Predicting Iris Flower Species based on their Features

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, we intend to predict the species of each iris flower based on their features, and conclude by comparing our results to the actual data. To achieve this, we will use common algorithms in data analysis and machine learning. Firstly, we will reduce the dimensions of the Iris dataset using PCA to remove unnecessary noise. Subsequently, we will apply K-means clustering (also known as Lloyd's algorithm) to categorize each data point to a specific species. Finally, we will compare our results to the real categorization of each of the iris flowers using a confusion matrix to verify the efficacy of our procedure. Ultimately, our goal is to showcase the applications of advanced data that can reveal deep insights from seemingly simple datasets.

Application of PCA

In our analysis, we utilized Principal Component Analysis (PCA) to reduce the dimensionality of the iris to only two columns. The process involved several key steps. Firstly, we 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, we 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, we 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 2 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.

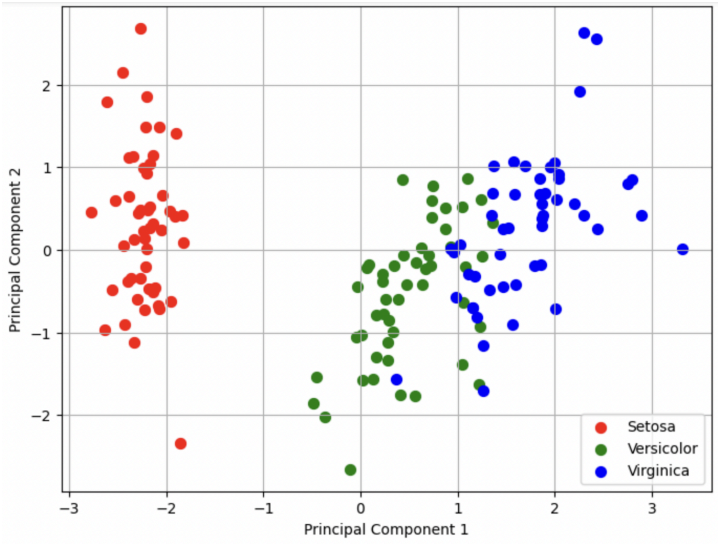

Figure 1

K-Means Clustering to Categorize the Data

Now we will apply the K-Means Clustering algorithm to the graph in Figure 1 to group each iris flower into one of three different clusters. In this case, we will choose K = 3, since we intend our output to predict the three different species of iris flowers that exist in our dataset. Figure 2 below graphically represents the output of the K-Means Clustering algorithm. Note that there are a few instances in which K-Means Clustering algorithm applied to a dataset with reduced dimensions outputs a wrong categorization of the iris flowers, indicating that the procedure is not fully correct.

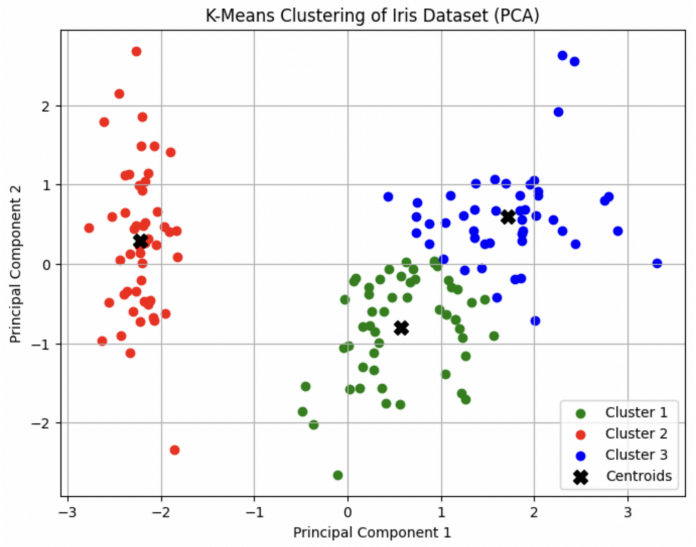

Figure 2

Evaluation using a Confusion Matrix

In order to verify the accuracy of our procedure, we created a confusion matrix that compares our categorization of the iris flowers 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 3 below indicates that our 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, our procedure categorized the iris flowers with an 86.7% 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.

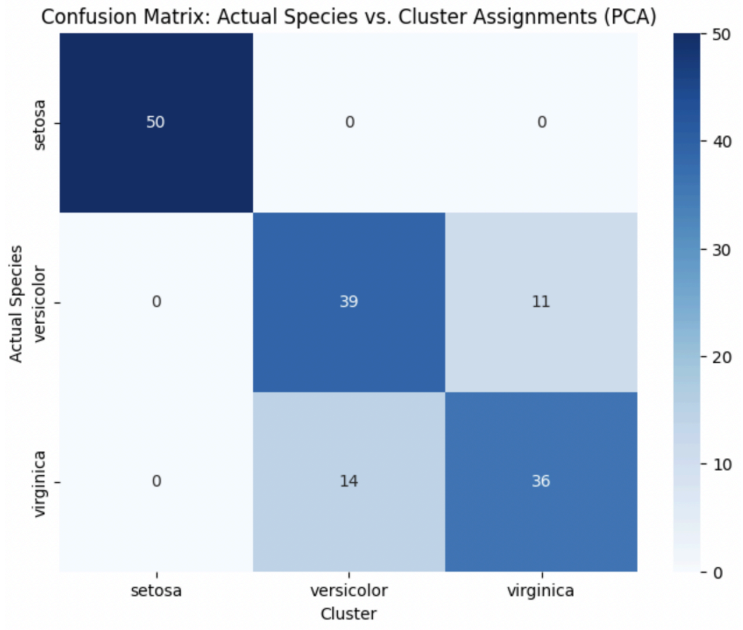

Figure 3

Conclusion

In conclusion, our analysis of the Iris dataset using Principal Component Analysis (PCA) for dimensionality reduction, K-Means clustering for identifying natural groupings, and evaluation with a confusion matrix has provided a comprehensive understanding of the dataset's structure and patterns. By using PCA, we 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 clustering allowed us to identify distinct clusters within the dataset, offering insights into its inherent groupings. Finally, we assessed the accuracy of our clustering algorithm in categorizing iris flowers into their respective species groups using a confusion matrix, which

yielded a satisfactory accuracy of 86.7%. Overall, our analysis showcases the efficacy of advanced data analysis techniques in classifying data.

Source References

- Scikit-learn documentation: *https://scikit-learn.org/stable/*
- Iris Dataset: *https://archive.ics.uci.edu/ml/datasets/iris*

Code

Link to Code