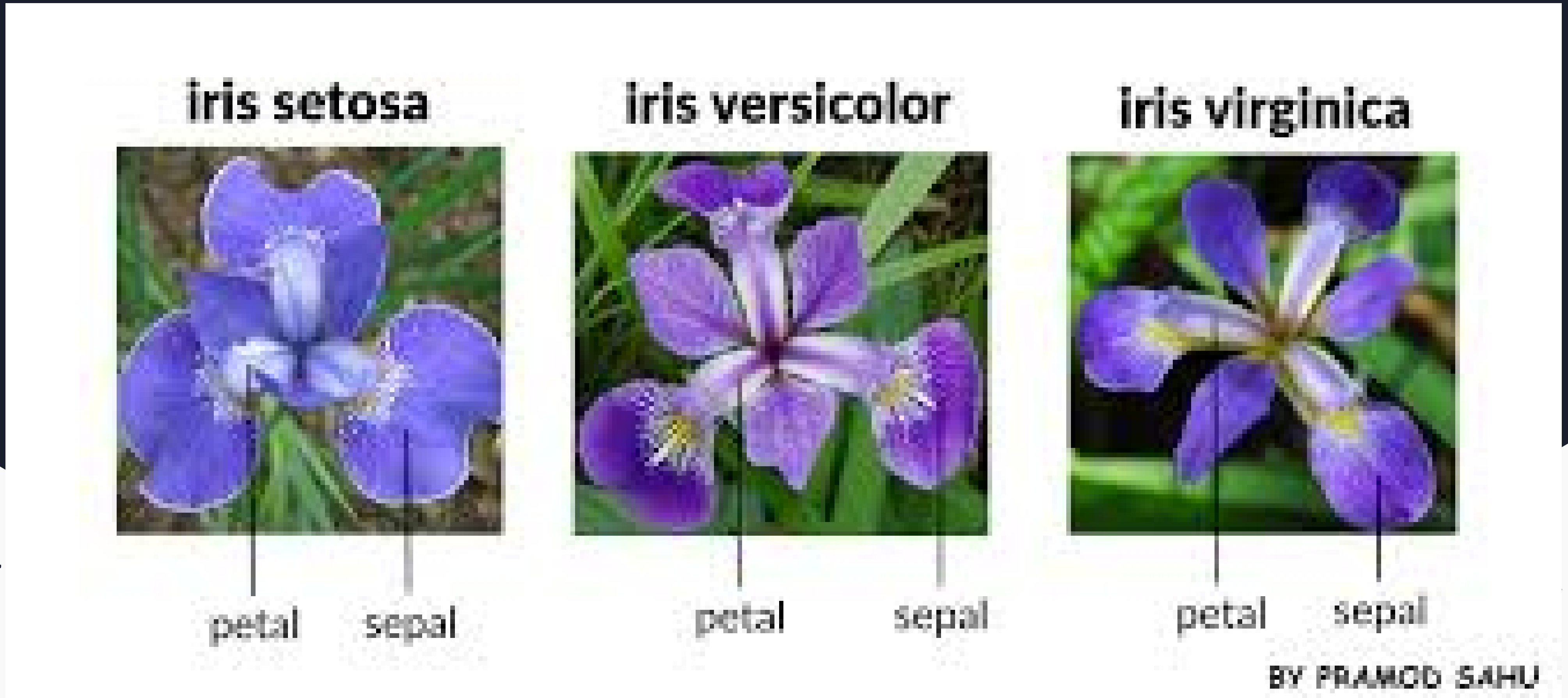
Self Learning

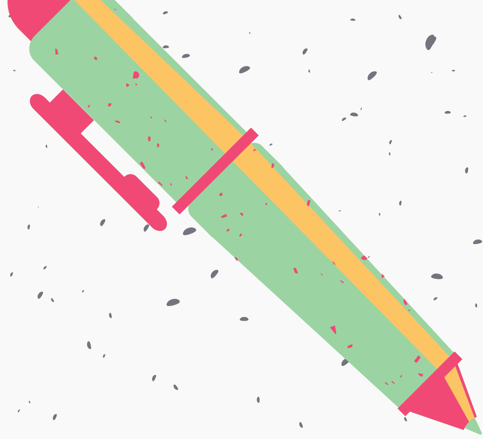
Presentation

CASE STUDY : USE OF IRIS DATASET FROM SCIKIT AND APPLY
K-MEANS CLUSTURING METHOD

Introduction to the Iris Dataset

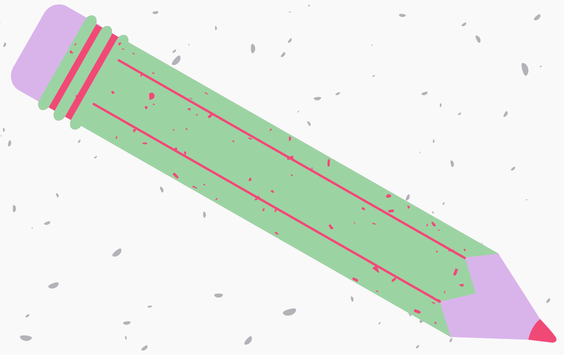
Description: The Iris dataset is a classic dataset in machine learning, containing 150 samples of iris flowers from three species: Setosa, Versicolor, and Virginica. Each sample has four features: sepal length, sepal width, petal length, and petal width.





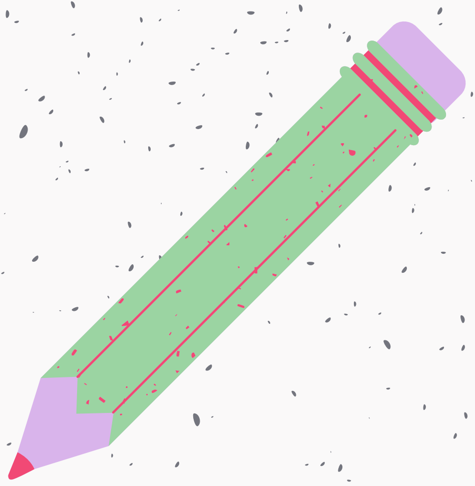
Features & Species

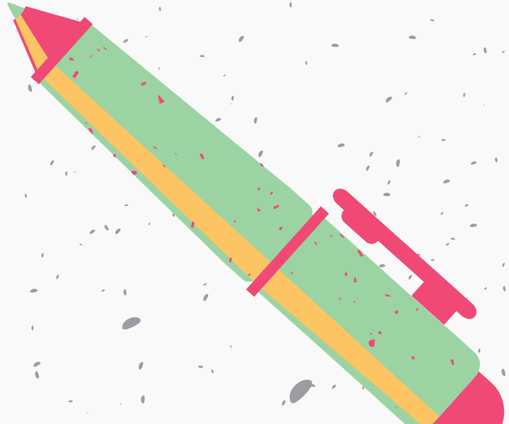
Species / Features



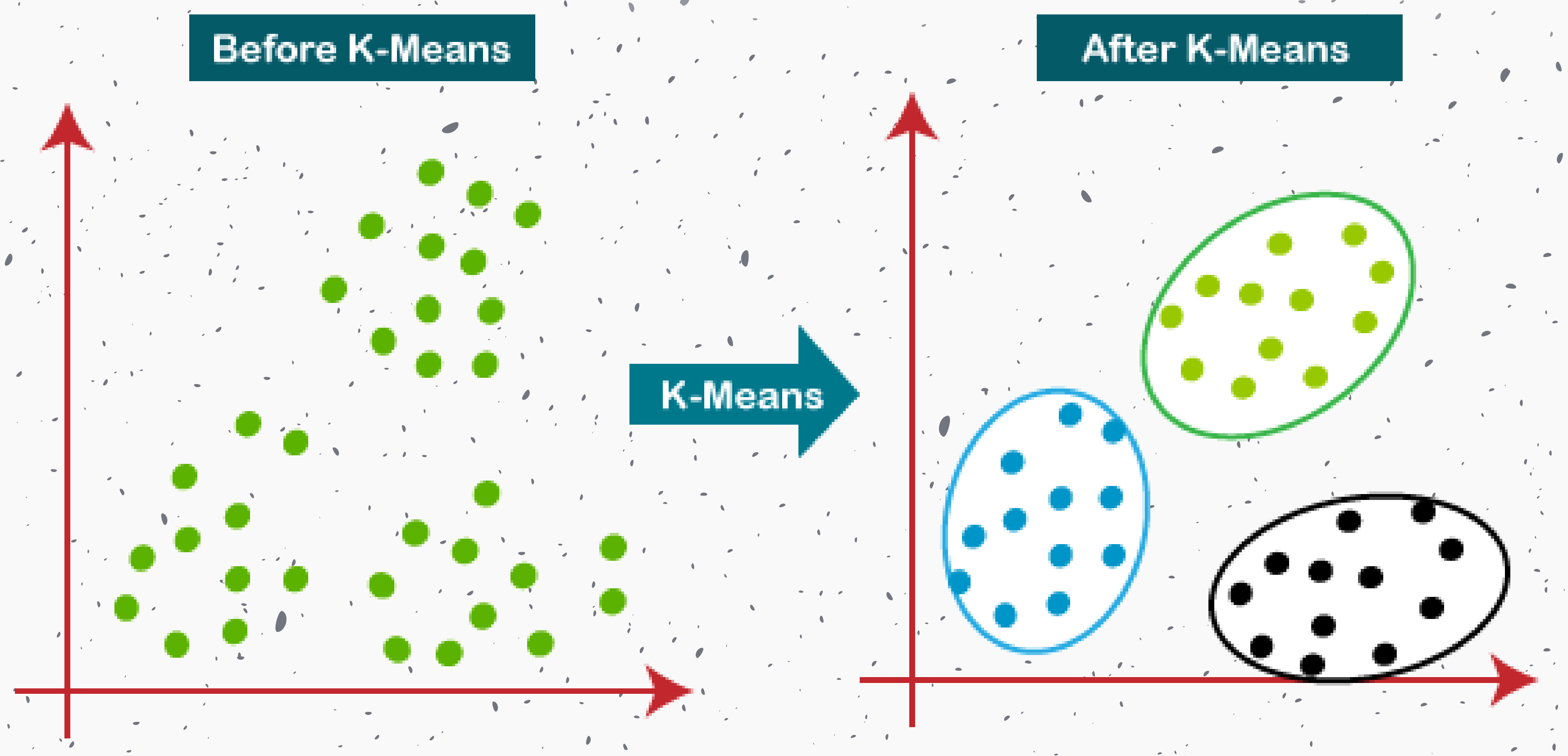
- Setosa
- Versicolor,
- Virginica

- sepal length
- sepal width
- petal length
- petal width





What is **K-Means Clustering?**



- Definition: K-Means is an unsupervised learning algorithm that partitions data into K distinct clusters based on feature similarity.
- Objective: Minimize the variance within each cluster.

Steps in K-Means Clustering

1. Choose the number of clusters (K).
2. Initialize centroids randomly.
3. Assign each data point to the nearest centroid.
4. Recalculate centroids as the mean of assigned points.
5. Repeat steps 3 and 4 until convergence.

Implementation: Step-by-Step

1. Import Necessary Libraries
2. Use libraries like NumPy, Matplotlib, and scikit-learn.
3. Load the Iris Dataset
4. Fetch the Iris dataset from scikit-learn’s datasets module.
5. Extract the feature data (sepal length, sepal width, petal length, petal width).
6. Choose Number of Clusters (K)
7. Set K = 3 because the Iris dataset has 3 species (Setosa, Versicolor, Virginica).
8. Initialize K-Means Model
9. Create a KMeans model by specifying the number of clusters and a random state (for reproducibility).
10. Fit the Model to the Data
11. Apply the KMeans model on the Iris data to compute cluster centers and predict cluster indices.
12. Obtain Cluster Labels
13. After fitting, the model assigns a cluster label (0, 1, or 2) to each data point.
14. Visualize the Results
15. Create a 2D scatter plot using two features (e.g., sepal length and sepal width).
16. Color the points according to their assigned cluster labels.
17. (Optional) Compare with Actual Species
18. Though K-Means is unsupervised, you can roughly compare the clustering output with the true labels of the Iris dataset.
19. Important: Cluster labels are arbitrary — label 0 doesn't necessarily mean "Setosa."

Conclusion

Summary:

- K-Means effectively clusters the Iris dataset into three groups.
- Some overlap may occur due to similarities between species.
- Choosing the right number of clusters (K) is crucial.