UNIVERSITÀ DEGLI STUDI DI MILANO

EXPLAINABLE CLUSTERING ASSIGNMENT 1

Course: Scientific Visualization

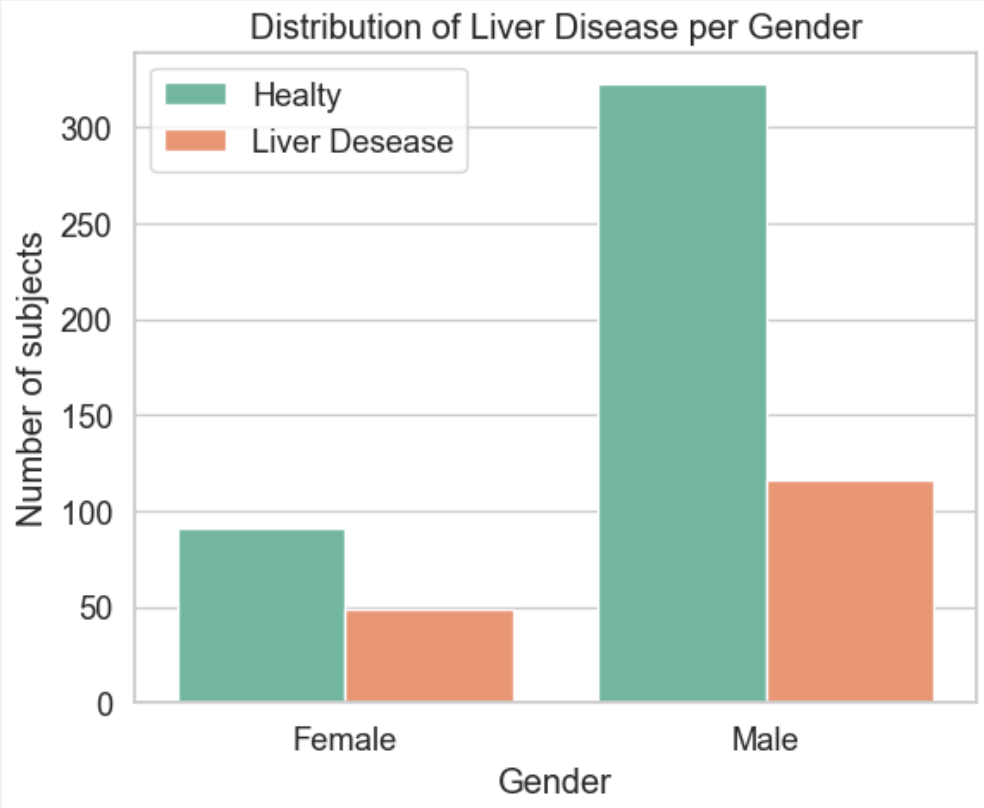
Corti Filippo
Dal Santo Giorgio
Donato Carlotta

Code and Visualizations available at
https://github.com/Filippo-Corti/ExplainableClustering

Dataset Overview

Dataset: ILPD (Indian Liver Patient Dataset)

- **579** patient samples
- **Features**:
  - *Continuous*: Age, Total Bilirubin, Direct Bilirubin, Alkaline Phosphotase, Sgpt, Sgot, Total Protein, Albumin, A/G ratio.
  - *Binary*: Gender (439 are Males and 140 are Females).
- **Label**: 1 = Liver disease, 2 = Healthy.
- **Goal**: identify and visualize the main features defining clusters in the ILPD dataset.



Finding the best Clustering Algorithm

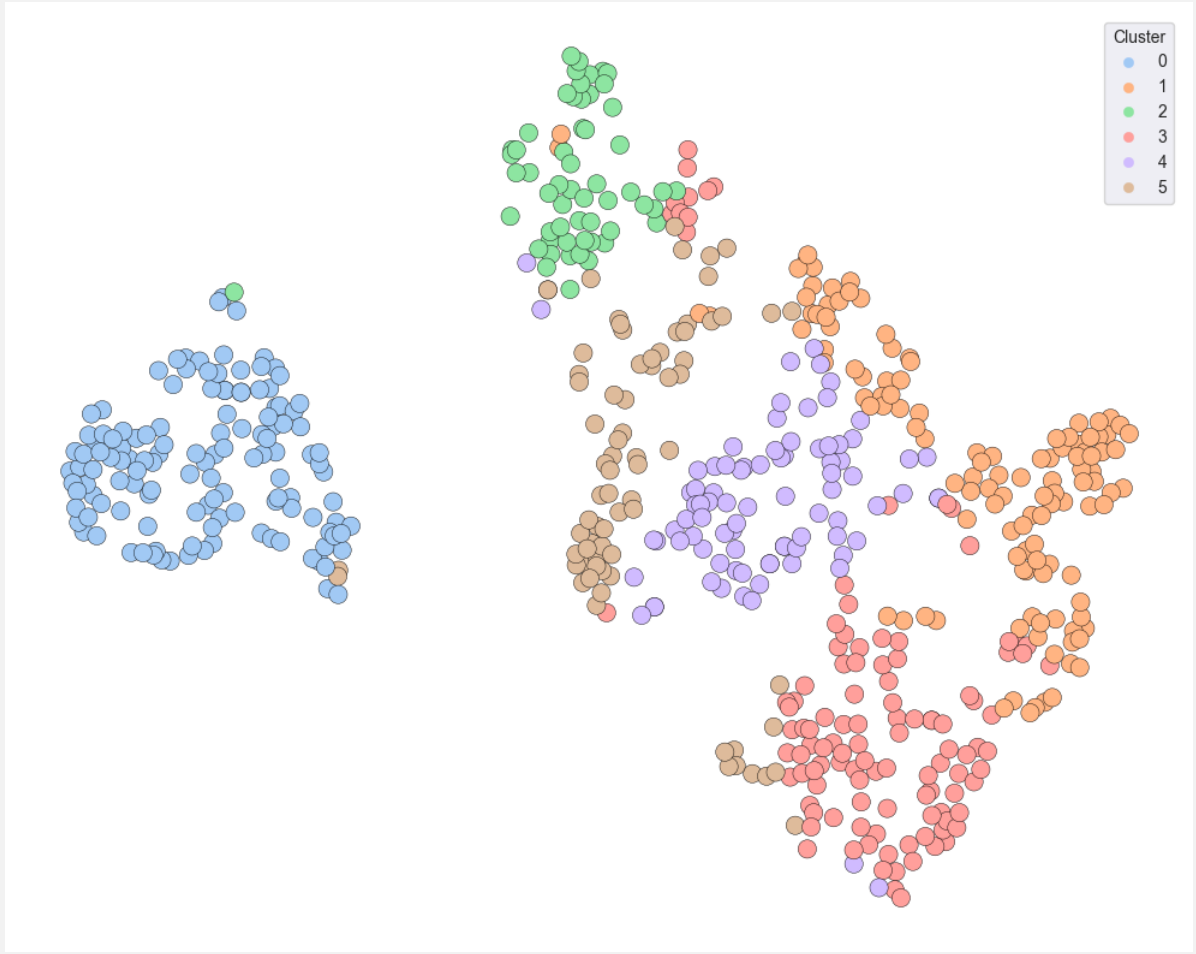
Method	Mixed Types	K Choice	Evaluation Metric	Cluster Sizes						
				6	5	4	3	2	1	0
K-Proto [K=4]	Yes	Elbow Method	Hamming distance (_cost)				2	15	43	519
K-Proto [K=5]	Yes	Elbow Method	Hamming distance (_cost)			2	9	33	41	494
GMM [K=4]	No	Elbow Method	Silhouette score				5	6	175	393
HDBSCAN [K=4]	No	Automatic	Silhouette score				5	8	132	434
Spectral [K=4]	Yes	Elbow Method	Silhouette score				75	132	168	204
Spectral [K=6]	Yes	Elbow Method	Silhouette score		56	75	85	112	121	130

Spectral Clustering with Gower Distance

1. Compute **Gower's Distance** between each Data Point.
2. Build a **Fully Connected Graph** having:
    - Data Points as Nodes.
    - Gower's Distance as Weights.
3. Compute the **Graph Laplacian Matrix** (L) as L = D – A.
4. Compute the **Eigenvalues** of L and take the **Eigenvectors** for the first n Eigenvalues.
5. Compose a **new Matrix** using the Eigenvectors.
6. Apply a classic **Clustering Algorithm** on the Points of the new Matrix.

Spectral Clustering Visualization

The **Elbow Rule** can help to visually find the point where increasing the Number of Clusters doesn't dramatically decrease the Silhouette Score anymore.

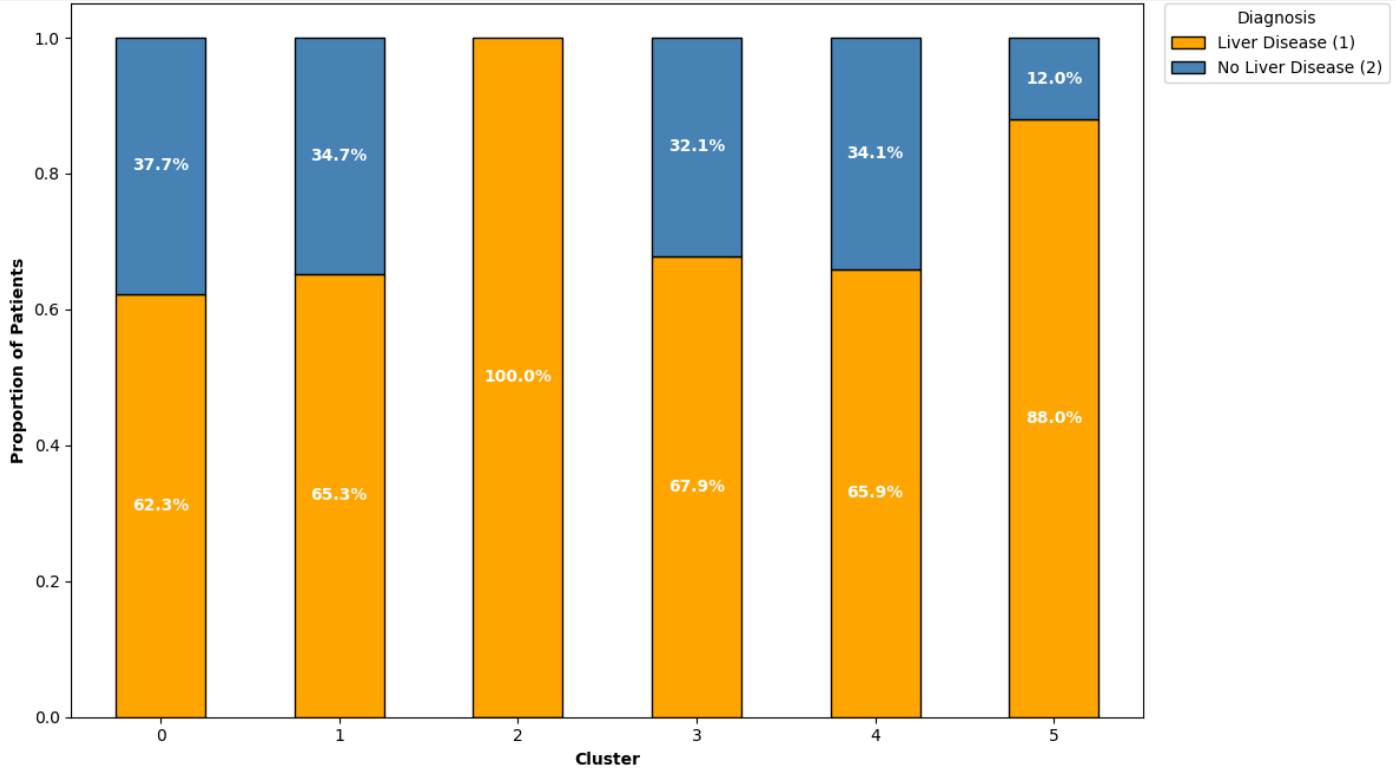


In order to **visualize** the Clusters, we applied 2 transformations:

1. From **10D** to **6D** with **Spectral** Embedding.
2. From **6D** to **2D** with **t-SNE**.

Cluster Validation: Goodness-of-Fit Test

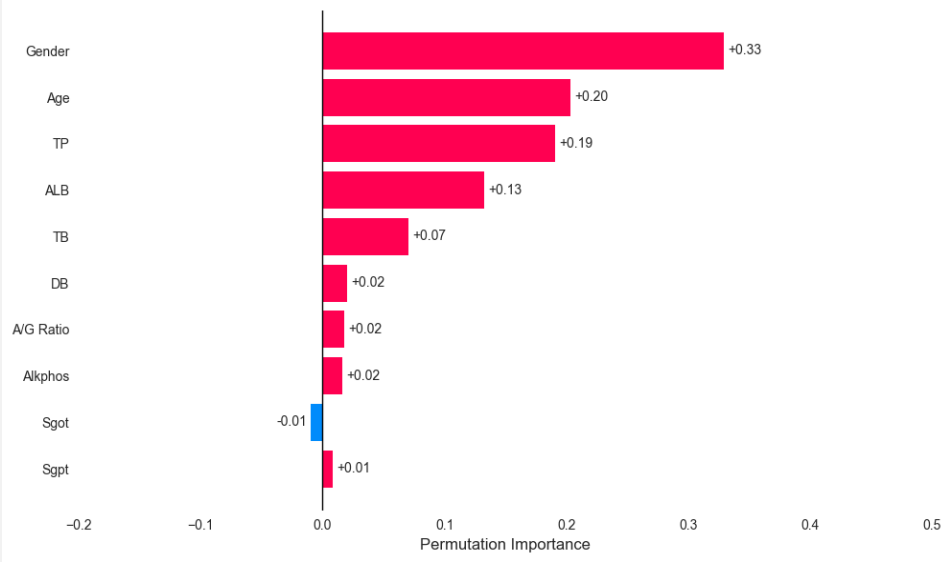
Question: How does each cluster's disease distribution compare to the overall population?



Cluster	p-value	Significant?
0	0.0202	Yes
1	0.1300	No
2	<0.0001	Yes
3	0.3927	No
4	0.2510	No
5	0.0016	Yes

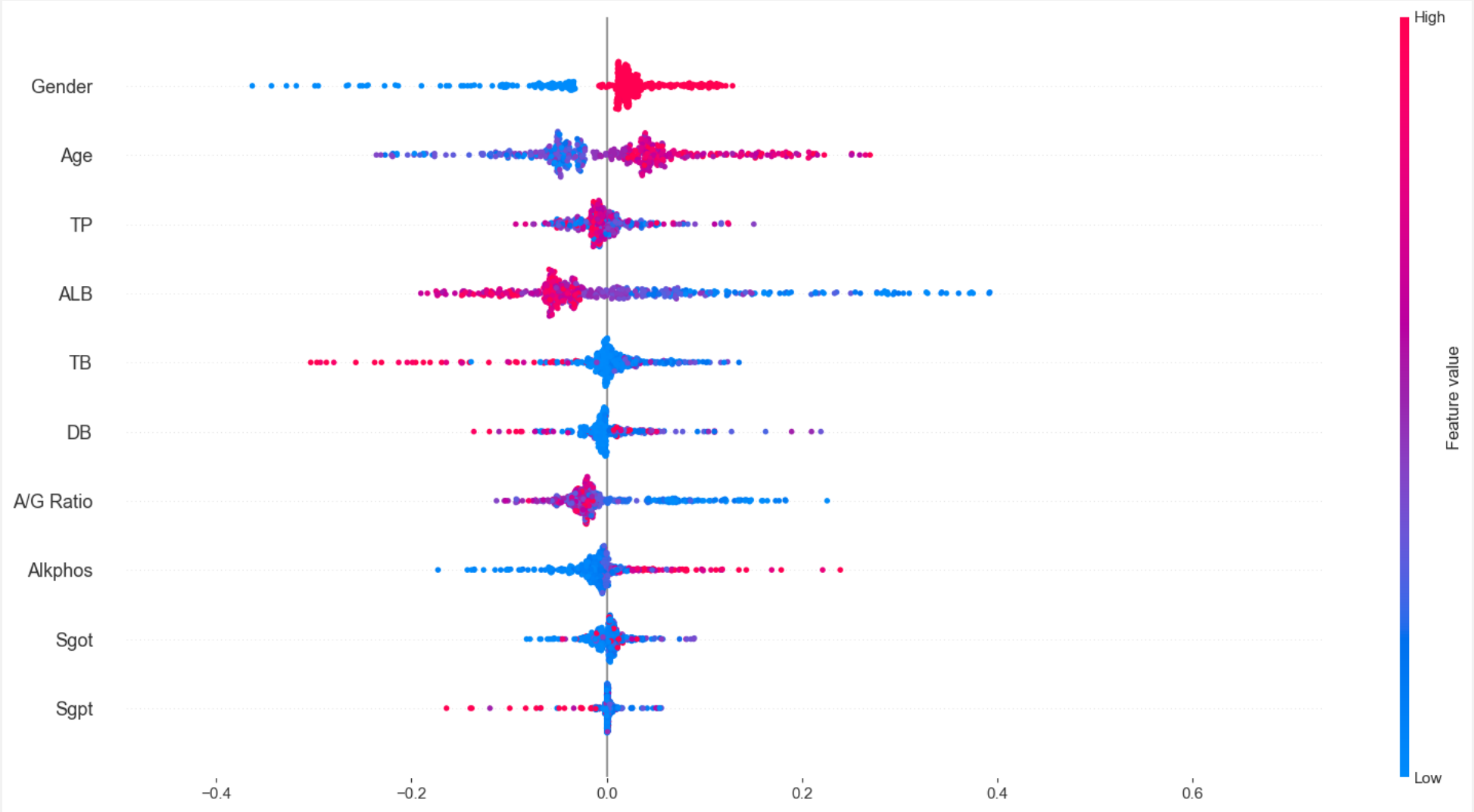
Clusters Analysis via Explainable ML

- After training a **Random Forest** to predict our Clusters Labels, we can use the learnt values to determine **Feature Importance**.
  - Accuracy: 0.978 (Train), 0.940 (Test)
  - Precision: 0.979 (Train), 0.940 (Test)
  - Recall: 0.978 (Train), 0.940 (Test)
  - F1: 0.978 (Train), 0.940 (Test)

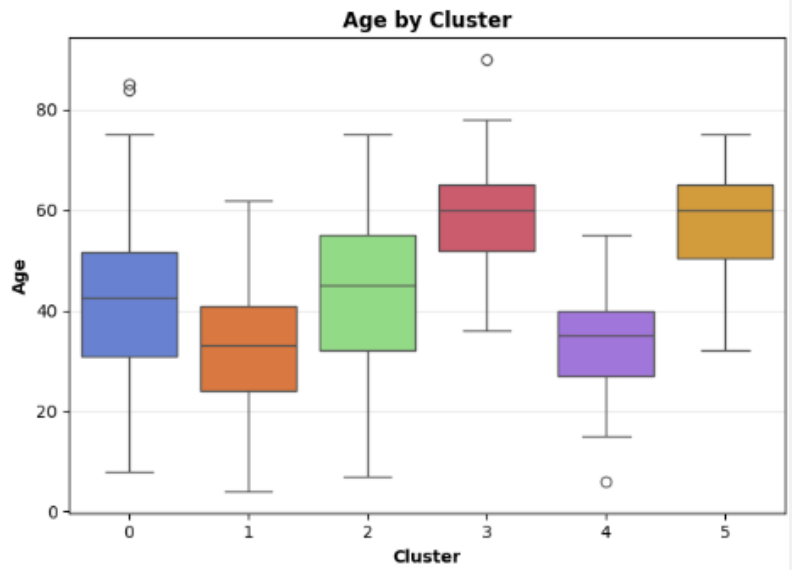


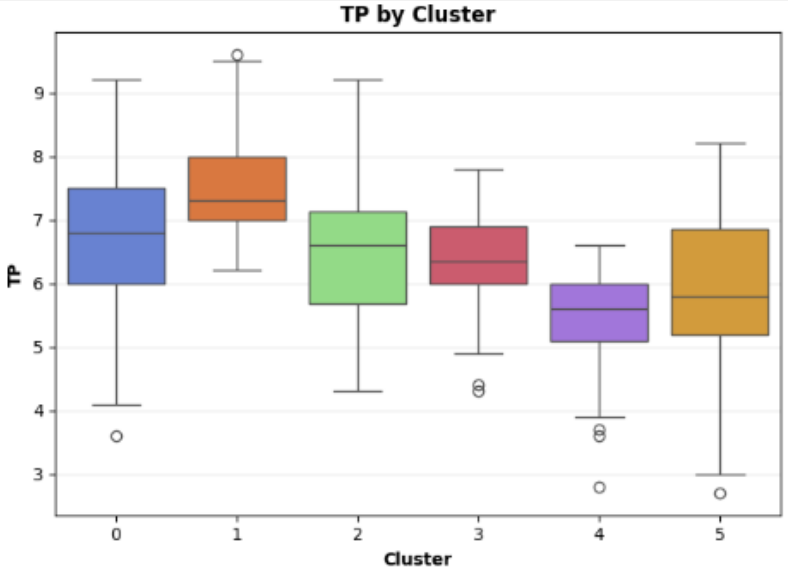
- Additionally, **SHAP values** can be used to show how each feature contributed to predicting a specific cluster label.
  - They do so by computing the impact of having a certain value for the feature, compared to its baseline.

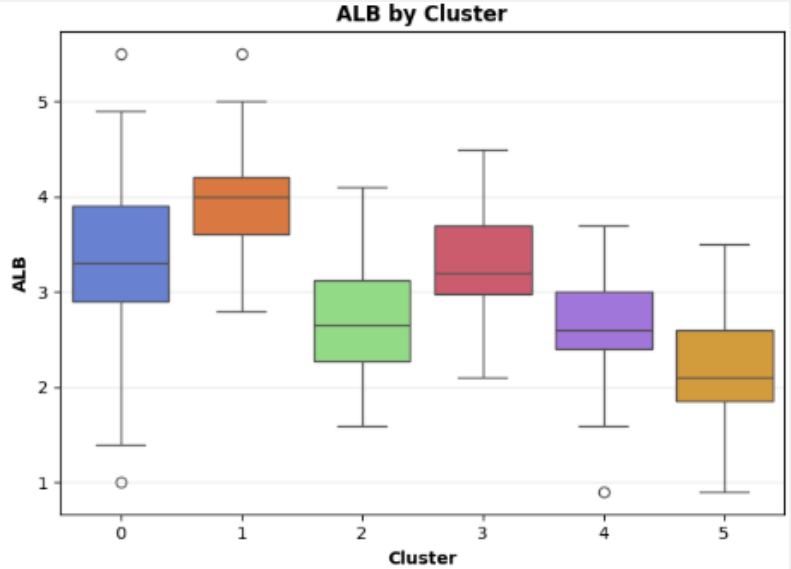
Beeswarm Plots of SHAP Values (Cluster 5)

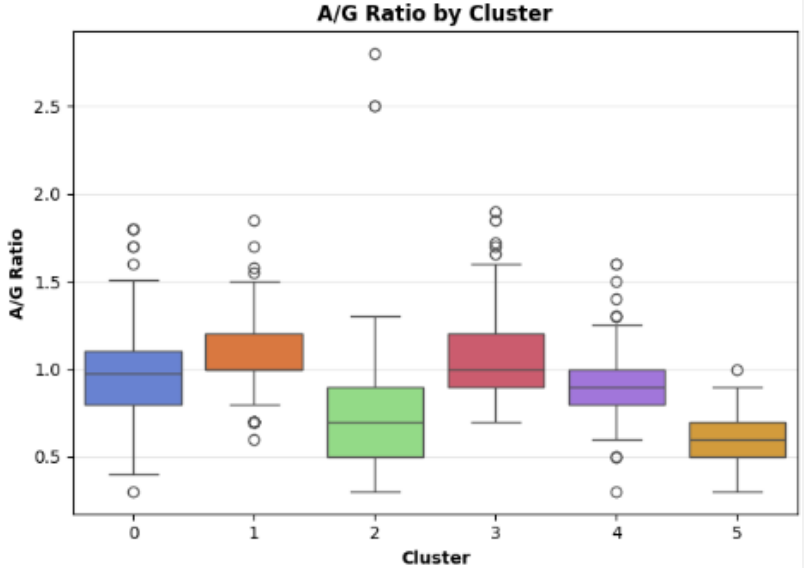


Boxplot – Feature Distribution Across Clusters

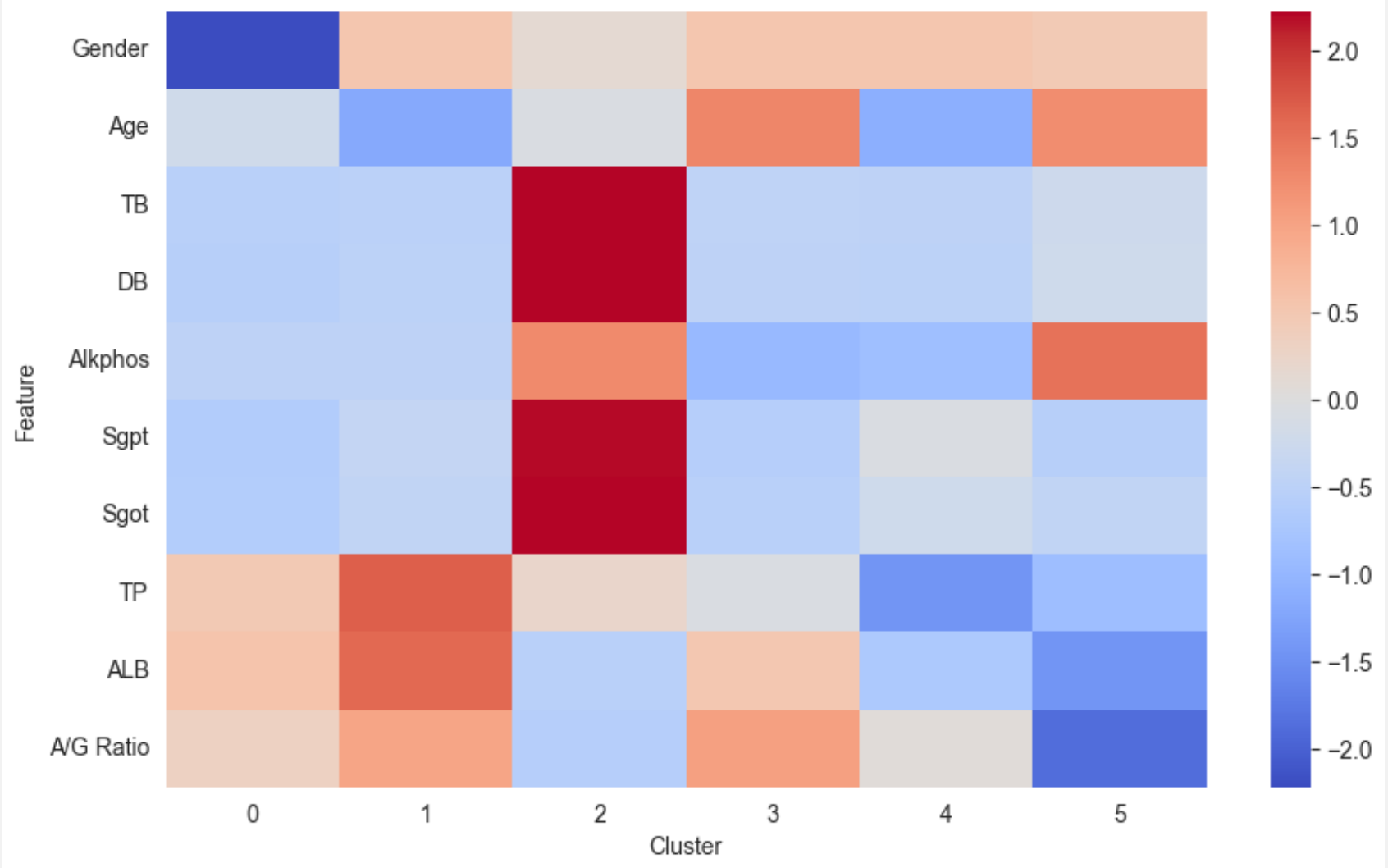




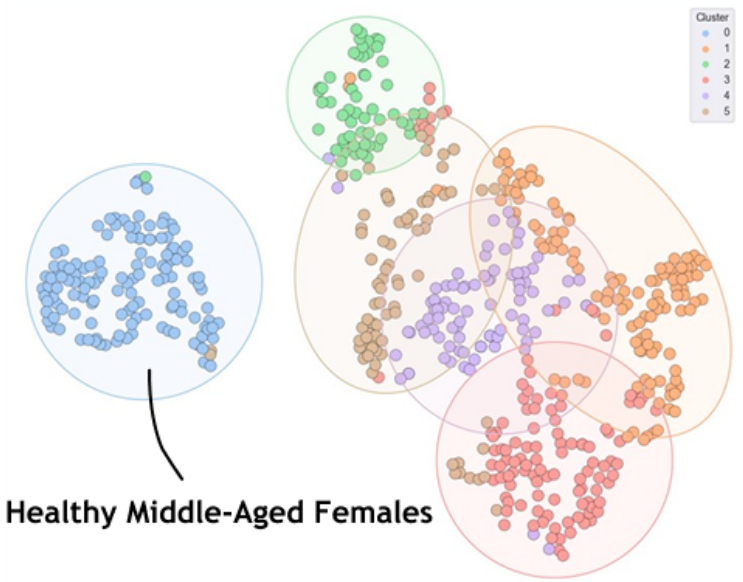


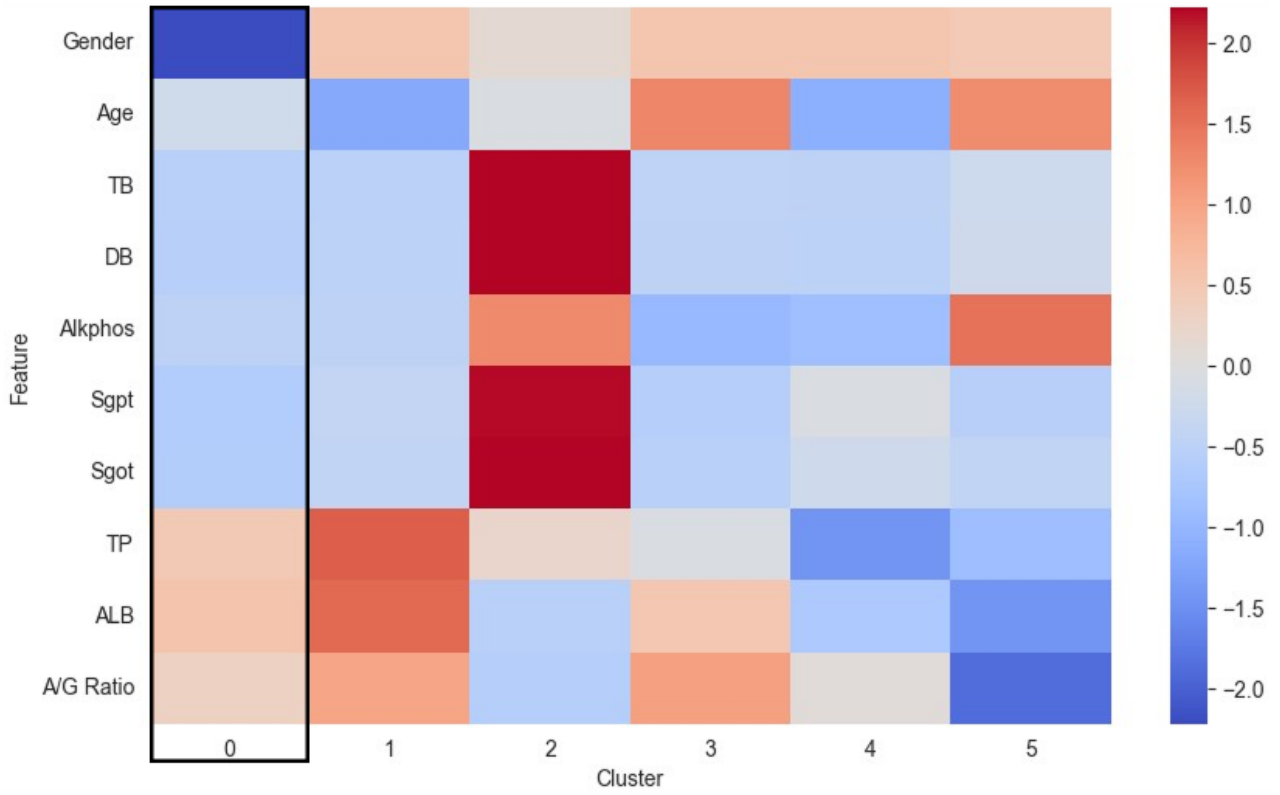


Heatmap - Cluster Means (Standardized)



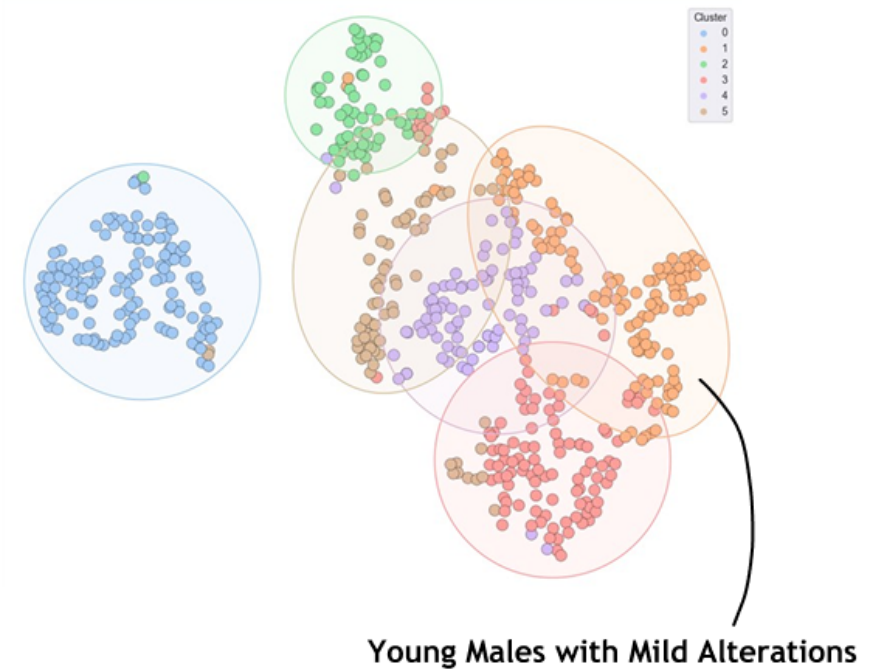
Final Clustering Results – 0

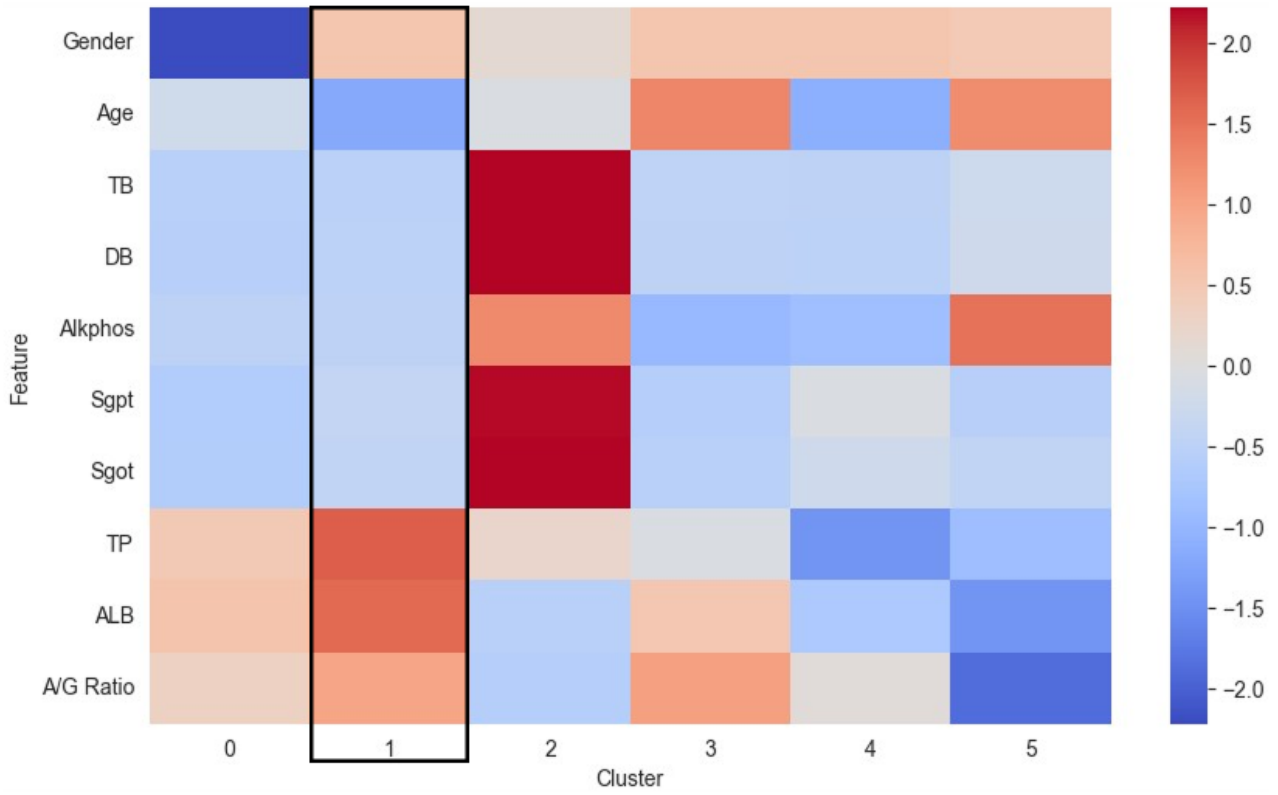




Size	Gender (%)		Age	Enzymes	Proteins	Liver Disease (%)	Key Traits
	F	M					
130	100	0	43.1 ± 16.0	Mostly normal, Alkphos slightly elevated	Normal	62	Middle-aged females, healthy liver

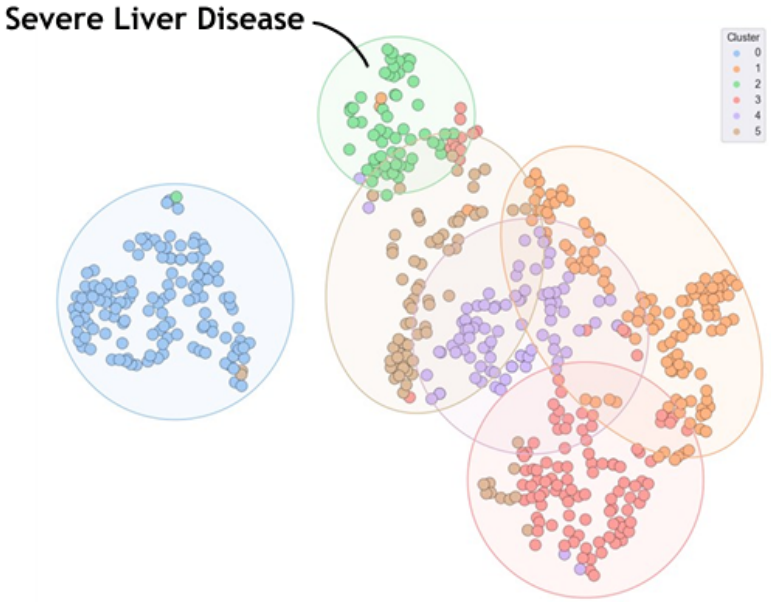
Final Clustering Results – 1

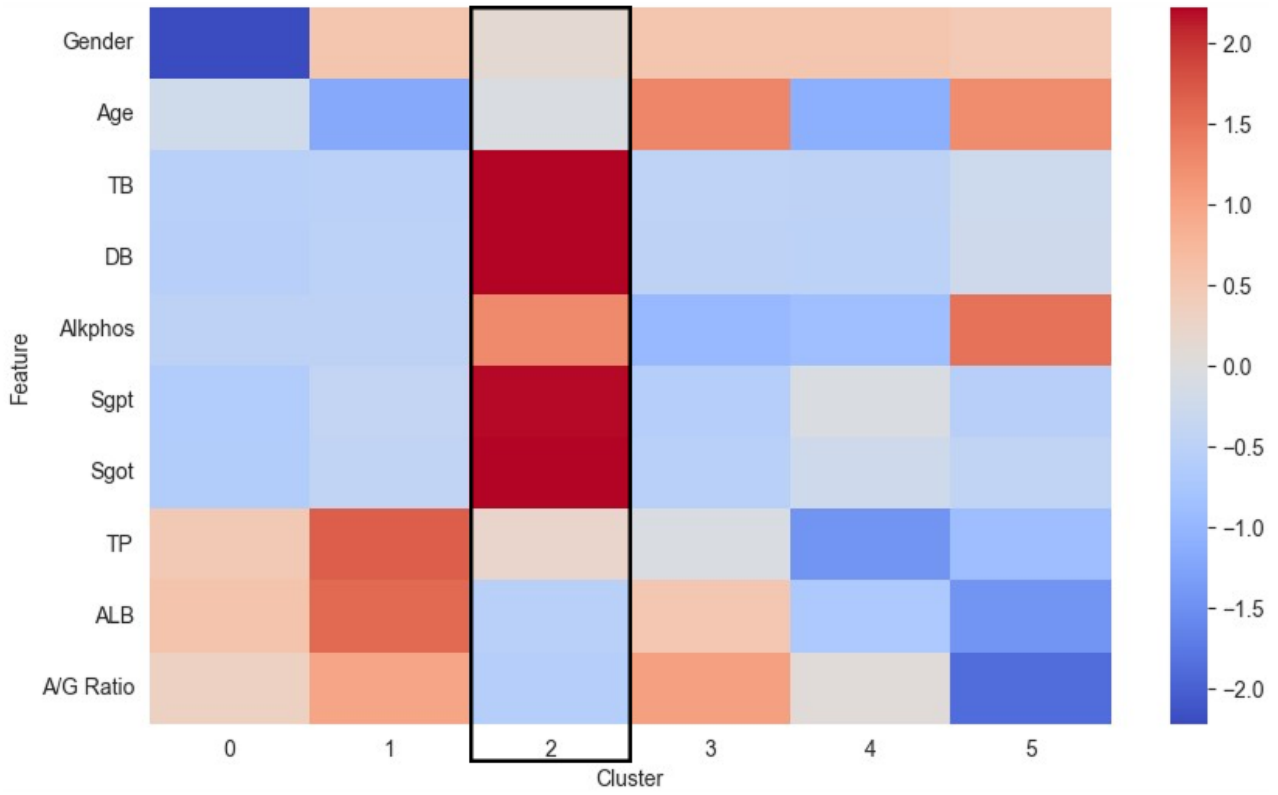




Size	Gender (%)		Age	Enzymes	Proteins	Liver Disease (%)	Key Traits
	F	M					
121	0	100	32.7 ± 12.2	DB, Sgpt/Sgot elevated	Higher TP and ALB	65	Young males, mild liver alterations

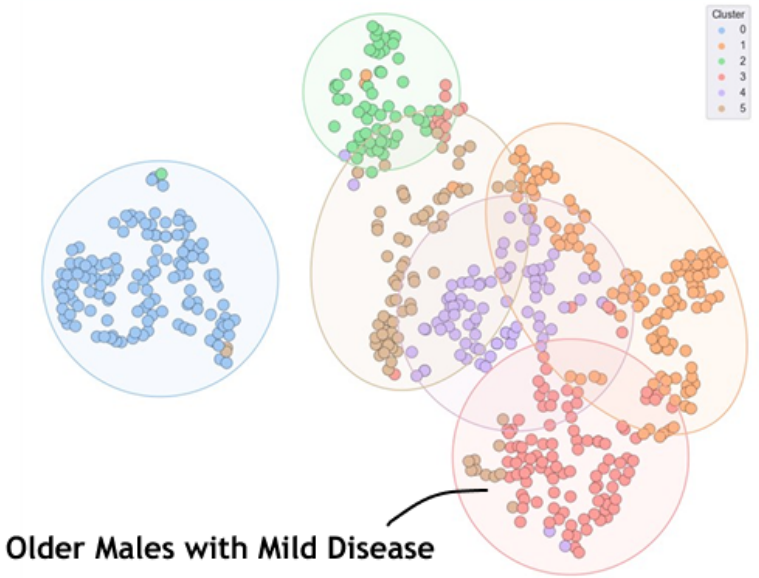
Final Clustering Results – 2

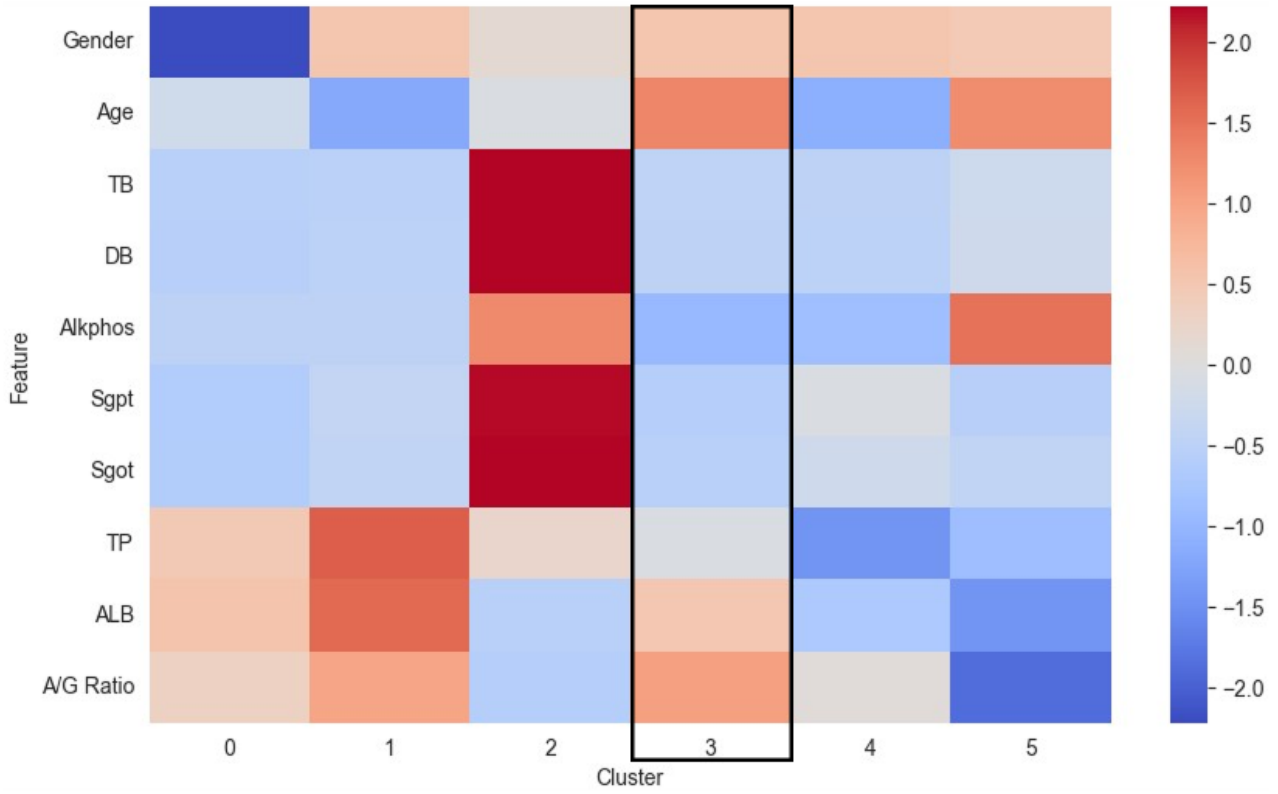




Size	Gender (%)		Age	Enzymes	Proteins	Liver Disease (%)	Key Traits
	F	M					
56	14	86	44.6 ± 13.8	Severely elevated enzymes	Low ALB	100	High liver damage

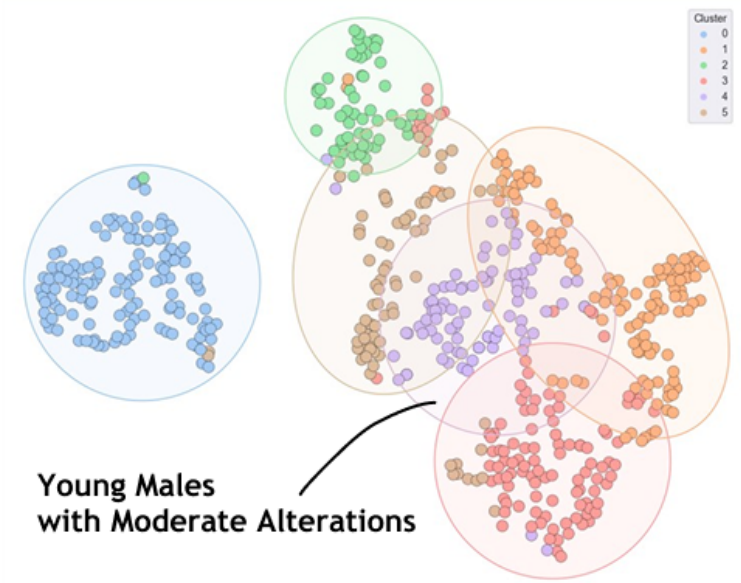
Final Clustering Results – 3

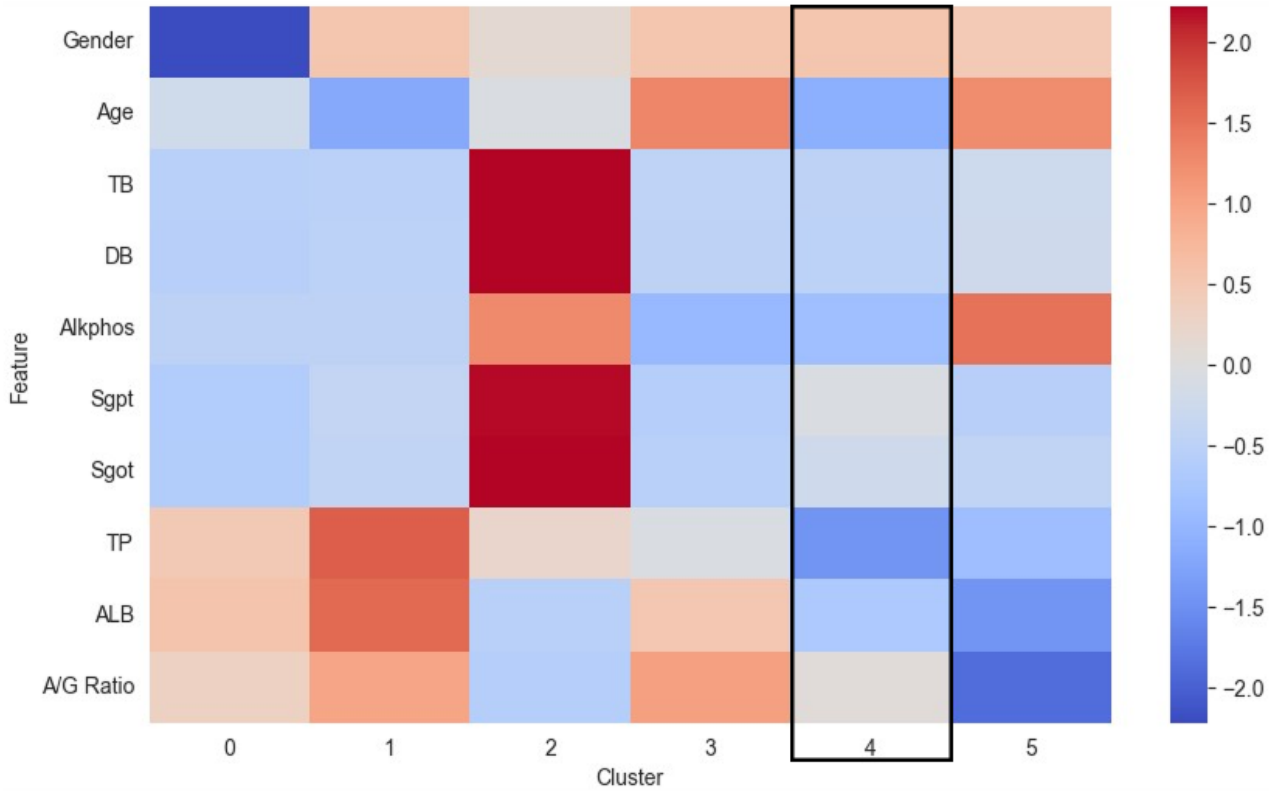




Size	Gender (%)		Age	Enzymes	Proteins	Liver Disease (%)	Key Traits
	F	M					
112	0	100	59.3 ± 09.3	Mild/moderate elevation	Normal	68	Older group, mild disease

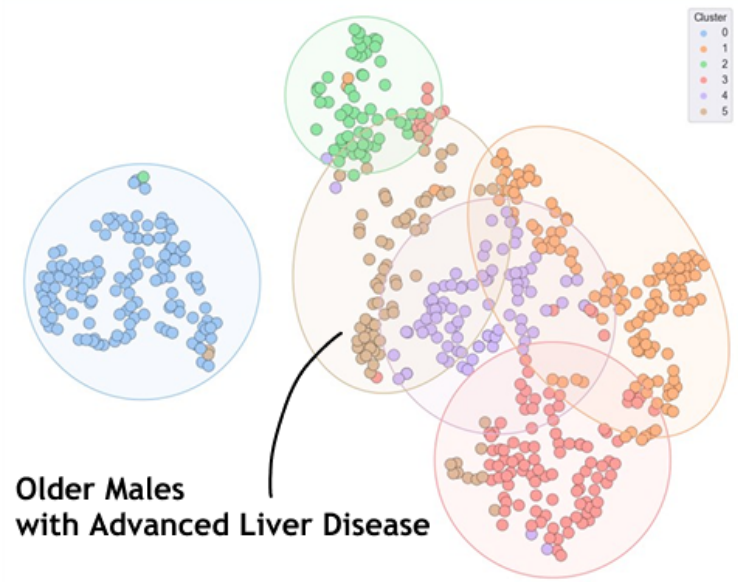
Final Clustering Results – 4

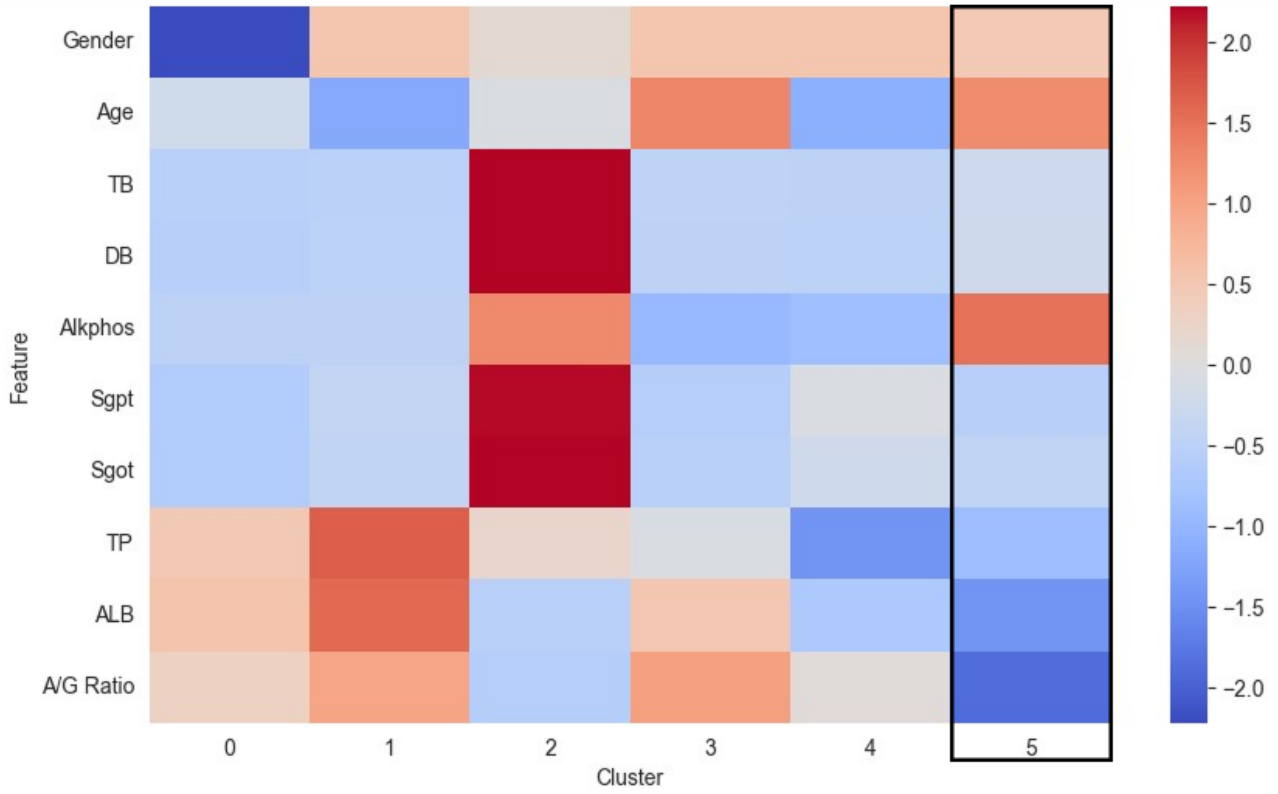




Size	Gender (%)		Age	Enzymes	Proteins	Liver Disease (%)	Key Traits
	F	M					
85	0	100	33.4 ± 09.4	Moderate enzyme elevation	Low-normal	66	Young, moderate enzyme elevations, low-normal protein levels

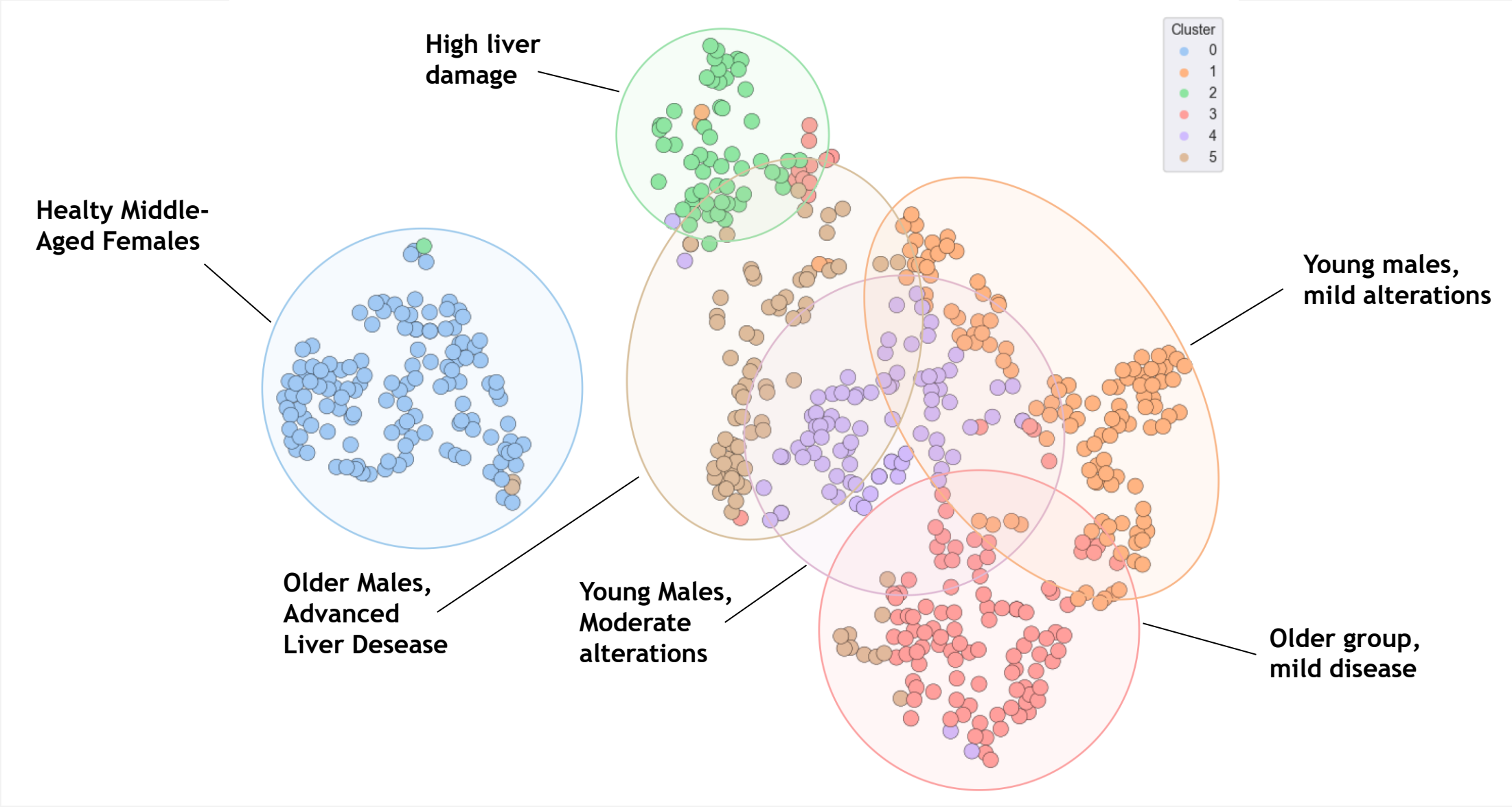
Final Clustering Results – 5



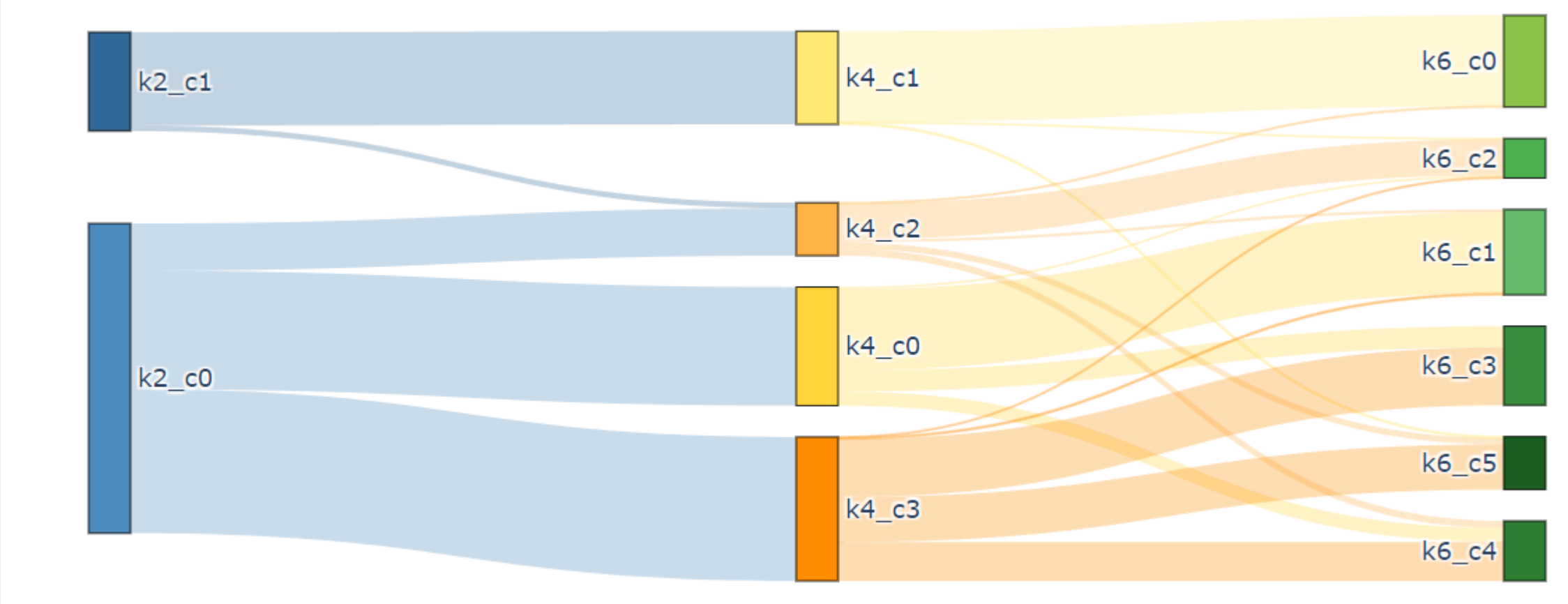


Size	Gender (%)		Age	Enzymes	Proteins	Liver Disease (%)	Key Traits
	F	M					
75	3	97	58.6 ± 09.8	High TB/DB, Alkphos	Very low ALB	88	Older, advanced disease

Final Clustering Results



Evolution of Spectral Clusters (k=2 → k=4 → k=6)



UNIVERSITÀ DEGLI STUDI DI MILANO

Thank you for your attention.

Clinical Reference for Standard Ranges:
https://www.ncbi.nlm.nih.gov/books/NBK482489/

Dataset: https://archive.ics.uci.edu/dataset/225/ilpd+indian+liver+patient+dataset