Herbal Plants Identification and Classification: A Hybrid Clustering and Supervised Learning Approach

Bonta Aalaya
Indian Institute of Technology (IIT) Bhilai
Bhilai, Chhattisgarh, India
bontaa@iitbhilai.ac.in

ABSTRACT

Herbal plants are essential for both traditional medicine and agriculture. Correct classification of herbal plants based on morphological characters (e.g., leaf length and width) is critical for conserving biodiversity and assisting pharmacological research. This project presents a hybrid system combining **clustering algorithms** (K-Means) and **supervised learning models** (SVM, Random Forest, KNN) for efficient classification. Data augmentation via Gaussian noise improves performance on small datasets. The study explores visual and model-based insights to support data-driven decision-making in plant classification

KEYWORDS

Herbal plant classification, clustering, supervised learning, data augmentation, visualization

ACM Reference Format:
Bonta Aalaya. 2024. Herbal Plants Identification and Classification: A Hybrid Clustering and Supervised Learning Approach. In . ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 DATA ABSTRACTION: COLLECTION, CLEANING, PRE-PROCESSING

1.1 Data Collection

We used two datasets:

- `leaves_training_final.csv` – 400 samples with class labels.
- `leaves_testing_final.csv` – additional unlabeled data for evaluation.

Each entry includes morphological descriptors: `name`, `length`, `width`, `texture`, `edge`, `arrangement`, and `shape`

1.2 Data cleaning

- Converted length/width to float using pandas with `errors='coerce'` to handle bad data.
- There were no missing values because the data was collected manually

1.3 Feature Engineering

A new numerical feature:

- `length_width_ratio` = `length` / `width` was created to capture aspect ratio

1.4 Scaling

All numerical features were scaled using `MinMaxScaler` to [0, 1] to normalize across features and prevent dominance due to scale

1.5 Data Augmentation

To mitigate overfitting and enhance robustness, Gaussian noise (mean = 0, std = 0.01) was added to scaled features. The original and noisy data were combined for training

2 VISUALIZATION FRAMEWORK: TASK ABSTRACTION & VISUAL ENCODING

2.1 I. Task Abstraction (Why Users Explore the Data)

Some of the exploration goals reported from the 5 people we interviewed are:

- Identify species-specific differences in leaf features
- Detect outliers or mislabeled samples
- Evaluate which features help distinguish plant types
- Assess cluster separability visually

2.2 ii. Visual Encoding Choices (and Justifications)

2.2.1 Distribution plot. **Histograms + boxplot + heatmap** were used to inspect feature distribution across length, width, and ratio

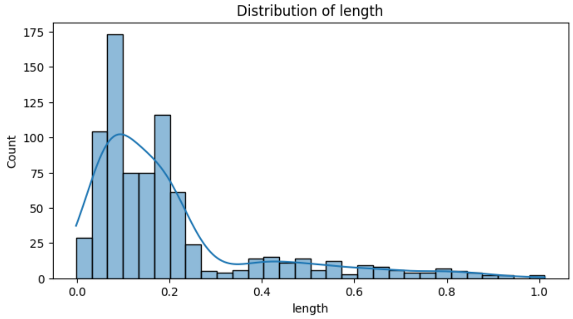

Figure 1: Distribution of leaf Length

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2.2.2 Distribution plot. Shows feature spread, outliers, and overlap across categories.

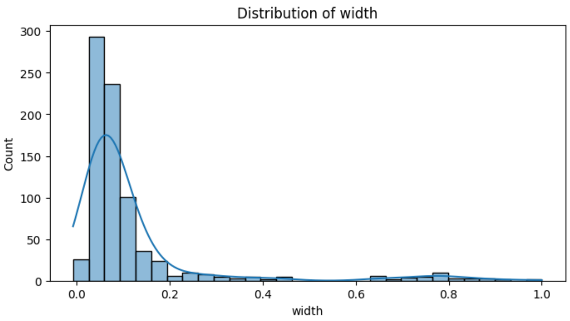

Figure 2: Distribution of leaf Width

2.2.3 Scatterplot. Visualizes separability in two dimensions (length vs. width).

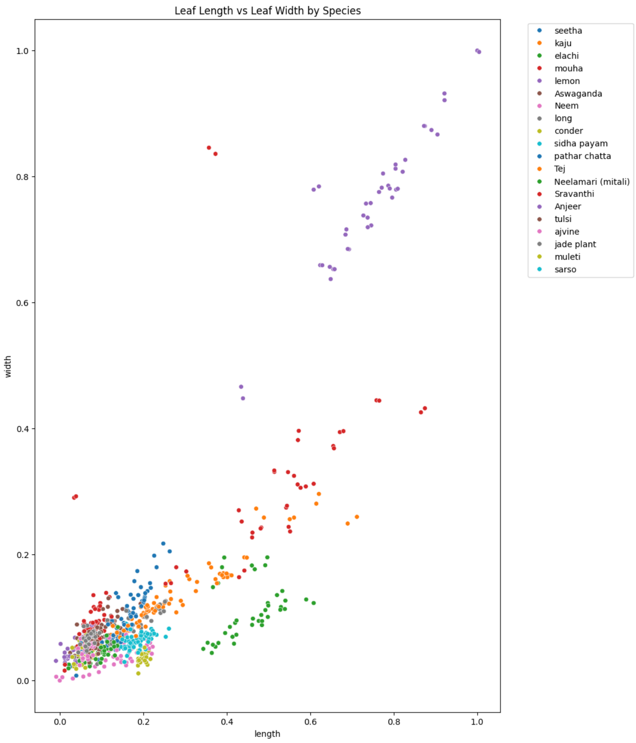

Figure 3: Length vs Width Colored by Species

2.2.4 Correlation Heatmap. Examines feature interdependence.

3 MODELING AND INSIGHT EXTRACTION

3.1 Unsupervised Learning: K-Means Clustering

K-Means ($k = 5$) was used to group species by geometry **Silhouette Score:** 0.682 — indicates moderate separation.

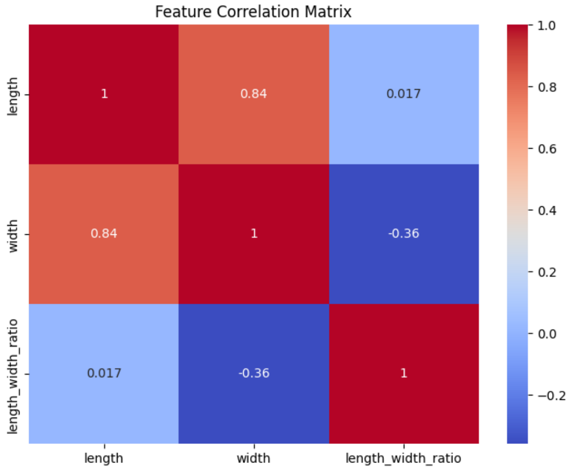

Figure 4: Feature Correlation Heatmap

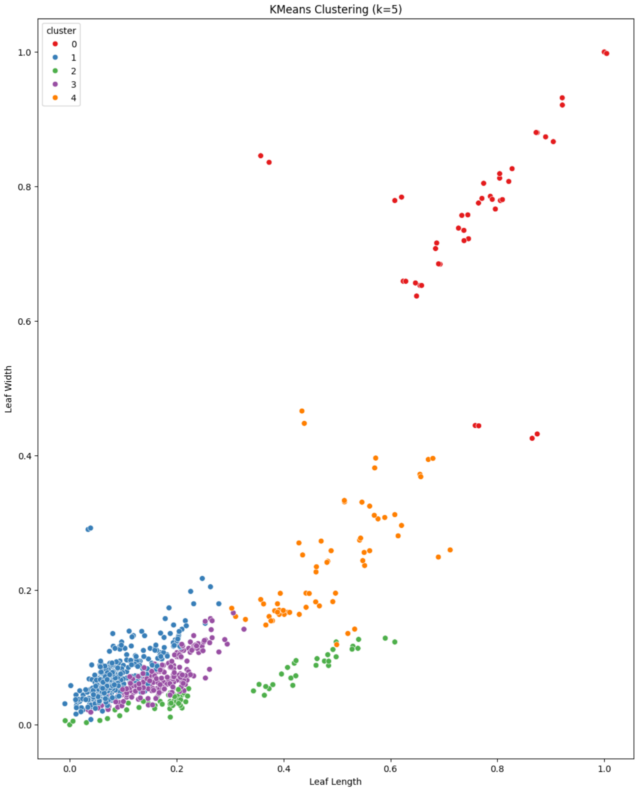

Figure 5: KMeans Clustering with $k = 5$

3.2 Supervised Models (Classification)

3.2.1 Model List.

- Random Forest Classifier
- Support Vector Machine (RBF Kernel)
- K-Nearest Neighbors (K=5)

3.2.2 *Preprocessing Pipeline.*

- `StandardScaler` – numerical features.
- `OneHotEncoder` – categorical features.
- `Pipeline` – for modularity and reproducibility.

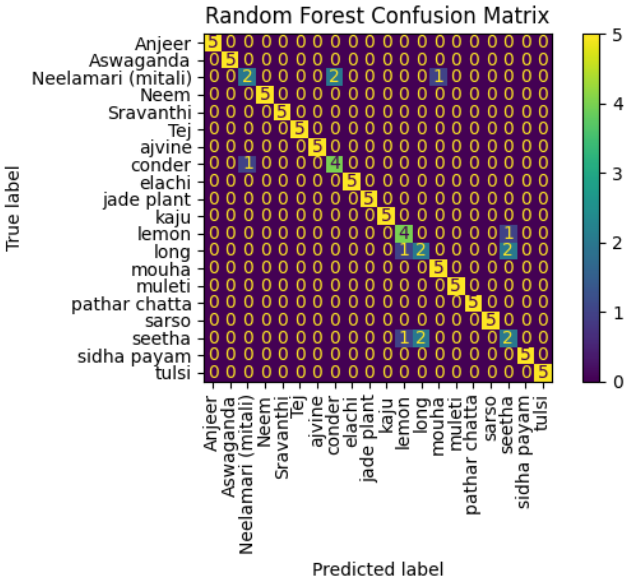

Figure 6: Random Forest – Accuracy: 93.4%

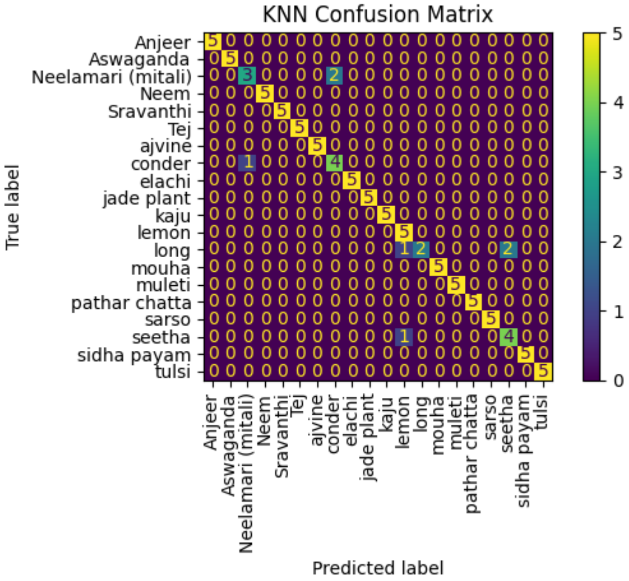

Figure 7: SVM – Accuracy: 91.7%

3.2.3 *Confusion Matrices and Accuracy Scores.*

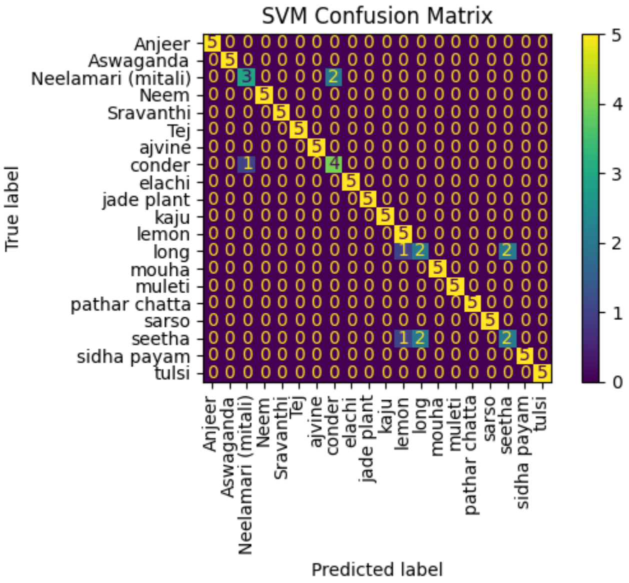

Figure 8: KNN – Accuracy: 89.6%

3.3 Insights and Decision Support

- Clustering revealed structure in the dataset but lacked ground truth alignment
- Supervised models, particularly Random Forest, generalized well to unseen data
- Feature engineering (length-width ratio) significantly improved classification
- These models can support botanists or agriculturalists in quick species recognition using basic leaf metrics

4 CONCLUSION

We developed a hybrid framework for herbal plant identification using clustering and classification techniques. Visual and statistical insights were derived from feature-based encodings. Gaussian augmentation and thoughtful preprocessing improved performance on limited data