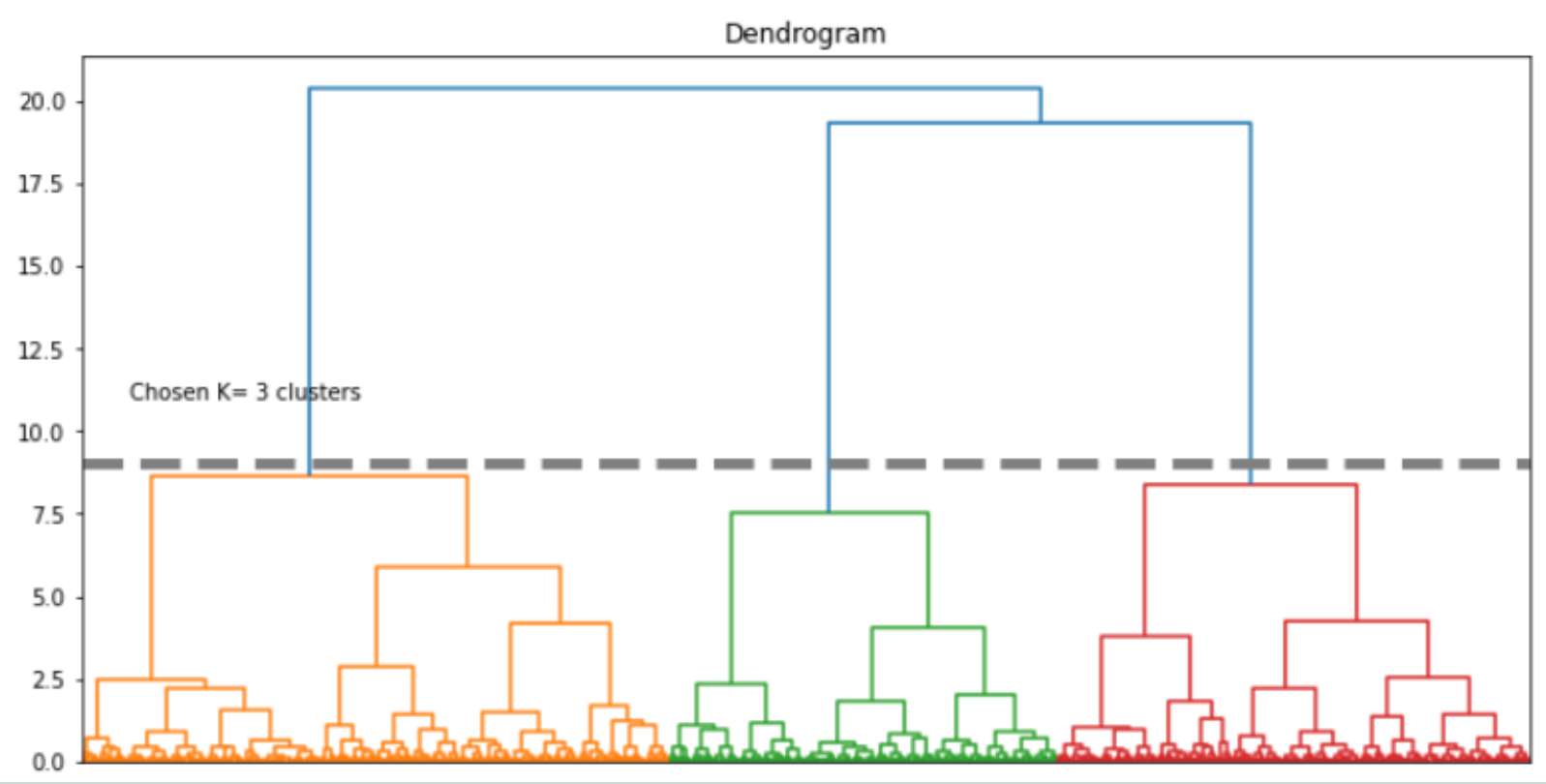
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1207 entries, 0 to 1206
Data columns (total 9 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   feedstock         1207 non-null   object
 1   ash               1207 non-null   float64
 2   c                 1207 non-null   float64
 3   n                 1207 non-null   float64
 4   h                 1207 non-null   float64
 5   o                 1207 non-null   float64
 6   ph                1207 non-null   float64
 7   delta_day14_nmin  1207 non-null   float64
 8   delta_day28_nmin  1207 non-null   float64
dtypes: float64(8), object(1)
memory usage: 85.0+ KB
```

| | feedstock | ash | c | n | h | o | ph | delta_day14_nmin |
|---|---|---|---|---|---|---|---|---|
| 0 | Algae | 32.100000 | 34.6 | 3.300000 | 1.50 | 11.680000 | 8.720000 | 13.955714 |
| 1 | Algae | 47.000000 | 20.3 | 1.700000 | 1.50 | 17.354286 | 9.120000 | 13.955714 |
| 2 | Algae | 73.500000 | 10.2 | 1.100000 | 0.80 | 8.071286 | 7.830000 | 13.955714 |
| 3 | Algae | 16.000000 | 23.6 | 2.400000 | 1.30 | 14.591429 | 9.610000 | 13.955714 |
| 4 | Algae | 46.500000 | 23.6 | 2.800000 | 1.50 | 14.468571 | 10.070000 | 13.955714 |



# PCA and Clustering

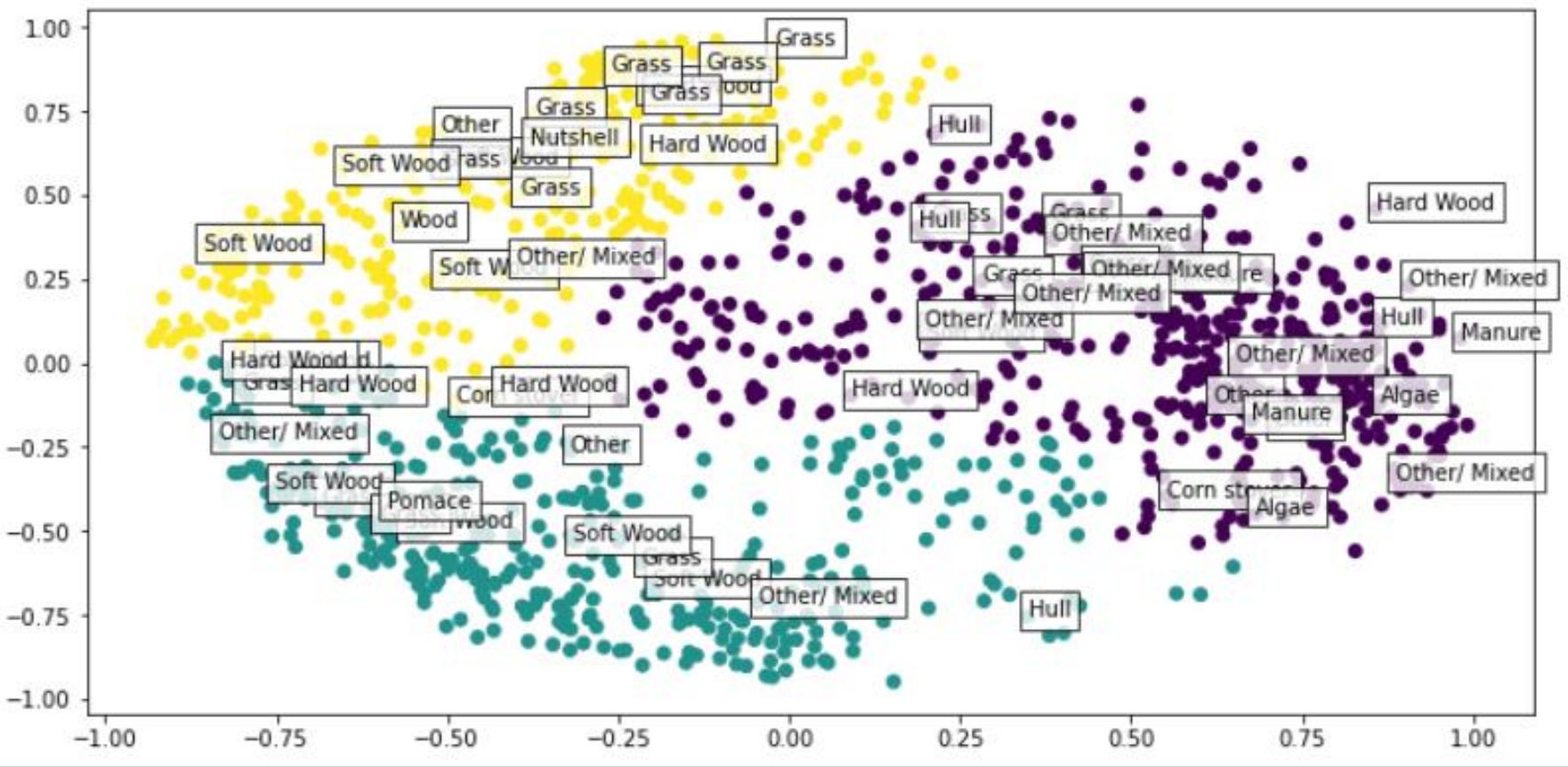
- We've done a principal components analysis to select for two components and visualize in a dendrogram to get the optimal number of clusters. In this case, we get 3 clusters.




```
         pc1       pc2
0   0.864851 -0.114919
1   0.895168 -0.009042
2   0.704555 -0.090203
            pc1       pc2
1204   0.034659 -0.598649
1205   0.106786 -0.347984
1206   0.524770 -0.457063
dimesions of pca df are: (1207, 2)
```

# Clustering (Continued)

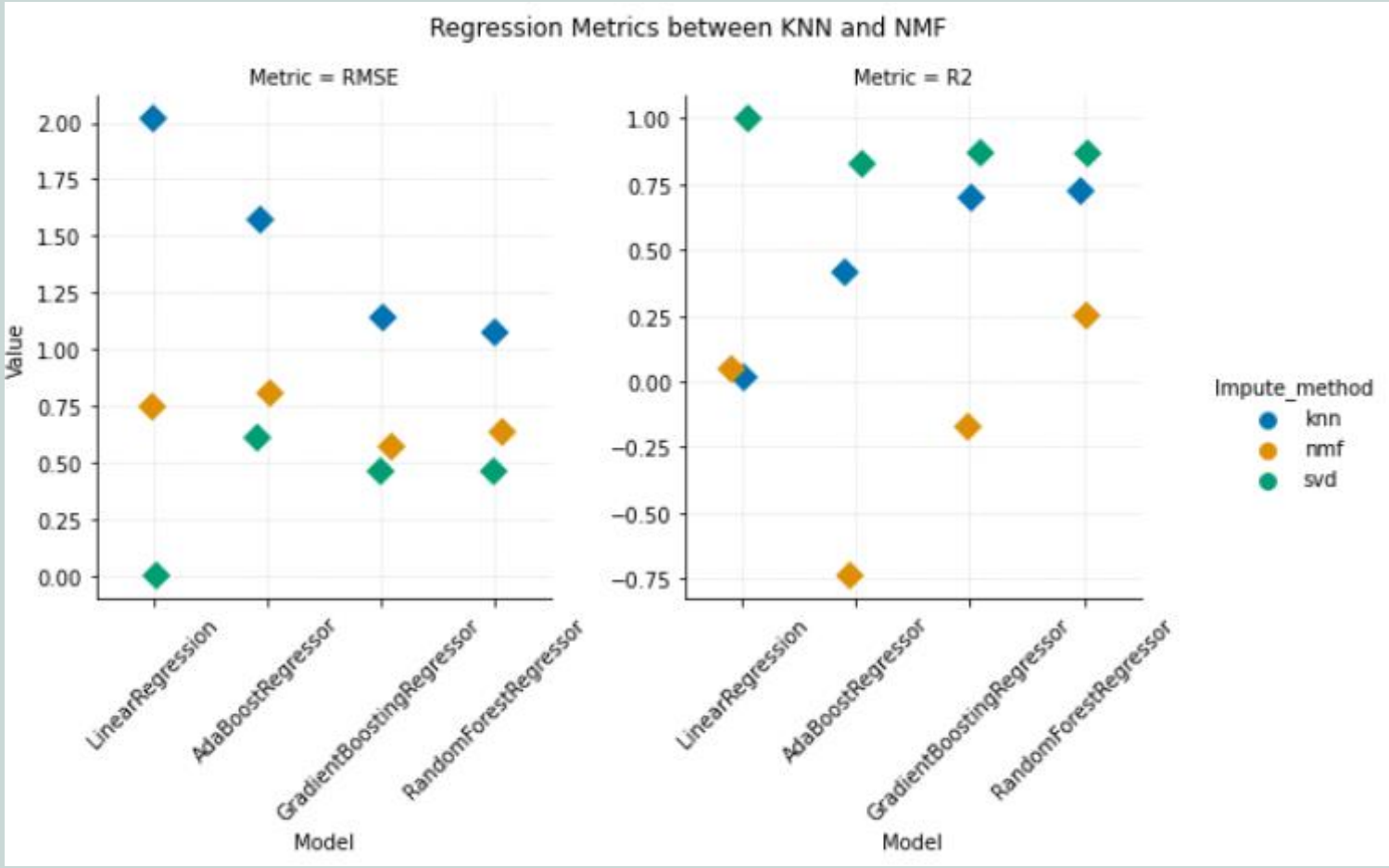
- Furthermore, the data can be visualized to see if there are any interesting trends or patterns within or between the clusters. Below is the PCA plot of the KNN imputed dataset.



# Evaluation

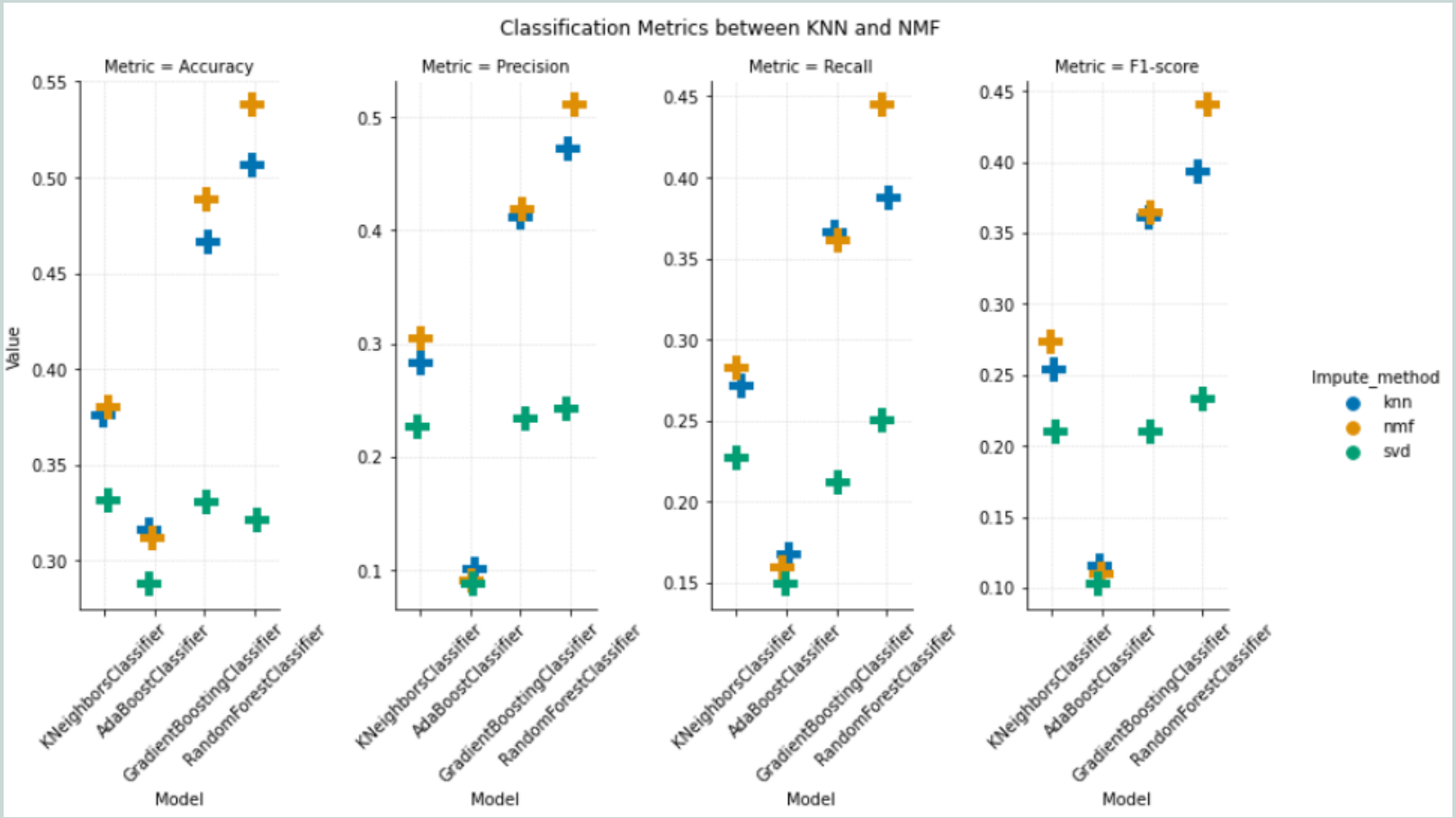
# Evaluation with Cross-Validation

- To assess how well soil nitrogen retention is predicted on new data, we used the following machine learning models: linear regression, adaboost regression, gradient boosting regression and random forest regression.
- A 10-kfold cross validation on these regressors and get the root mean squared error (RMSE) and coefficient of determination (R^2) was run.



# Evaluation with Cross-Validation (Continued)

- Similarly, we used classification models and evaluated their performance via measures for accuracy, precision, recall and F1-score.
- The following models were used for classification: K-nearest neighbors, adaboot, gradient boot and random forest classifier.
- A 10-kfold cross validation with these classifiers predicting on the feedstock label was run.



# Discussion

# 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.

- 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.

```
1 makeclusters(df_knn, 3)
```

| cluster | 0 | 1 | 2 |
|---|---|---|---|
| feedstock | | | |
| Algae | 12.50 | 87.50 | 0.00 |
| Corn stover | 9.62 | 36.54 | 53.85 |
| Grass | 40.22 | 14.76 | 45.02 |
| Hard Wood | 36.36 | 4.55 | 59.09 |
| Hull | 15.15 | 24.24 | 60.61 |
| Manure | 6.48 | 78.70 | 14.81 |
| Nutshell | 52.17 | 6.52 | 41.30 |
| Other | 26.42 | 28.30 | 45.28 |
| Other/ Mixed | 26.84 | 29.44 | 43.72 |
| Pomace | 31.58 | 0.00 | 68.42 |
| Sludge | 11.11 | 77.78 | 11.11 |
| Soft Wood | 49.67 | 1.32 | 49.01 |
| Wood | 28.57 | 0.00 | 71.43 |

# Discussion (Continued)

- 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.

# Conclusion

# Conclusion

- This project highlights the importance data mining and machine learning to understand the properties of biochar and its soil nitrogen retention potential. We can use these methods to 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.

# Thank You!

Image credit: https://www.odevatagardshotell.se/en/what-is-biochar-how-is-biochar-made-and-its-benefits/