Predicting Iris Flower Species based on their Features

By Marc Walden

Introduction

The Iris dataset contains measurements of sepal length, sepal width, petal length, and petal width for three species of iris flowers: Setosa, Versicolor, and Virginica. In this report, I intend to predict the species of each iris flower based on their features, and conclude by comparing my results to the actual data. To achieve this, I will use common algorithms in data analysis and machine learning. Firstly, I will reduce the dimensions of the Iris dataset using PCA to remove unnecessary noise. Subsequently, I will apply K-means clustering (also known as Lloyd's algorithm), density-based clustering (DBSCAN), agglomerative clustering, and divisive clustering to categorize each data point to a specific species. Finally, I will compare the results of each algorithm to the real categorization of each of the iris flowers using confusion matrices to see which one worked best. Ultimately, my goal is to showcase the applications of unsupervised machine learning that can reveal deep insights from seemingly simple datasets.

Application of PCA

In my analysis, I utilized Principal Component Analysis (PCA) to reduce the dimensionality of the iris to only two columns. The process involved several key steps. Firstly, I standardized the features of the dataset to ensure uniform contribution to the PCA analysis. This standardization step is crucial as it prevents features with larger scales from dominating the analysis. Next, I applied PCA to the standardized dataset, specifying to retain only two principal components derived from linear combinations of the original features to simplify visualization. By transforming the dataset into a lower-dimensional space, I effectively reduced the dataset's dimensionality while preserving its essential characteristics.

Figure 1 below shows the Iris dataset reduced to two dimensions after applying PCA. Note that, although a color code reveals which species of flower each data point is, the PCA algorithm only outputs the location of these data points with respect to the axes PC1 and PC2, which are linear combinations of the variables sepal length, sepal width, petal length, and petal width of the data. It does not use the species of each iris flower to reduce the dimensions of the dataset. Keeping in mind that PCA has completely removed two dimensions of the dataset, it has done a fair job in visualizing the dataset. In other words, in Figure 1 it can be seen how the physical separation of the data points accurately represent the actual categorization of the data set. Although the separation of versicolor and virginica iris flowers are not clearly distinguishable for every single data point, this is because they do indeed have somewhat similar

characteristics. However, as we will see in the next section of this report, we can use K-means clustering to predict the species of each iris with a high degree of accuracy.

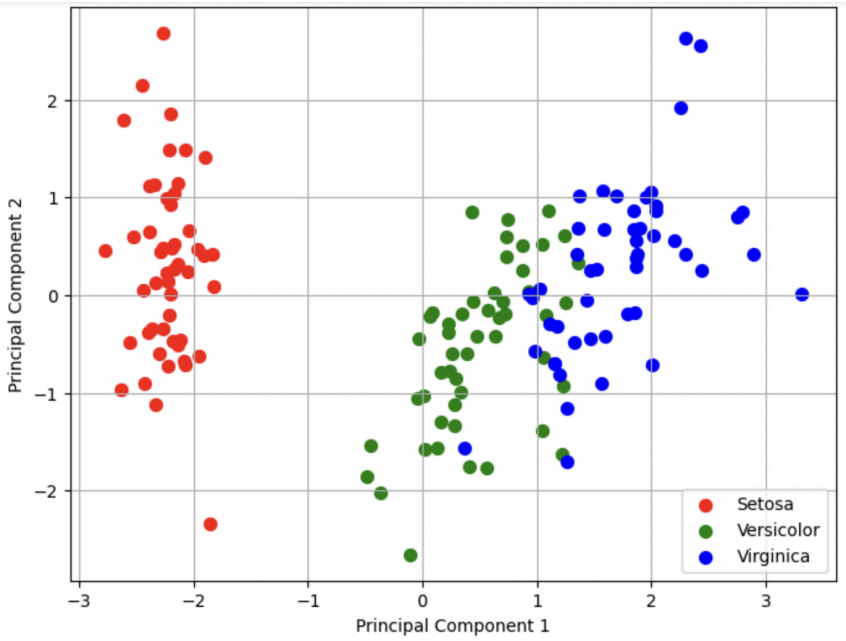

Figure 1: *The Iris dataset reduced to 2 dimensions using PCA*

Applying Clustering Algorithms to categorize the data

Now I will apply the four clustering algorithms to the graph in Figure 1 to group each iris flower into one of three different clusters. Each cluster intends to represent the species of the Iris flower: Setosa, Versicolor or Verginica. For k-means and agglomerative clustering, setting n_clusters = 3 ensured that I was separating the data into three clusters. However, to achieve three clusters using DBSCAN, I needed to play around with the epsilon variable which determines the maximum distance between two samples for them to be considered to be in the same neighborhood, and the min_samples variable which specifies the minimum number of samples required in one neighborhood for it to be considered a cluster. Moreover, for divisive clustering, I needed to adjust the maximum distance threshold at which the dendrogram will be cut to obtain the desired number of clusters (three in this case). Figure 2, 3, 4 and 5 below shows the output of the K-Means Clustering algorithm, the density-based clustering algorithm (DBSCAN), the agglomerative clustering algorithm and the divisive clustering algorithm, respectively.

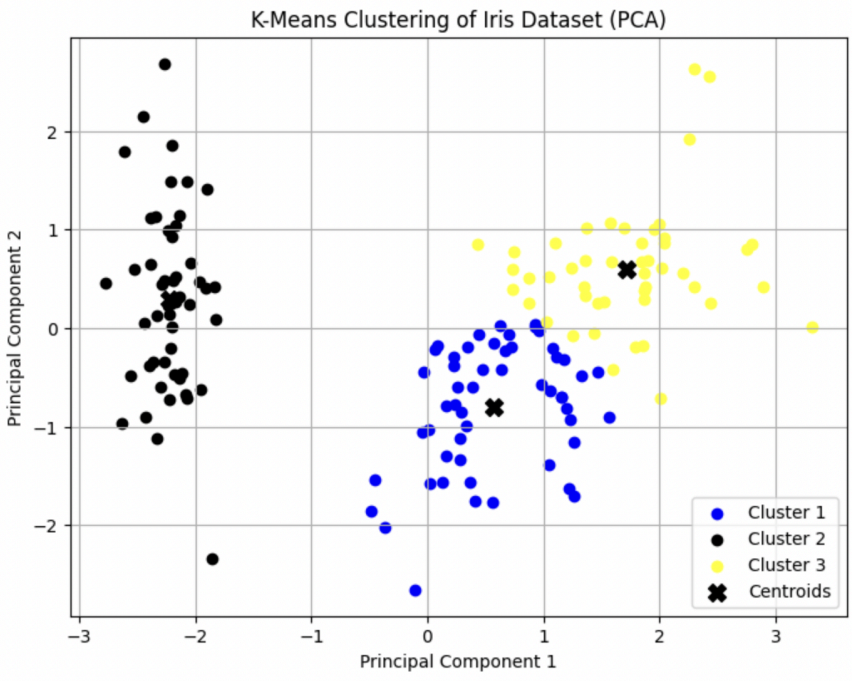

Figure 2: *K-Means Clustering*

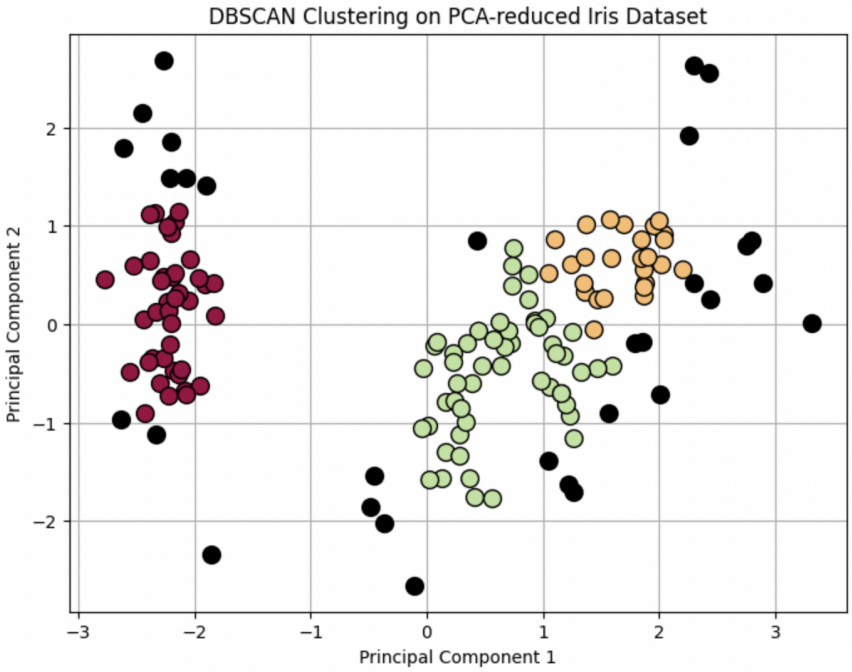

Figure 3: *Density-Based Clustering (DBSCAN) with eps = 0.32 and min_samples = 5. Black points are categorized as noise.*

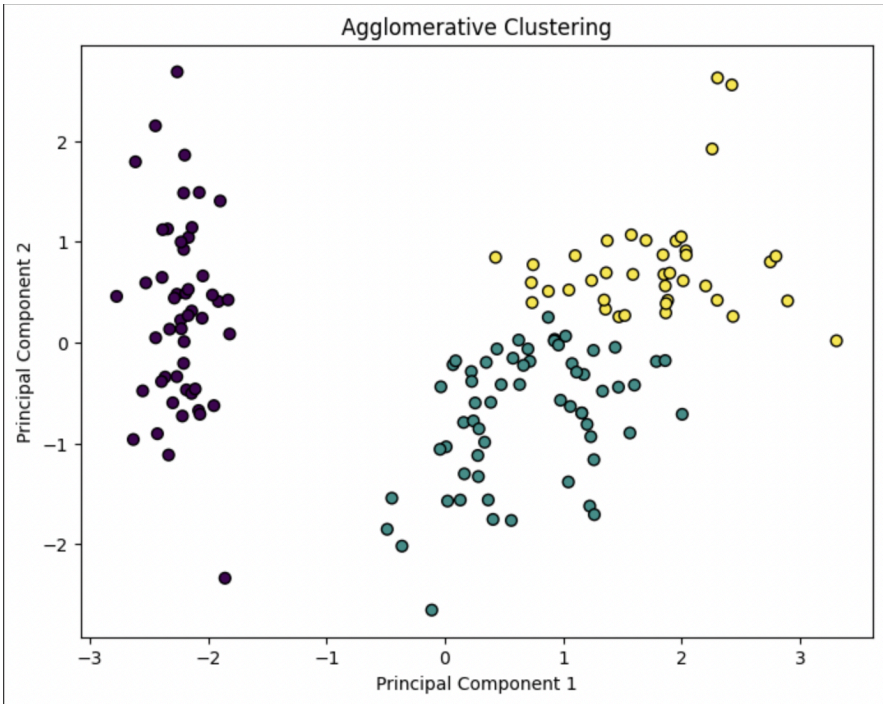

Figure 4: *Agglomerative Clustering*

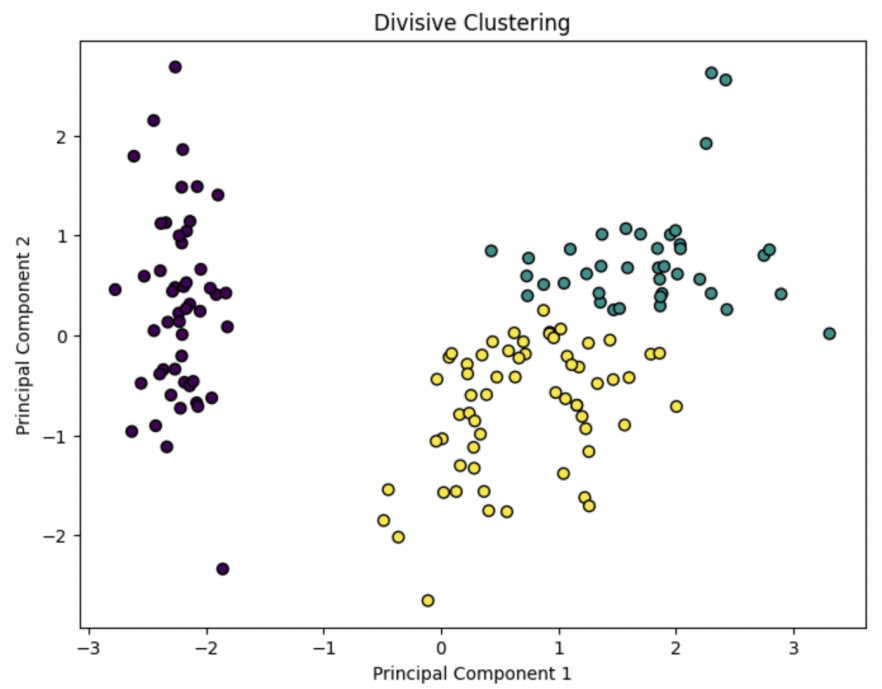

Figure 5: *Divisive Clustering with max distance = 8.*

Evaluation using a Confusion Matrix

In order to verify the accuracy of these four algorithms, their output needs to be compared to the true species. I created four confusion matrices that compare the categorization of the iris flowers using the four algorithms with their actual categorization. The columns of the matrix represent the species predicted, whereas the rows represent the actual species of the iris flower. Figure 6, 7, 8 and 9 below show the confusion matrices for the K-Means Clustering algorithm, the density-based clustering algorithm (DBSCAN), the agglomerative clustering algorithm and the divisive clustering algorithm, respectively. As seen in figure 6 below the K-means algorithm accurately classified all 50 Setosa irises. Moreover, it correctly identified 39 out of 50 Versicolor irises, but misclassified 11 Versicolor irises as Virginica. It also accurately classified 36 out of 50 Virginica irises, but incorrectly labeled 14 Virginica irises as Versicolors. Overall, k-means categorized the iris flowers with an 83.3% accuracy, derived by dividing the number of correct classifications by the total number of data points. Considering that some Versicolor and Virginica iris flowers have very similar features, this percentage is satisfactory. On the other hand, figure 7 shows how DBSCAN didn't perform as well due to the large amount of flowers it misclassified, including those that it treated as noise. It obtained an accuracy of 67%. Finally, as shown by figure 9, the agglomerative and divisive clustering methods, which outputs the same categorization were very accurate at predicting the species of the iris flowers with a precision of 87%.

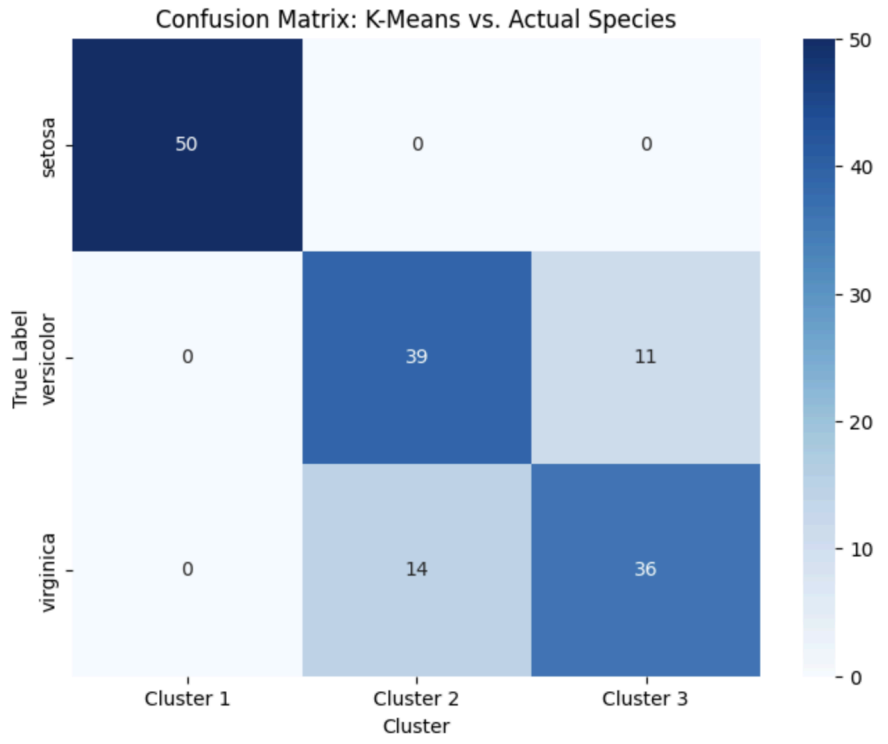

Figure 6: Confusion matrix for K-Means clustering

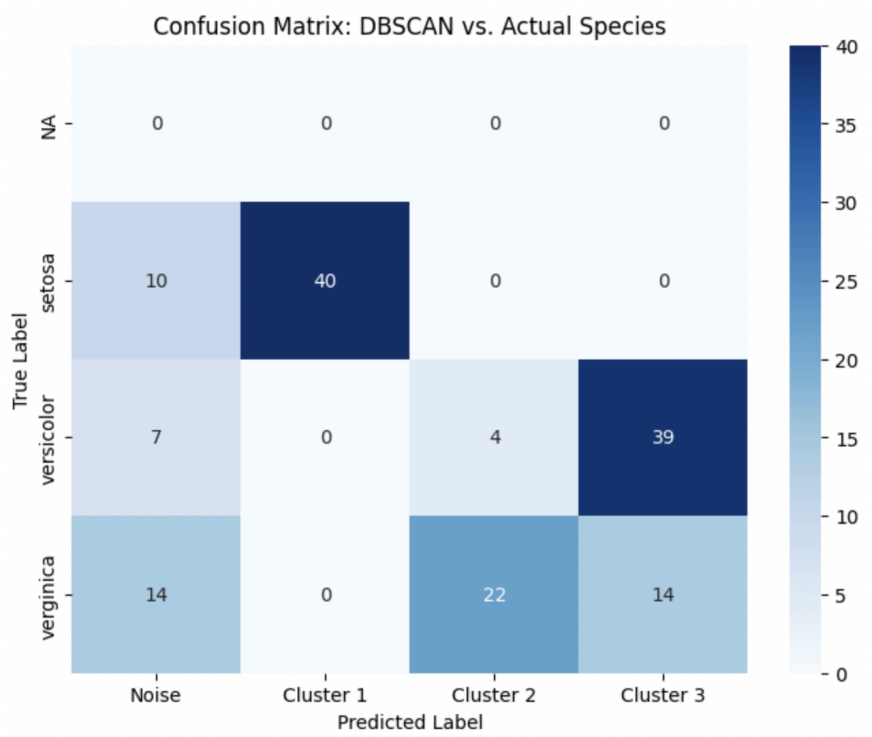

Figure 7: Confusion matrix for density-based (DBSCAN) clustering

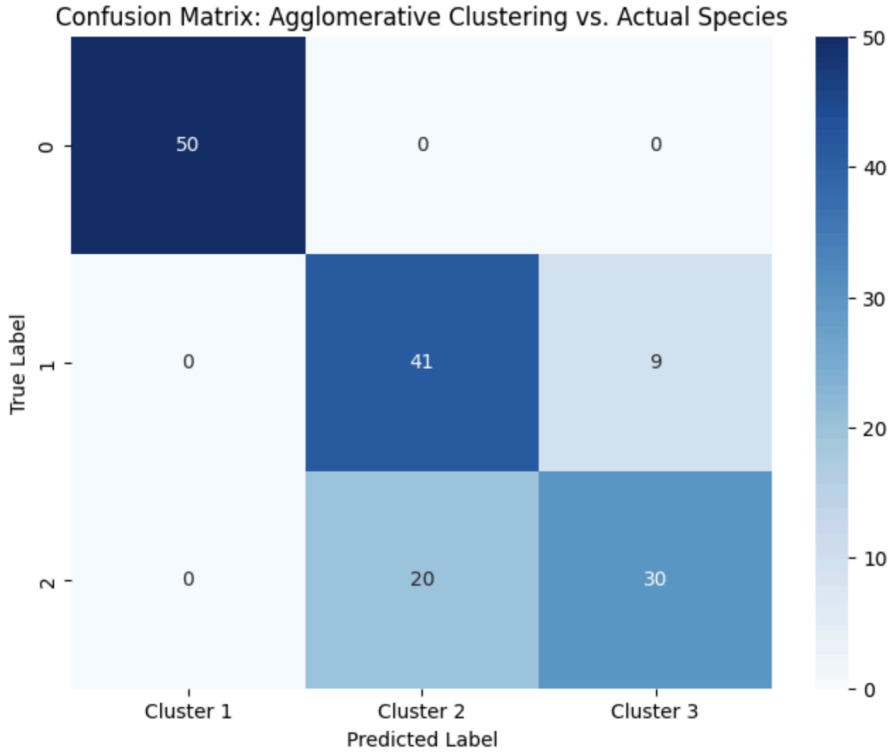

Figure 8: Confusion matrix for agglomerative clustering

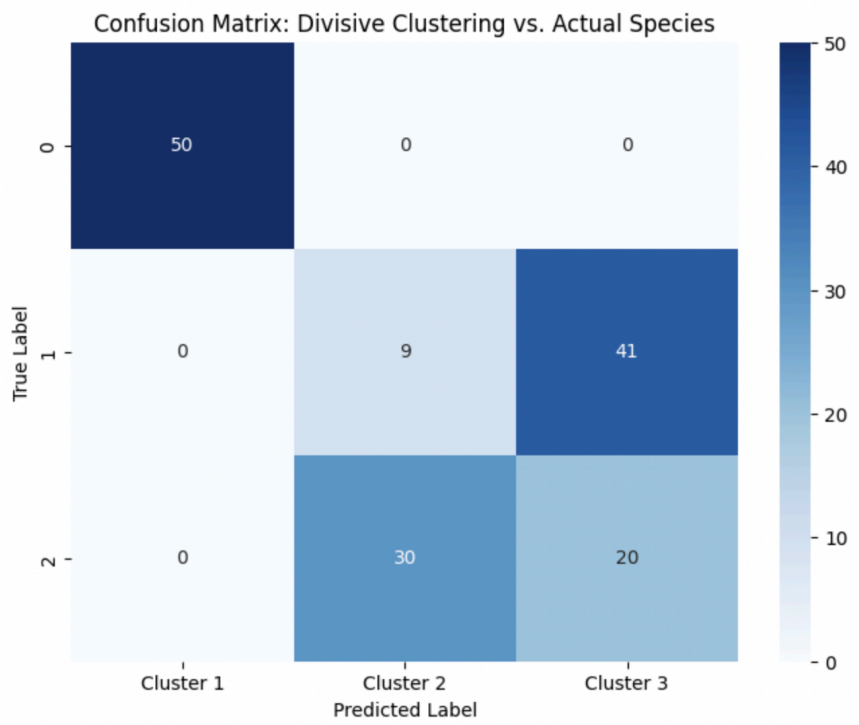

Figure 9: Confusion matrix for divisive clustering

Conclusion

In conclusion, reducing the Iris data set to two dimensions in order to then use four clustering algorithms to predict their species has provided me with great insight about unsupervised machine learning methods. By using PCA, I successfully condensed the dataset's features into a lower-dimensional space while retaining essential information, enabling its visualization and interpretation in only two dimensions. Subsequently, the application of K-Means, DBSCAN, agglomerative and divisive clustering allowed us to identify distinct clusters within the dataset, offering insights into its inherent groupings. The accuracy of each of the algorithms was measured using a confusion matrix to compare the groupings from the four algorithms from the actual groupings. It is interesting to note how DBSCAN was the algorithm that performed worse due to its inherent properties of detecting natural groupings by density. The versicolor and virginica iris flowers often had similar properties of sepal and petal length and width that made it hard to tell them apart. Whereas, setosa flowers were way easier to identify by all the clustering algorithms since they had very distinct features from the other two. However, in other settings where the clusters are non-spherical, DBSCAN algorithms can be more effective at discovering clusters in data with

non-globular (e.g., crescent moon, concentric circles) shapes. The other algorithms perform poorly when there are outliers in the data, the clusters are overlapping, or the natural clusters have imbalanced sizes.

Sources References

- **Scikit-learn documentation**:
  - "Scikit-learn: Machine Learning in Python." Scikit-learn Documentation, scikit-learn.org/stable/.
  - Direct link: Scikit-learn Documentation
- **Iris Dataset**:
  - Fisher, R.A. "The Use of Multiple Measurements in Taxonomic Problems." Annals of Eugenics, vol. 7, no. 2, 1936, pp. 179-188. UCI Machine Learning Repository, archive.ics.uci.edu/ml/datasets/iris.
  - Direct link: Iris Dataset at UCI Machine Learning Repository

Link to Code:

Iris.ipynb