Transformational Pathways of **Household Farms in Tanzania** Based on **Machine Learning Analysis**: **Key Factors** from income structure to sustainable practices

Rui-An Lin, Fabian Steinmetz, Hwong-Wen Ma*

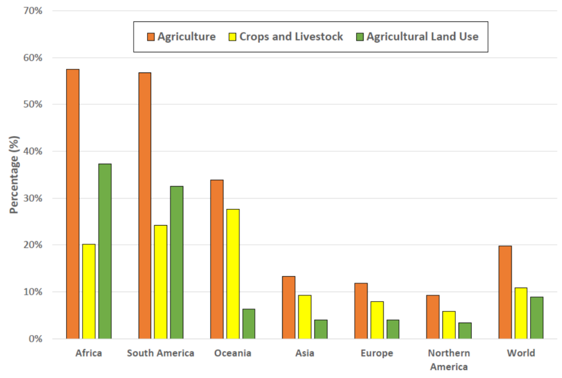
2024.12.20

1. Introduction – Background & Challenge

Background

This project investigates farm-level and household characteristics to **assess sustainability** (structural and practical) and evaluate how Machine Learning can play a role in **deriving targeted policy recommendations**.
Sustainable practices aim to reduce greenhouse gas emissions, prevent soil degradation, and enhance farms' self-sufficiency, but adoption often depends on overlooked structural conditions.
Pastoral systems, critical to food security, face challenges like enteric fermentation, overgrazing, and soil degradation. Balancing productivity with environmental conservation requires **understanding trade-offs and leveraging synergies**, such as the role of education and technological support. Previous studies on the agricultural sustainability in countries like Argentina used **traditional methods** (e.g. CART) to provide insights into synergies, for instance between mating strategies and decreasing emissions. By incorporating machine learning methods into our analysis, the project **explores multi-causal relationships and identifies strategies** to promote a sustainable and resilient agricultural system.

Figure 1. Agricultural emissions shares of regional total GHG emissions

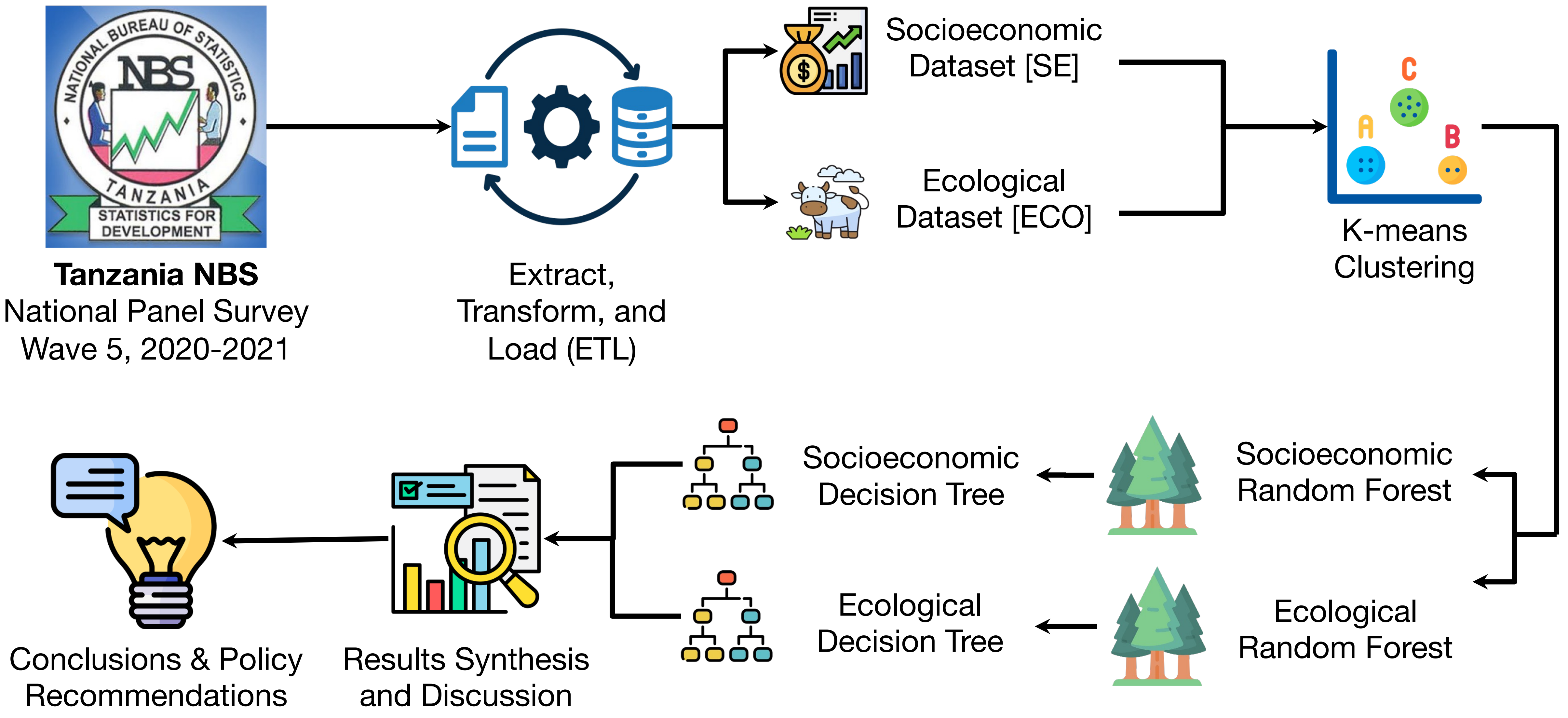


Source: *Food and Agriculture Organization of the United Nations* *(2017)*

Scikit-learn

Machine Learning tools have become widely and publicly accessible, including the scikit-learn library, which is a free and open-source ML-library we used in Python to outperform traditional statistical tools.

2. Method – Research Framework and Analysis Process



This illustration has been designed using images from Flaticon.com

2. Method – NPS Wave-5 ETL

Socioeconomic Features		Agricultural Features			
Code	**Feature Meaning**	**Code**	**Feature Meaning**	**Code**	**Feature Meaning**
SE1	Received agricultural advice	*AG1*	Erosion control methods	*LF1*	Livestock Vaccination
SE2	Received Livestock Advice	*AG2*	Irrigation type	*LF2*	Livestock Deworming
SE3	**Education Level**	*AG3*	Water acquisition method	*LF3*	Preventative measures for livestock
SE4	Media Devices Used	*AG4*	Water Source	*LF4*	Livestock Tick Treatment
SE5	Number of Farm Workers	*AG5*	**Used organic fertilizer**	*LF5*	Curative Treatment for Livestock
SE6	Hired labor farm work	*AG6.1*	Used 2nd type of inorganic fertilizer	*LF6*	Purchased feed/fooder
SE7	Household members in farm work	*AG6.2*	Type of inorganic fertilizer	*LF7*	Livestock watering frequency
SE8	Total Household Members	*AG6.3*	Use a 2nd type of inorganic fertilizer	*LF8*	Livestock Water Sources
SE9	Agricultural Income	*AG7*	Used Pesticides	*LF9*	Livestock housing system
SE10	Other income sources	*AG8.1*	Received seeds on credit	*LF10.1*	Controlled breeding strategies
SE11	Land tenure type	*AG8.2*	Input type (seeds)	*LF10.2*	Breeding strategies used
SE12	**Land size (acres)**	*AG8.3*	Input type (organic fertilizers)	*LF11*	Used Livestock Dung
SE13	Animal stock (LSU)	*AG8.4*	Input type (inorganic fertilizers)	*LF12*	Livestock used for Transportation
SE14	Cattle rearing	*AG8.5*	Input type (pesticides)	*LF13*	**Livestock Ploughing**
SE15	Goat rearing	*AG9*	Used animal traction	*LF14*	Produced Livestock Products
SE16	Sheep rearing	*AG10.1*	**Cultivation Intercropping**		
SE17	Other livestock	*AG10.2*	Reason for intercropping		
SE18	Household members caring for LS	*AG11*	Crop storage method		
SE19	Hired labor for LS	*AG12*	Crop protection method		
SE20	Livestock income	*AG13*	Crop Transport Method		
		AG14	Crop Residue Management		
		AG15	Crop Storage Method		
		AG16	Crop Protection Method		
		AG17	Crop By-products		
		AG18	Owned agricultural tools		
		AG19	Number of agricultural tools		

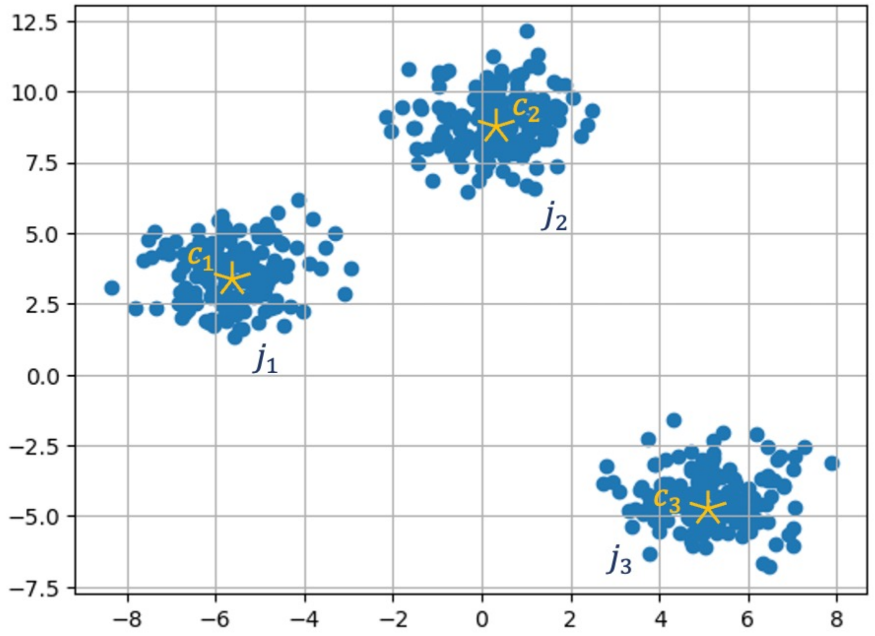
2. Method – K-means Clustering

Goals

Identifying and **classifying** (*quantitatively* and *qualitatively*) farms into clusters to (1) **assess sustainability** (structural and practical) and (2) derive targeted **policy recommendations**.

Determining k

The **Silhouette Coefficient** and **Davis-Bouldin Index** were used to identify the optimal number of clusters, maximizing *separation* and *cohesion*.



Method

Initial centroids are chosen using the ***k-means++*** method, where the first j_1 is selected randomly, and subsequent centroids j_k are the farthest from the ones already chosen. Each data point p is then assigned to the cluster j_k with the nearest centroid c_k. Once assignments are made, new centroids are calculated by averaging the coordinates of all points in each cluster. This process repeats iteratively, maintaining a total of **k** centroids, until the clusters stabilize or convergence is reached (using *max_it* in Python).

Distance:

$$d(p_{ij}) = \sqrt{\Delta x^2 + y^2}$$

New centroid creation:

$$c_j(a) = \frac{1}{n_j} \sum_{x_i \to c_j} x_i(a)$$

Cluster evaluation:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^j - c_j \right\|^2$$

Code Implementation *scikit-learn*

```
# Input Features (without household ID)
features = data_raw.drop(['y5_hhid'], axis=1)

# Setting up the Clustering Method
kmeans = KMeans(n_clusters=4, init='k-means++', max_iter=300, random_state=42)

# Apply kmeans clustering to the chosen features
kmeans.fit(features)

# creating cluster labels (1,2,3...)
labels = kmeans.labels_

# Create a column with the results
data_raw['cluster_label'] = labels
```

2. Method – Random Forest Classification

Goals

(1) Identifying the variables with the **largest effect** on cluster assignment as well as (2) **analyzing the structural components** of the four clusters identified to assess synergies or trade-offs.

Method

RF is an **ensemble learning** method that combines multiple decision trees to make predictions. Unlike a single decision tree, which can overfit the data, Random Forest creates a collection of n trees using **bootstrapped samples**, where each tree is created from a random subset of the data (with replacement). When splitting each node, only a random subset of features is considered, introducing further diversity among trees. Predictions are aggregated through majority voting in classification tasks. Feature Importance (FI) in RF is determined by evaluating the reduction in impurity (e.g. Gini) brought about by each feature, **aggregated across all the trees** in the forest.

Code Implementation *scikit-learn*

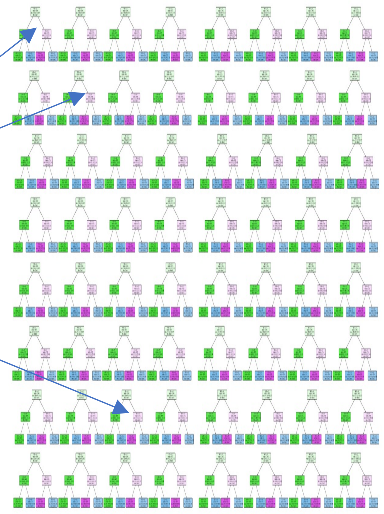
```
# Define input features X and target variable Y
X = data_raw.drop(['y5_hhid', 'cluster_label'], axis=1)
Y = data_raw['cluster_label']

# Set up Random Forest Classification
clf = RandomForestClassifier(n_estimators=1000)

# Apply RF Classification
clf = clf.fit(X,Y)

# Calculate Feature Importance
importance = clf.feature_importances_
```

Tanzania NBS
(bootstrapped samples)

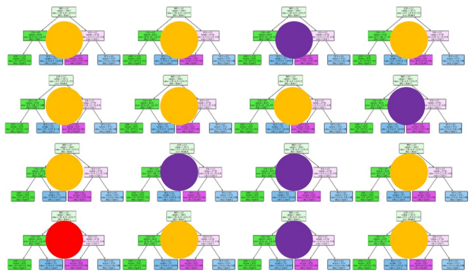


Random Forest

(1) Determining FI

Feature	Importance
A	16.4734
B	14.8405
C	11.4105
D	9.6377
E	7.7580
…	…

(2) Majority Vote Classification



P_i is

3. Result – Key Indicator Overview of Automated Clustering

Results

Two recognized indicators were employed to determine the optimal number of clusters for farm-type classification. Across both indicators, the initialization method (k-means++ vs. random) introduces some variability in clustering outcomes, but ***k-means++* consistently produces superior results**. In contrast, changes to the maximum number of iterations (*max_it*) have minimal impact, suggesting that cluster assignments stabilize even with fewer iterations.

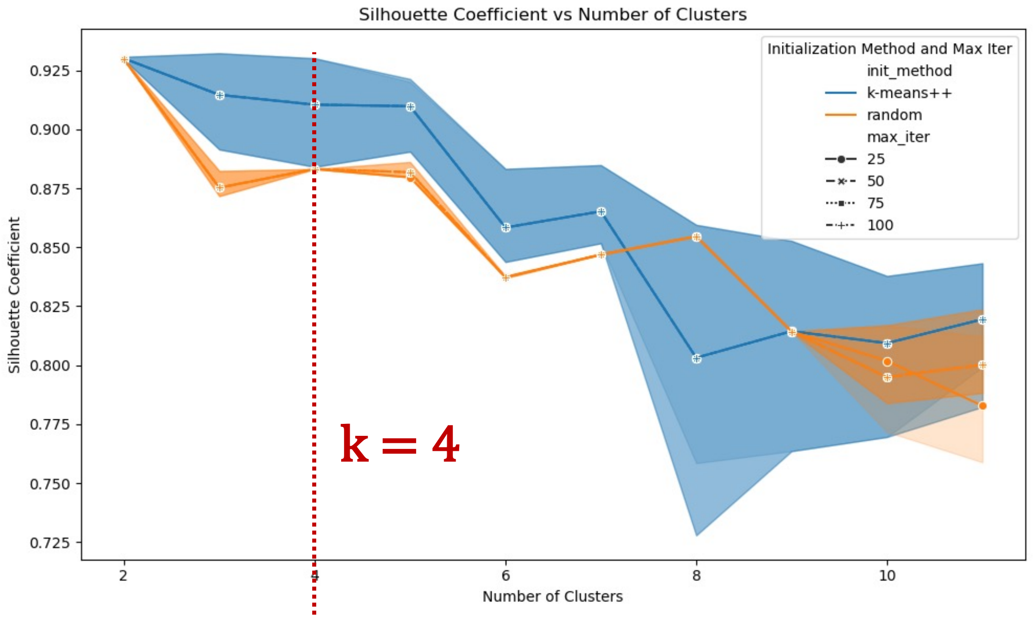
1. Silhouette Coefficient

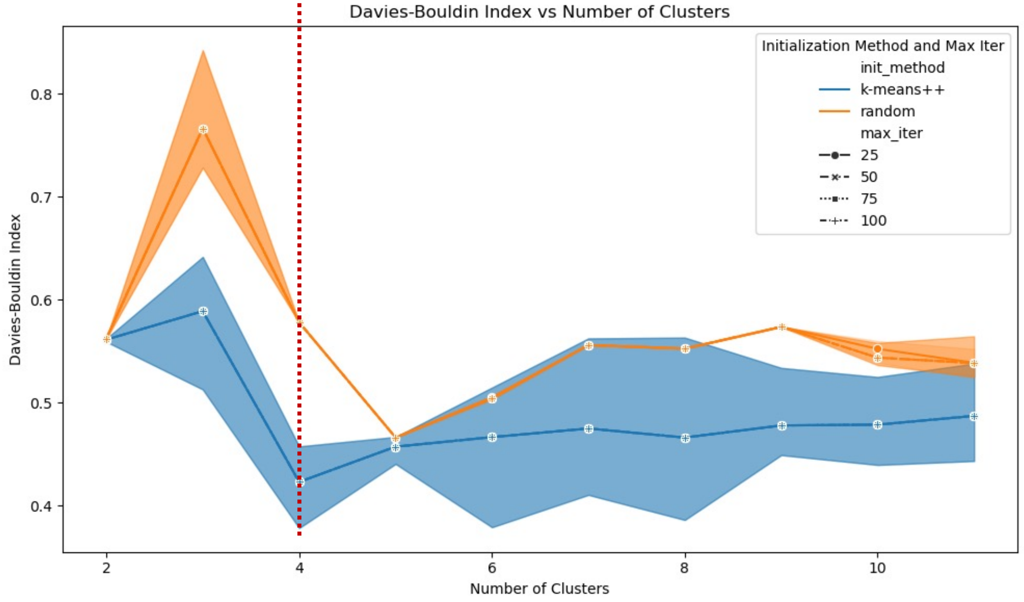
Higher *SC* values are preferable, as they indicate stronger separation between clusters and greater cohesion within clusters. The *SC* values generally decrease as the number of clusters increases, with a pronounced drop after $\boldsymbol{k = 5}$ clusters, suggesting that fewer clusters yield more coherent groupings.

2. Davies-Bouldin Index

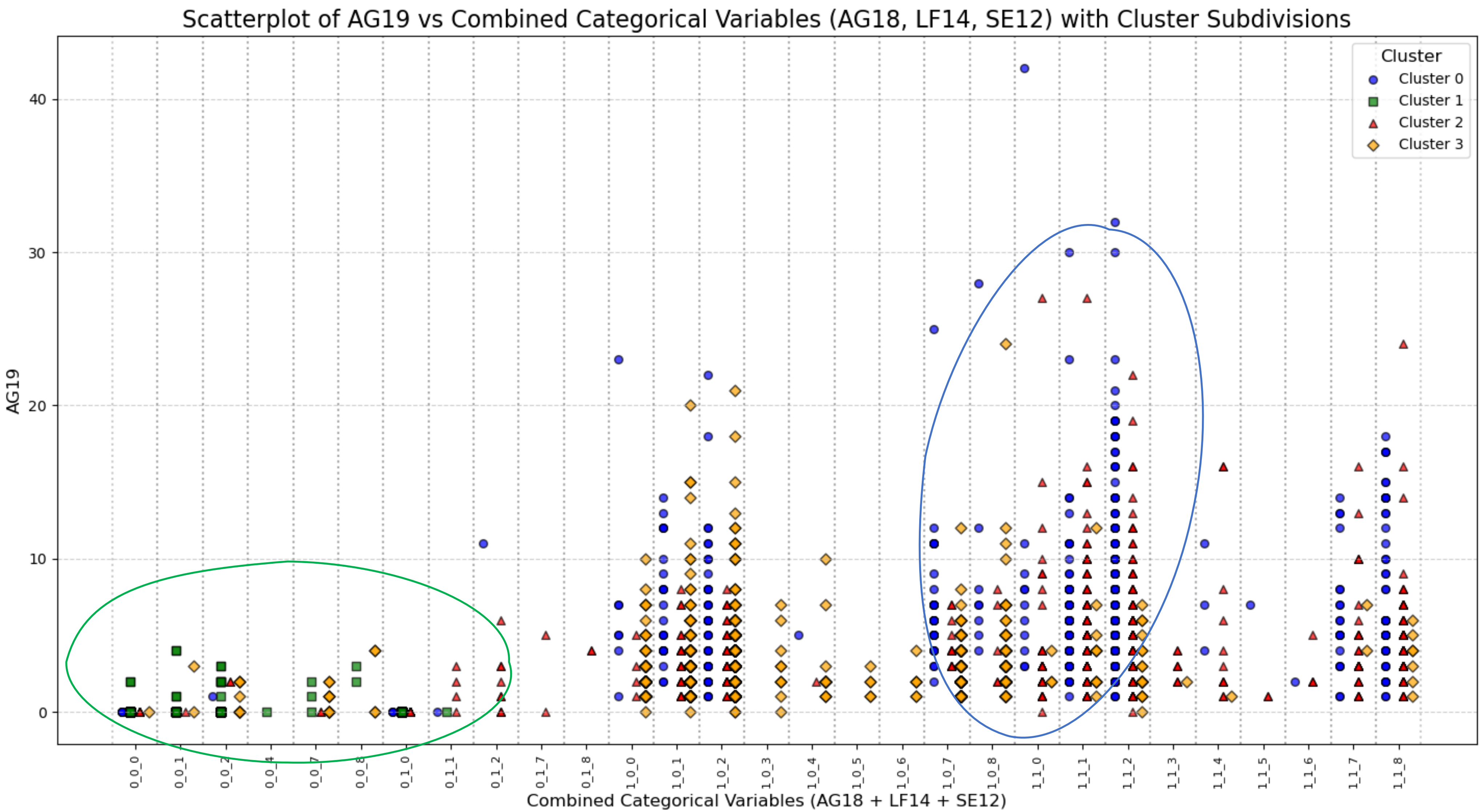
Lower *DBI* values indicate better-defined clusters. For this dataset, *DBI* values consistently support the selection of $\boldsymbol{k = 4}$ to $\boldsymbol{k = 6}$ clusters as the optimal range, highlighting a balance between compactness and separation.

Our choice of $\boldsymbol{k = 4}$ offers a **balance between interpretability and performance**, making it well-suited for segmenting farm types.





3. Result – Scatter plot of farm clusters of SE/ECO Indicators



Legend

AG19: *Number of agricultural tools*

AG18: *Tool Ownership* (Y/N)

LF14: *Livestock Plowing (Y/N)*

SE12: *Land Tenure Condition* (1-8)

3. Result – Descriptive Statistics by Feature Importance

Cluster 0

Household Size μ : **8.43** (1.0~29.0)
Animal Stock μ : **10.16** (0.0~92.22)
Livestock Income μ : **465460.57** (0.0~7170000.0)
No. Agricultural Tools μ : **7.03** (0.0~42.0)
Livestock Plowing *Mo*: Yes (76.39%)

Large, high-income, cattle-farming households

Cluster 1

Household Size μ : **3.97** (1.0~17.0)
Animal Stock μ : 0.04 (0.0~8.4)
Livestock Income μ : **6275.74** (0.0~2325000.0)
No. Agricultural Tools μ : **0.02** (0.0~4.0)
Livestock Plowing *Mo*: No (97.65%)

Small, low-income, ill-equipped, agricultural households

Cluster 2

Household Size μ : **5.34** (1.0~17.0)
Animal Stock μ : **0.43** (0.0~13.12))
Livestock Income μ : **164318.16** (0.0~43635000.0)
No. Agricultural Tools μ : **10.7** (0.0~5000.0)
Media Use μ: **2.8** (0.0~14.0)

Medium-sized to large, technological, wealthy farms with livestock farming but less sustainable practices

Cluster 3

Household Size μ : **4.82** (1.0~18.0)
Animal Stock μ : **0.43** (0.0~13.12)
Livestock Income μ : **14107.28** (0.0~4480000.0)´
No. Agricultural Tools μ : **2.87** (0.0~24.0)
Land Tenure Type *Mo*: **Freehold** (43.22%)

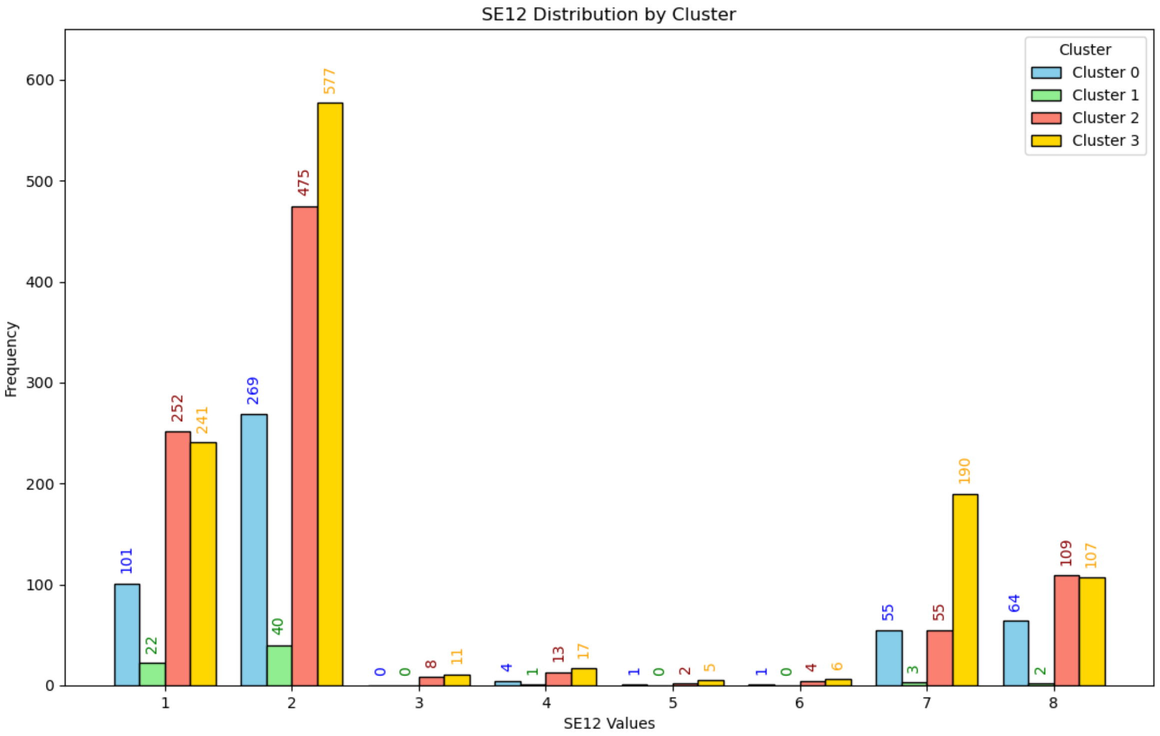
Smaller, (family), wealthy farms with reliance on agriculture

Feature Description	Importance	Cluster 0	Cluster 1	Cluster 2	Cluster 3
	Count	1826	1263	1099	521
Number of agricultural tools	16.4734	7.03 (0.0~42.0)	0.02 (0.0~4.0)	10.7 (0.0~5000.0)	2.87 (0.0~24.0)
Agricultural tool ownership	14.8405	1 (97.89%)	0 (100.0%)	1 (86.35%)	1 (97.94%)
Livestock Plowing	11.4105	1 (76.39%)	0 (97.65%)	1 (89.9%)	0 (96.28%)
Livestock Housing System	9.6377	11 (87.52%)	0 (92.17%)	2 (42.04%)	0 (77.12%)
Livestock watering frequency	7.7580	7 (75.62%)	0 (92.17%)	5 (31.67%)	0 (77.12%)
Crop residue treatment	5.8543	9 (39.73%)	0 (98.52%)	2 (41.4%)	2 (45.05%)
Livestock water sources	4.3584	4 (20.54%)	0 (95.13%)	1 (27.21%)	0 (87.65%)
Agricultural by-products	3.8630	19 (55.85%)	0 (98.96%)	19 (36.67%)	0 (39.98%)
Livestock use for transportation	3.0956	1 (58.16%)	0 (99.95%)	0 (99.09%)	0 (98.89%)
Use of intercropping	2.8353	0 (41.27%)	0 (98.63%)	0 (48.13%)	0 (41.41%)

Description	Importance	Cluster 0	Cluster 1	Cluster 2	Cluster 3
	Count	1826	1263	1099	521
Land tenure condition	22.2430	2 (51.63%)	0 (96.28%)	2 (43.22%)	2 (45.68%)
Land tenure size	14.3911	10.5 (0.0~500.0)	0.13 (0.0~100.0)	3.88 (0.0~103.0)	3.06 (0.0~90.0)
Number of livestock workers	12.9610	2.16 (1.0~5.0)	0.1 (0.0~3.0)	1.34 (0.0~5.0)	0.31 (0.0~3.0)
Animal Stock (in LSU)	11.5770	10.16 (0.0~92.22)	0.04 (0.0~8.4)	0.43 (0.0~13.12)	0.16 (0.0~32.5)
Cattle Rearing	7.2085	1 (89.44%)	0 (99.23%)	0 (97.0%)	0 (97.15%)
Farming multiple livestock	7.1352	1 (82.34%)	0 (98.74%)	1 (74.61%)	0 (95.8%)
Livestock Income	4.8894	465460.57 (0.0~7170000.0)	6275.74 (0.0~2325000.0)	164318.16 (0.0~43635000.0)	14107.28 (0.0~4480000.0)
Household size	3.8184	8.43 (1.0~29.0)	3.97 (1.0~17.0)	5.34 (1.0~17.0)	4.82 (1.0~18.0)
Education Level	3.3798	18 (56.24%)	18 (27.82%)	18 (49.23%)	18 (44.81%)
Media Use	3.2661	2.78 (0.0~18.0)	3.4 (0.0~24.0)	2.8 (0.0~14.0)	2.21 (0.0~17.0)

3. Result – Bar Plots of SE/ECO features with highest FI

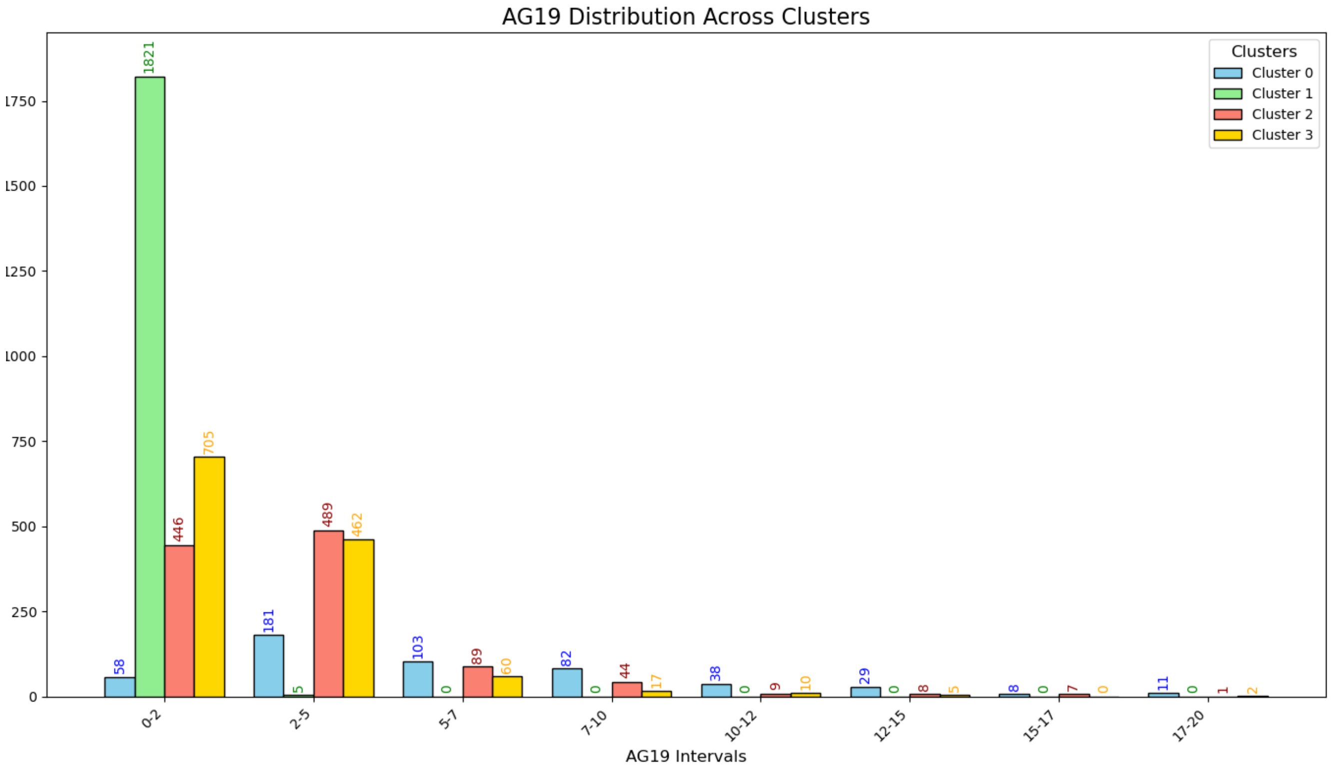
Land Tenure Type (SE12)



Description	Importance	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Land tenure condition	22.2430	2 (51.63%)	0 (96.28%)	2 (43.22%)	2 (45.68%)

While little data exists for cluster 1, the proportion of farms with Customary tenures ('1') is very high. While clusters 2 and 3 have similar levels of freehold farms, Cluster 0 has the largest proportion of its households on freehold land.

Number of agricultural tools (SE19)



Feature Description	Importance	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Number of agricultural tools	16.4734	7.03 (0.0~42.0)	0.02 (0.0~4.0)	10.7 (0.0~5000.0)	2.87 (0.0~24.0)

Cluster 1's low agricultural tools are emphasized again, implying lower productivity and technology adaption. While all clusters also contain smaller farms, the right skew of clusters 0, 2, and 3 is notable.

4. Discussion – Key Socioeconomic and Ecological Factors

Interaction Between Socioeconomic and Ecological Factors: For example, households with higher livestock are more likely to adopt environmentally friendly agricultural practices (e.g. Clusters 0 and 2), thereby improving soil and water management and achieving sustainable income growth.

Distinguishing farms through their clusters: Using the results from the Random Forest Classification, clusters can be distinguished in terms of their *farm size, cattle stock, household size,* and some *sustainable practices*.

Synergies between variables: The plot highlights the interplay between resource availability—such as agricultural tools (AG18/19) and livestock ownership (LF14)—and land tenure conditions (SE12). Clusters with better access to agricultural tools (AG18 = 1) and sustainable livestock practices (LF14 = 1) are associated with more secure land tenure arrangements, such as "freehold" (SE12 = 2) or "leasehold" (SE12 = 3). This also indicates the underlying need for structural policies that improve land ownership security.

"In particular, Cluster 1 underscores the challenges faced by resource-constrained households, emphasizing the need for targeted interventions to enhance their access to agricultural tools and livestock-based farming practices."

Discussion of Data Improvements: Further development of more specific and comprehensive sustainability assessment indicators is necessary to guide individuals towards achieving ecological sustainability and climate resilience. Relationships between the structural features of household farms and sustainability features, including carbon emissions, remains unexplored. Challenges arising from limited insights through binary data could be mitigated through 'numericalization' of data (e.g. LF10.1 and LF10.2).

4. Discussion – Role of Machine Learning in Policy Design

Precise Policy Targeting

• Allows for targeted interventions tailored to group-specific needs (using *k-means*).

• Random Forest models reveal key factors affecting income and sustainability for policy focus.

•Very large datasets (with new waves being released in regular time intervals) can be analyzed efficiently

•Can be used to identify resource needs (e.g., training, funding) for targeted interventions

•ML tools (like *skicit*-learn) are widely available and intuitive to use

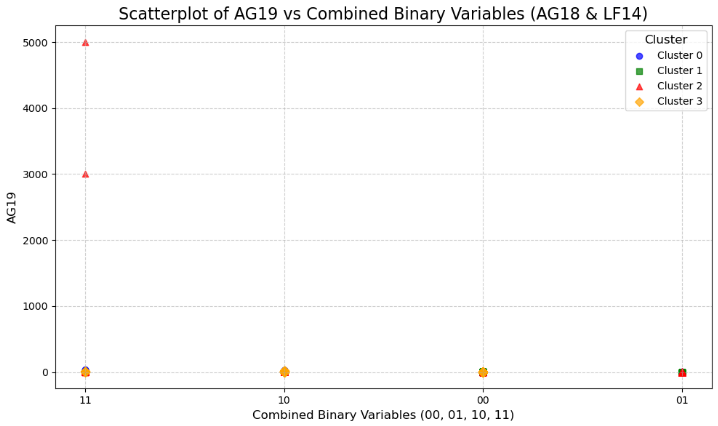
Challenges in Implementation

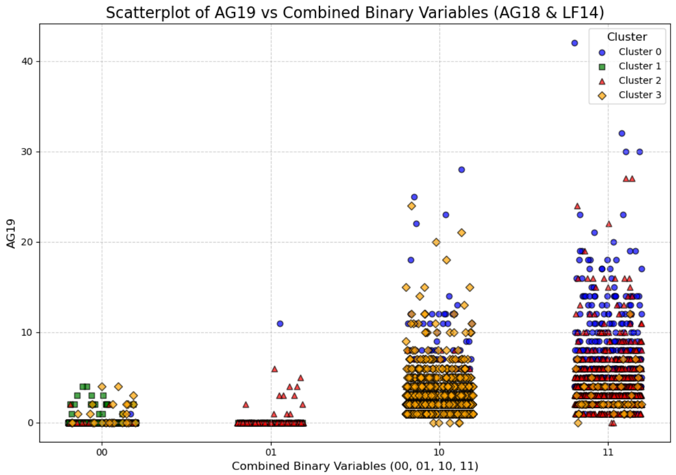
• **Interpretability**: Difficulty to Interpret Random Forest Results, lacking automation, provides slightly different values every time that it is used. If not used for classification, it is less intuitive to use. Needs to be translated back into natural language.

• **Visualization:** Dealing with combinations of binary, categorical and numerical variables was challenging.

•**High-dimensional analysis:** Less intuitive visualizations when 2D space only represents a low amount of aggregate Feature Importance.

Visualization Challenges





5. Project Outlook – Timeline & Future Applications

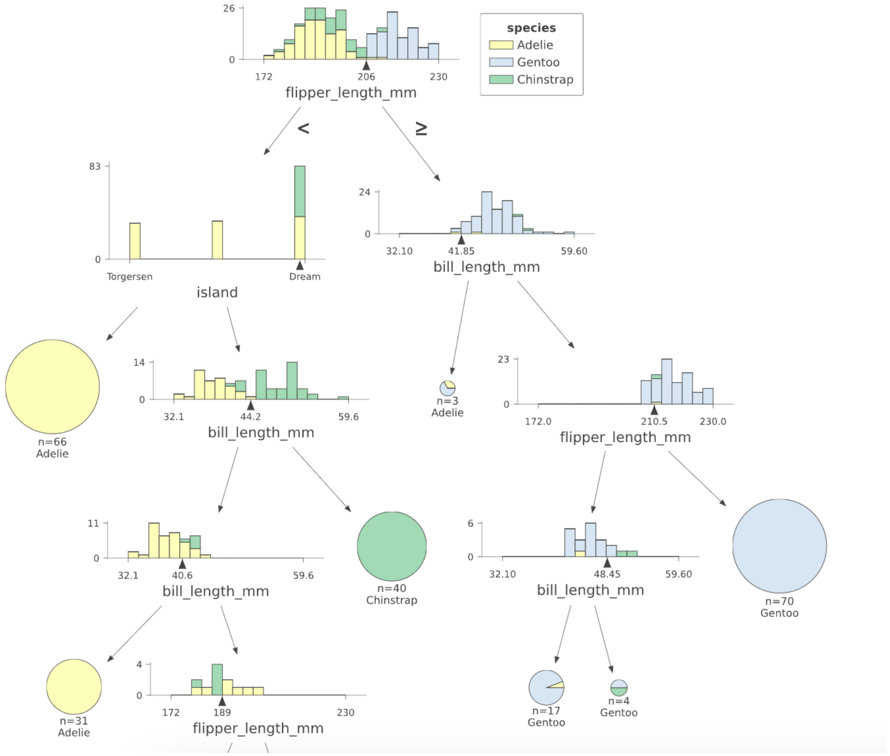
Next Steps

1. Decision Tree Reconstruction

- Creating a comprehensible decision tree structure to easily classify new data points.
- *Example:* A decision tree with sufficient structure with sufficient aggregate feature importance can help quickly assign new farms to the existing clusters.

2. Random Forest Automation

- Automating the K-means Clustering and Random Forest Classification
- *Example*: Extending this case study to the next NPS-Wave.



Source: *Tensor Flow (2024)*

[January 2025]
Finish the Paper & Reconstruct Decision Tree

[Jan 31st]
Review by Prof. Ma

[February 2025]
Submission Goal.

Summary

- **Land tenure condition** and **size**, as well as the **number of livestock workers** are the the most important socioeconomic features in cluster assignment.
- **Agricultural tool ownership** as well as the sustainable practice of livestock plowing are the most important ecological features in cluster assignment.
- **Livestock Ownership** and the **sustainable practices** associated with it are associated with safer land ownership, higher income, and higher levels of diversification.
- Cluster 0 and Cluster 1 highlight the difference between wealthy households embracing sustainable practices and resource-constrained, agricultural households.
- Numerical, quality data on specific sustainable practices and associated GHG-emissions would extend our analysis *using the same methodology.*

Thank you for your attention! For further inquiries, please contact:

Email: beck1017@hotmail.com and fabian@uni.minerva.edu

Transformational Pathways of Household Farms in Tanzania Based on Machine Learning Analysis: Key Factors from income structure to sustainable practices

2024.12.20