Unsupervised Algorithms in Machine Learning Final Project

By Albert Diaz

Problem Statement

- Automate segmentation of unlabeled data to eliminate manual labeling overhead.
  - Manual annotation is slow, expensive, and inconsistent.
  - Clusters enable targeted marketing, anomaly detection, and pre-segmentation for supervised models.

Iris Dataset Overview

- 150 samples, 3 species (50 each): Setosa, Versicolor, Virginica.
  - Features: sepal length, sepal width, petal length, petal width (cm).
  - No missing values; mild outliers; petal length & width highly correlated.
  - Balanced, non-trivial, ideal for demonstration of clustering and PCA.

Methodology: EDA & Preprocessing

- Exploratory Data Analysis:
  - Summary statistics, histograms, boxplots, correlation heatmap.
  - Identify skewness, outliers, and feature correlations.
- Preprocessing:
  - Impute/drop missing values.
  - Standardize features to zero mean & unit variance.
  - Apply PCA when needed for dimensionality reduction.

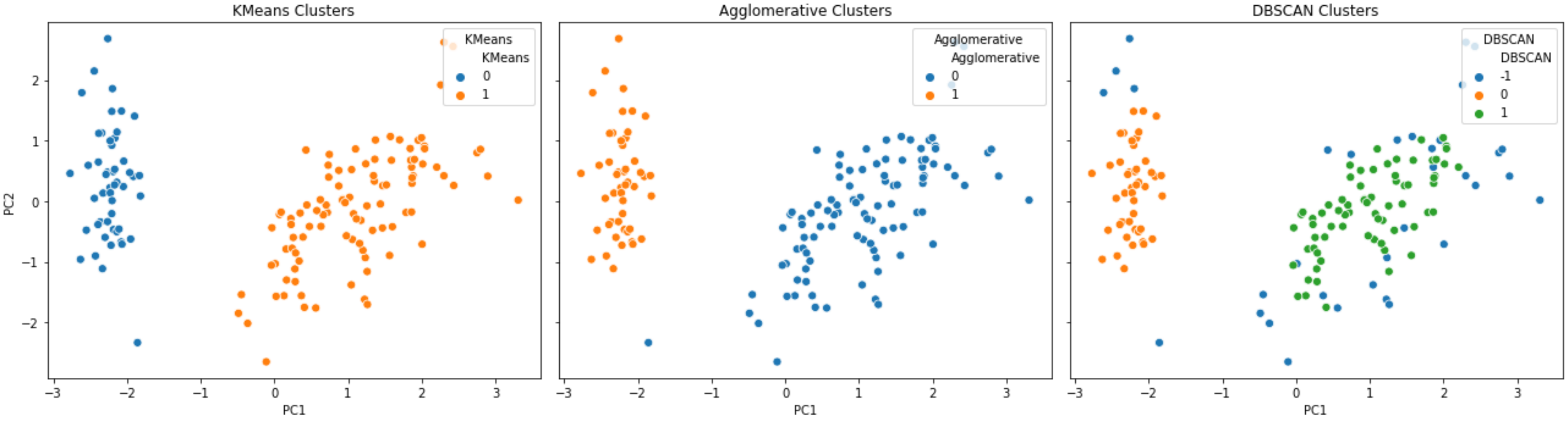
Methodology: Clustering Algorithms

1. K-Means (k =2)
2. Agglomerative Hierarchical (Ward linkage)
3. DBSCAN (ε=0.5, min_samples=5); labeled noise points.

Key Results & Metrics

Algorithm	Parameter	Silhouette Score	Notes
K-Means	k = 2	0.58	Isolated Setosa; merged others
Agglomerative	2 clusters	0.55	Similar to K-Means
DBSCAN	ε=0.5, min_samples=5	0.45	2 clusters + 34 noise points

Visualization



Conclusion & Next Steps

- Conclusions:
  - Pipeline uncovers structure without labels.
  - K-Means & Hierarchical produced robust clusters.
- Next Steps:
  - Explore Gaussian Mixture Models & hyperparameter tuning.
  - Integrate into production workflows for scalable insights.