Identifying leaves using Cluster Analysis

1. Introduction

The primary objective of this analysis is to segment a collection of unlabeled leaf specimens into homogeneous groups based on shape and texture measurements using the k-means clustering algorithm. Specifically, **how many distinct clusters best describe variation in leaf shape and texture, and what biological insights can be drawn from these groupings?**

This dataset was sourced from the UC Irvine Machine Learning Repository. It was originally created by Pedro Silva and Andr Maral in the context of a mathematical engineering master's program to establish an automated system for identifying plants through the "automatic extraction of morphometric information from images of plant leaves" in place of taxonomers (2013).

Problem Domain

In the past, the identification of plants was completed by specialized technicians called taxonomers. Clustering leaf-image features supports biodiversity monitoring by automating vegetation surveys over large spatial scales. For example, remotely sensed or in-situ leaf scans can be clustered to detect changes in community composition, track invasive species incursions, or assess responses to environmental stressors (Eklundh & Olsson, 2003).

Method Rationale

k-Means cluster analysis was chosen for its simplicity, scalability, and interpretability. It partitions observations into k groups by minimizing within-cluster variance (Lloyd, 1982; UMGC, 2024). While hierarchical and density-based methods can capture non-spherical clusters, our exploratory analysis indicates roughly globular feature distributions, making k-Means an appropriate choice. Our exploratory data analysis – especially the pairwise scatterplots and PCA projection – showed that leaf specimens form pretty compact groups along the main shape and texture axes. K-Means also allows straightforward parametric tuning via the elbow method (aka inertia) and silhouette analysis (James et al., 2013)

In our k-means implementation, we first determined the optimal number of clusters *k* by examining the inertia ("elbow") curve and silhouette scores for *k* values between one and ten, ultimately selecting k=4 as the most parsimonious choice that preserved cluster cohesion. Second, we employed the "k-means++" initialization scheme to place initial centroids in a manner that accelerates convergence and typically produces more stable solutions than purely random seeding (Arthur & Vassilvitskii, 2007). Finally, we allowed up to 300 iterations (max_iter=300) to ensure sufficient centroid refinement without excessive computational expense.

2. Exploratory Analysis

2.1 Data Description and Summary Statistics

The publicly available 'leaf' data set is comprised of 340 observations and 14 numeric features extracted from scanned leaf images (ReadMe, 2025). There are no missing values in the dataset.

After dropping the first two identifier columns (Class and Specimen Number), the dataset's remaining features include shape and texture descriptions based on statistical moments and analyses, with more complete descriptions shared in the master's thesis. (Silva, 2013; see Appendix A for more complete descriptions)

Shape Descriptions –

- Eccentricity,
- Aspect Ratio,
- Elongation,
- Solidity,
- Stochastic Convexity,
- Isoperimetrtic Factor,
- Maximal Indentation Depth,
- Lobedness,

and **Texture Descriptions –**

- Average Intensity,
- Average Contrast,
- Smoothness,
- Third moment,
- Uniformity,
- Entropy.

Summary Statistics

Summary statistics of the raw (non-standardized) dataset indicate moderate variability across features,

	count	mean	std	min	25%	50%	75%	max
Eccentricity	340.0	0.719854	0.208311	0.117080	0.550623	0.763450	0.895097	0.998710
Aspect Ratio	340.0	2.440210	2.599043	1.006600	1.211300	1.570750	2.343100	19.038000
Elongation	340.0	0.513760	0.195583	0.107610	0.349623	0.501855	0.633373	0.948340
Solidity	340.0	0.904158	0.114639	0.485490	0.890667	0.948130	0.976897	0.993880
Stochastic Convexity	340.0	0.943793	0.115047	0.396490	0.966230	0.992980	1.000000	1.000000
Isoperimetric Factor	340.0	0.531234	0.217532	0.078376	0.346818	0.579160	0.700713	0.858160
Max Indentation Depth	340.0	0.037345	0.038575	0.002837	0.009521	0.023860	0.047834	0.198980
Lobedness	340.0	0.523845	1.039639	0.001464	0.016500	0.103615	0.416432	7.206200
Average Intensity	340.0	0.051346	0.035965	0.005022	0.022843	0.042087	0.073046	0.190670
Average Contrast	340.0	0.124535	0.051860	0.033415	0.083362	0.119375	0.163795	0.280810
Smoothness	340.0	0.017670	0.013755	0.001115	0.006901	0.014050	0.026127	0.073089
Third Moment	340.0	0.005928	0.005294	0.000229	0.002080	0.004447	0.008307	0.029786
Uniformity	340.0	0.000387	0.000431	0.000007	0.000102	0.000239	0.000516	0.002936
Entropy	340.0	1.162630	0.584854	0.169400	0.718900	1.077450	1.554575	2.708500

Table 1 *Summary statistics (count, mean, std, min, 25%, median, 75%, max) for the 14 quantitative shape and texture features in the Leaf dataset.*

3. Preprocessing

3.1 Data Cleaning, Transformation

The raw dataset had no missing or duplicate values, and outliers were retained by design; no rows were dropped prior to analysis.

However, "because most clustering algorithms (including k-means) are based on distance measures, variables with larger numeric ranges will dominate the distance calculations and hence the resulting cluster assignments. It is therefore essential to standardize all features—typically to zero mean and unit variance—so that each variable contributes equally to the Euclidean distances used by the algorithm." (James et al., 2021, p. 354)

To support more accurate clustering, the Leaf dataset was transformed using standardization techniques.

	count	mean	std	min	25%	50%	75%	max
Eccentricity	340.0	-3.657205e-17	1.001474	-2.897890	-0.813596	0.209591	0.842498	1.340624
Aspect Ratio	340.0	0.000000e+00	1.001474	-0.552405	-0.473529	-0.335024	-0.037419	6.395529
Elongation	340.0	-5.224579e-18	1.001474	-2.079675	-0.840461	-0.060961	0.612468	2.225245
Solidity	340.0	-4.597629e-16	1.001474	-3.657431	-0.117849	0.384136	0.635446	0.783803
Stochastic Convexity	340.0	8.568309e-16	1.001474	-4.764242	0.195313	0.428170	0.489279	0.489279
Isoperimetric Factor	340.0	-4.911104e-16	1.001474	-2.084870	-0.849016	0.220644	0.780249	1.505107
Max Indentation Depth	340.0	7.314411e-17	1.001474	-0.895892	-0.722357	-0.350084	0.272311	4.196344
Lobedness	340.0	-1.044916e-17	1.001474	-0.503204	-0.488720	-0.404803	-0.103469	6.437043
Average Intensity	340.0	-9.828739e-17	1.001474	-1.289944	-0.793693	-0.257821	0.604238	3.879586
Average Contrast	340.0	-8.359326e-17	1.001474	-1.759637	-0.795103	-0.099643	0.758163	3.017868
Smoothness	340.0	1.018793e-16	1.001474	-1.205309	-0.784032	-0.263563	0.615773	4.034934
Third Moment	340.0	8.881784e-17	1.001474	-1.077931	-0.727942	-0.280150	0.450068	4.513203
Uniformity	340.0	-6.269495e-17	1.001474	-0.882763	-0.661391	-0.344780	0.299334	5.915846
Entropy	340.0	1.671865e-16	1.001474	-1.700756	-0.759821	-0.145858	0.671147	2.647069

Table 2 *Summary statistics (count, mean, std, min, 25%, median, 75%, max) for the 14 quantitative shape and texture features in the standardized Leaf dataset.*

After standardizing, features don't show wide range of variance. Instead, the interquartile range and dimensionality-reduction (PCA) techniques were completed to inform relative spread and which features contribute most to the top components. Based on the results of the IQR and PCA analysis, the following key features were identified for further exploration:

- Max Indentation Depth (PCA 0.598, IQR ≈ 1.168)
- Solidity (PCA 0.593, IQR ≈ 1.123)
- Average Intensity (PCA 0.572, IQR ≈ 1.397)
- Average Contrast (PCA 0.563, IQR ≈ 1.548)
- Isoperimetric Factor (PCA 0.555, IQR ≈ 1.629)

During exploratory analysis, visualizations of key features revealed further insights:

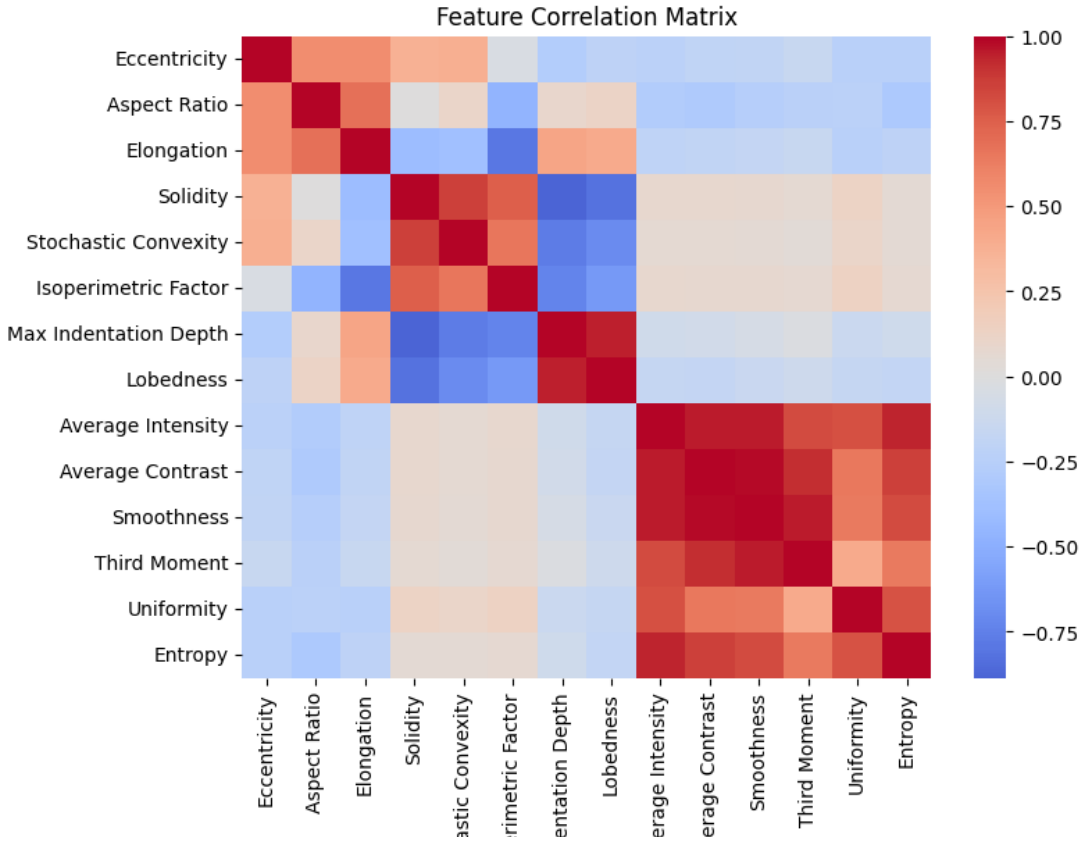

Figure 2 –Feature Correlation Matrix *shows correlation between features; selecting one or two redundant features may suffice for clustering without losing too much information.* ***Highly correlated features should be reduced prior to use in clustering algorithm.***

A correlation matrix (figure 1) indicated high redundancy within texture features (Average Intensity, Average Contrast, Smoothness, Third Moment, Uniformity, Entropy). There were also indications of moderate shape interrelationships, with eccentricity, aspect ratio, and elongation all positively correlated. Max Indentation Depth and Lobedness also strongly correlated.

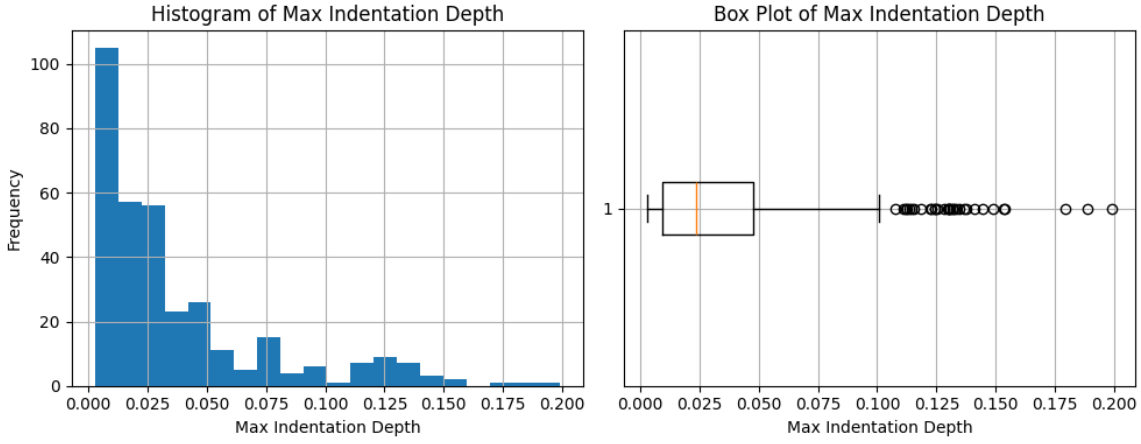

Figure 1 - *Max Indentation* Histogram and Boxplot (raw data): *Histogram – most values cluster around zero with long right tail extending towards max, indicating most leaves have shallow indentations. Boxplot – median sits near 25% quartile with box towards lower end of scale, confirming skew. Outliers above upper quartile indicates deep indentations.*

Histograms and Boxplots of the raw data highlighted skew in *Max Indentation Depth* (e.g. figure 2) and *Average Intensity* key features, whereas *Average Cost (figure 3)* and *Isoperimetric Factor* provided the most reliability variation across the full dataset.

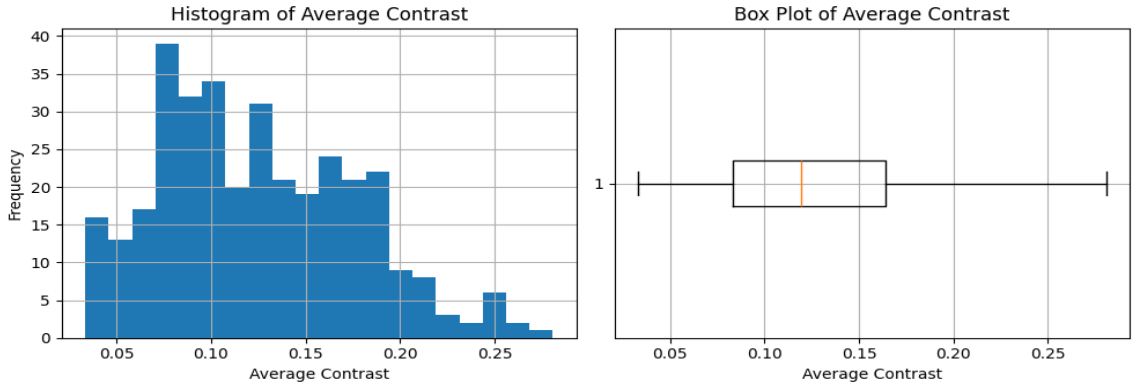

Figure 3 - *Average Contrast* Histogram and Boxplot (raw data): *Histogram shows more standard bell-curve shape, with moderate tailing but no extreme contrasts. Boxplot is centered, with fairly evenly distributed whiskers and few points lying beyond them, indicating limited outlier influence.*

Pairwise plots (figure 4) of the scaled data revealed strong bivariate splits between Max Indentation vs. Solidity (shape) and Intensity vs. Contrast (texture), which produced two visually unambiguous groups (clusters). Isoperimetric Factor interacted with both shape and texture features.

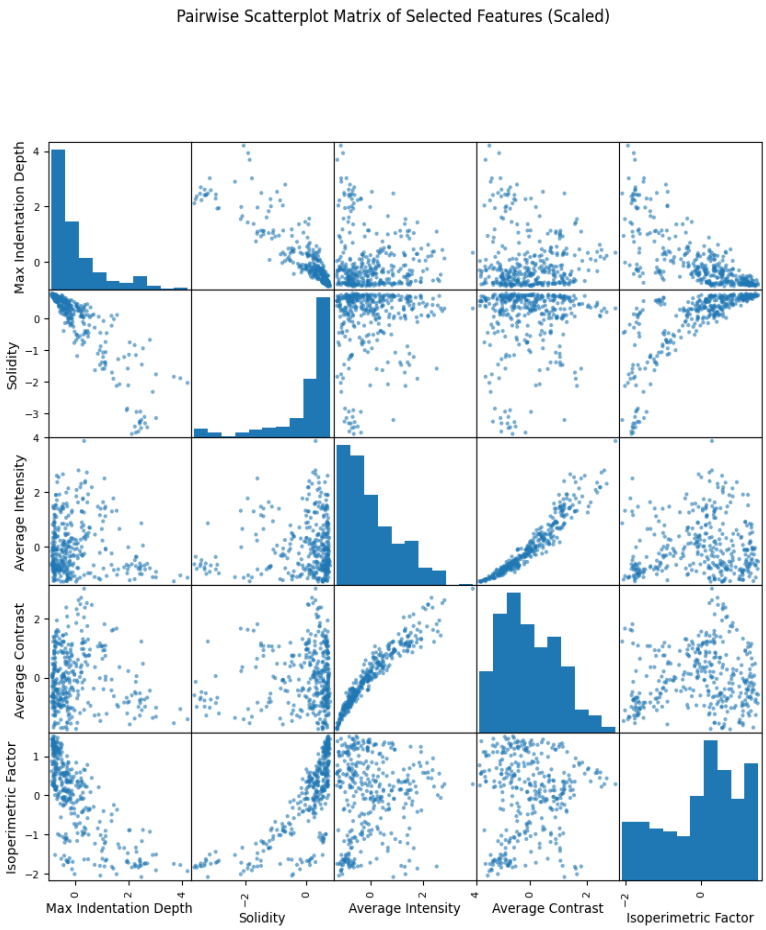

Figure 4 -Pairwise plots of Key Features (scaled)

4. Determining Optimal *k* for k-Means

4.1 Elbow Method

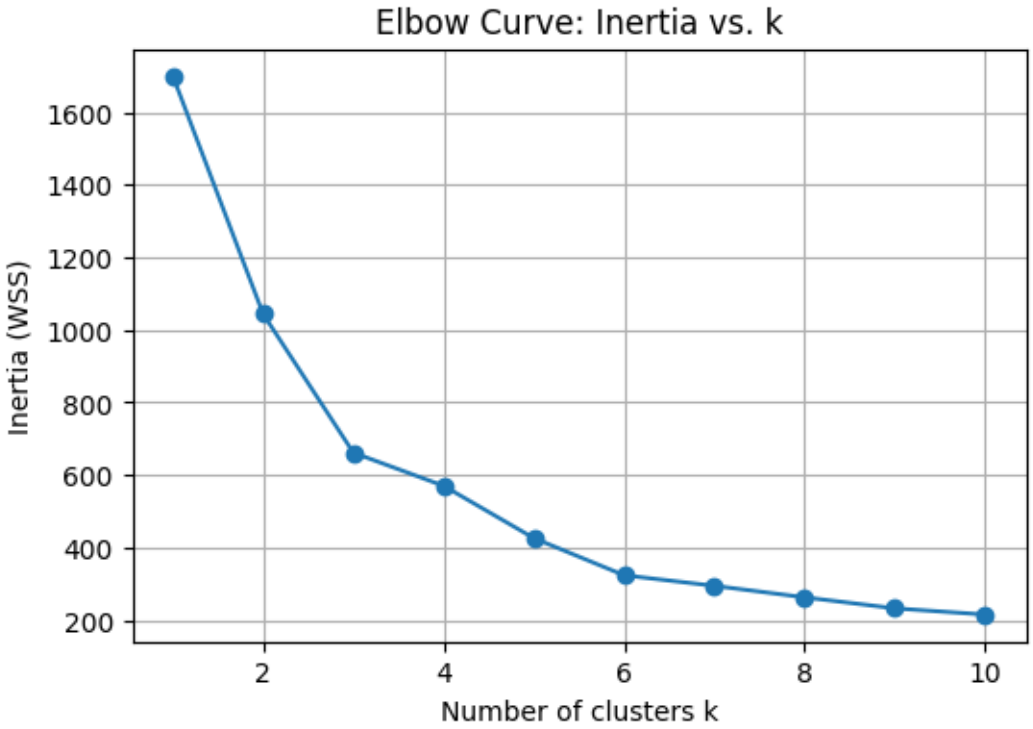

Figure 5 Elbow Curve: Inertia vs k (five key features) *Elbow curve shows bend at k=4*

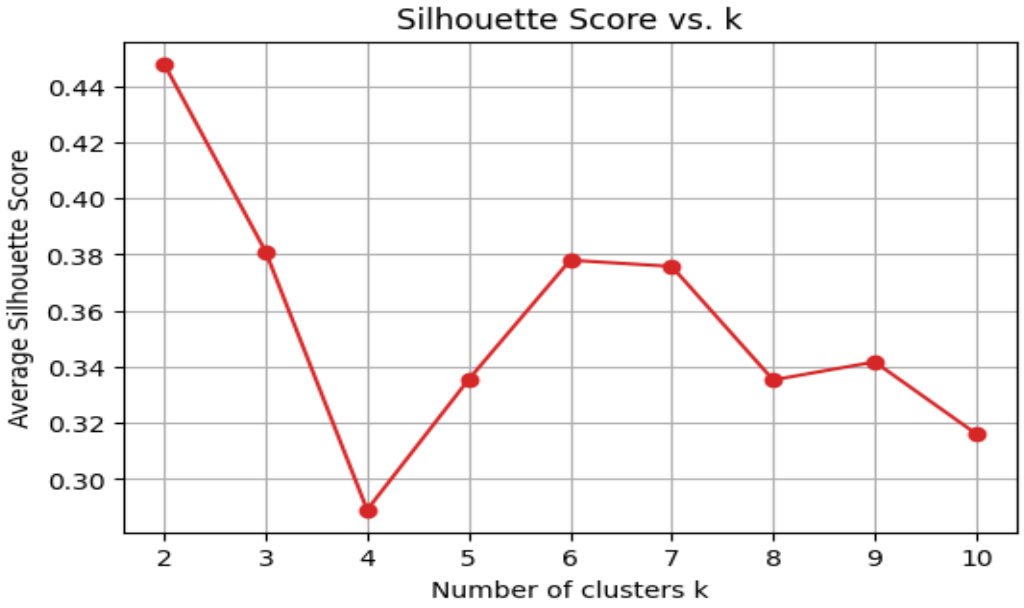

Figure 6 Silhouette Score vs k (5 key features)

The plot of inertia versus the number of clusters k (figure 5) shows steep decline from k=1 to k=3, and elbowing around k=4. After 4 clusters, the inertia reduction does not appear to be large.

However, the plot of average silhouette scores vs number of k clusters indicated that the most cohesive and well-separate groups occurred at k=2. However, splitting the leaves into just group 1 and group 2 doesn't appear to capture meaningful biological or morphological nuance.

To ensure we didn't miss any orthogonal structures, we reran the elbow method and silhouette analysis with all 14 features:

- With all 14 features, ,the average silhouette scores drop across most k, meaning that including the additional features slightly blur the cluster boundaries and reduce cohesion.
- While the peak at k=2 still remains the highest, it is too broad to consider.
- With all 14 features, the trough moves from k=4 to k=5, with a peak occuring at k=4, meaning that including all features strengthens distinction at four clusters. Some secondary features (most likely Lobedness or Elongation, per initial exploration) complements the primary dimensions and adds structure.
- All 14 features more neatly aligns with the elbow plot interpretation (k=4), albeit introduces noise.

Based on our analyses, we decided on analyzing the cluster plot at k=4.

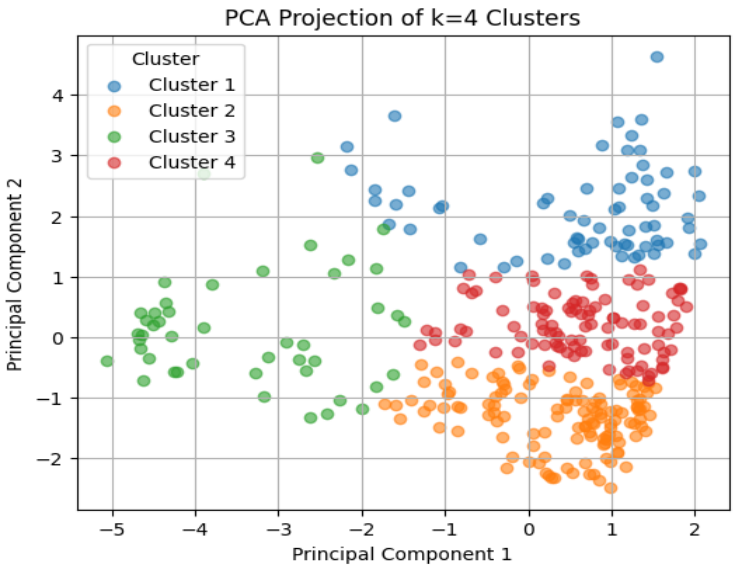

Figure 7 PCA Projection of k=4 clusters

	PC1	PC2
Max Indentation Depth	-0.562895	0.164466
Solidity	0.566091	-0.170663
Average Intensity	0.202575	0.677381
Average Contrast	0.201230	0.677946
Isoperimetric Factor	0.530246	-0.159276

Table 3 PCA Loadings

For interpretation purposes, we reduced the five shape and texture measurements into two principal dimensions (2D shape) – the horizontal shape representing leaf shape and the vertical axis representing leaf texture.

Shape (PC1):

- Values towards the left correspond to leaves with pronounced lobing or indentations
- Values toward the right indicate smoother, more convex leaves

Texture (PC2):

- Lower values denote uniform, less textured surfaces
- Higher values denote greater contrast and intensity variations

When each leaf is plotted in this 2D space, four clusters emerge:

Cluster 1: Lobed and Highly Textured (upper left)

Cluster 2: Lobed and subtly textured (lower left)

Cluster 3: Smooth and Highly Textured (upper right)

Cluster 4: Moderate shape and texture (central-right)

Each cluster occupies its own quadrant, with minimal overlap at boundaries, indicating well-separated groupings.

Anomaly Detection (Outliers skewing cluster centroids?)

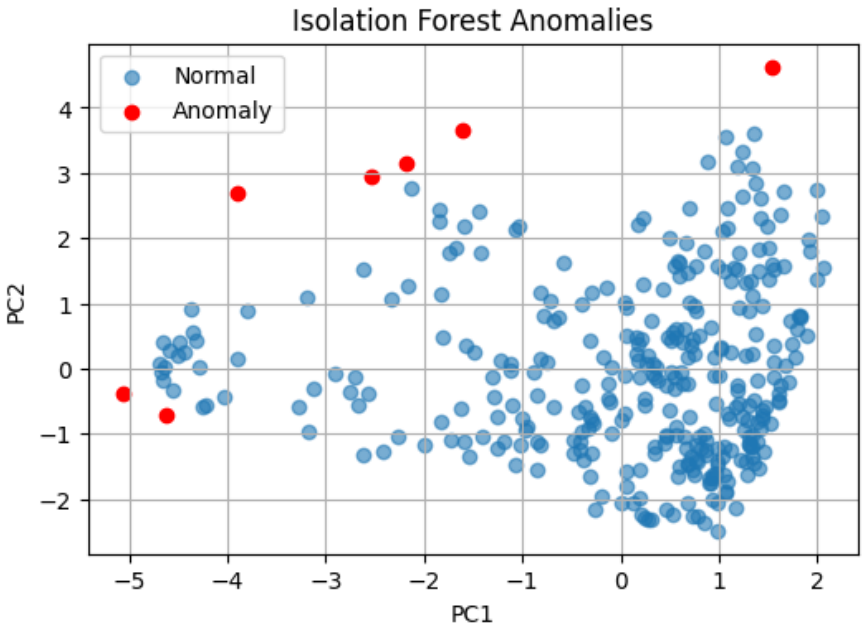

Figure 8 Isolation Forest flags in PCA Plot

An isolation forest flagged approximately 2% (7 out of 340) of leaf observations (in red) as anomalies when projecting into the first two principal components. Anomalies lie on the periphery rather than intermingled with the main clusters, posing minimal risk to the overall cluster structure.

Distance-based metrics

TSS: 1700.00
WSS (inertia): 570.40
BSS: 1129.60
R^2 (BSS/TSS): 0.664

R^2 = 0.664 indicates that 66.4% of the total variance in the selected feature space is explained by the clustering. The four‐cluster solution captures a substantial portion of the structure in leaf morphology and texture, corroborating the elbow and silhouette diagnostics and the clear separation seen in our PCA and feature‐pair visualizations.

Together, R^2 = 0.664, silhouette ≈ 0.30, and inertia reduction indicate robust separation.

5. Conclusion

4.1 Summary of Findings

In this study, we applied k-means clustering to five carefully selected, standardized leaf-morphology and texture features (Max Indentation Depth, Solidity, Average Intensity, Average Contrast, and Isoperimetric Factor) from the Leaf dataset. Diagnostic analyses—including the elbow method, silhouette scores, and distance-based metrics—consistently identified **k=4** as the optimal number of clusters. These four clusters explain **66.4%** of the total variance in feature space (R^2 = 0.664) and exhibit clear separation in both PCA projections and feature-pair scatterplots. Each cluster corresponds to a distinct combination of margin complexity and texture:

1. **Cluster 1:** Smooth, convex leaves (high solidity; low indentation) with moderate texture.
2. **Cluster 2:** Deeply lobed leaves (high indentation; low solidity) with muted texture.
3. **Cluster 3:** Smooth margins combined with strong texture (high intensity and contrast).
4. **Cluster 4:** Moderately lobed, moderately textured leaves with intermediate contour complexity.

4.2 Limitations

- **Unsupervised Nature:** Without ground-truth species labels, cluster validity relies on internal metrics and visual inspection. External validation (e.g., expert-annotated species) would strengthen confidence.
- **Cluster Assumptions:** k-means assumes spherical clusters of similar size. Although diagnostics supported this assumption, any non-globular groupings remain unmodeled.

4.3 Areas for Future Improvement

- **Alternative Clustering Methods:** Applying hierarchical clustering or DBSCAN could uncover non-spherical patterns or density-based groupings, offering complementary insights.
- **Semi-Supervised & Ensemble Approaches:** Incorporating a small set of labeled specimens or combining multiple clustering algorithms may improve robustness and interpretability.

5. References

Arthur, D., & Vassilvitskii, S. (2007). *k*-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 1027–1035). Society for Industrial and Applied Mathematics.

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1989). Learnability and the Vapnik–Chervonenkis dimension. *Journal of the ACM, 36*(4), 929–965. https://doi.org/10.1145/18565.18578

Eklundh, L., & Olsson, L. (2003). Vegetation index trends for the African Sahel 1982–1999. *Geophysical Research Letters, 30*(9), 1–4. https://doi.org/10.1029/2003GL017376

Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques (3rd ed.). Elsevier Science.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning with applications in Python* (1st printing July 5, 2023). Retrieved from https://www.statlearning.com/

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Silva, P. F. (2008). *Development of a system for automatic plant species recognition* (Master's thesis). Faculdade de Ciências, Universidade do Porto. http://hdl.handle.net/10216/67734

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks, 16*(3), 645–678. https://doi.org/10.1109/TNN.2005.845141