Data Preparation and Clustering Analysis on the Iris Dataset

Name: Farid Hossain, Student id: 23006446, Github repo:
github.com/fariduk/clustering-fitting-23006446

Introduction

This report analyzes the famous Iris dataset containing measurements of sepal length, sepal width, petal length, and petal width for 150 iris flowers from 3 different species (Iris-setosa, Iris-versicolor, Iris-virginica). The goals are to prepare the data, perform clustering using k-means, and evaluate the clustering results visually and quantitatively.

Data Preparation

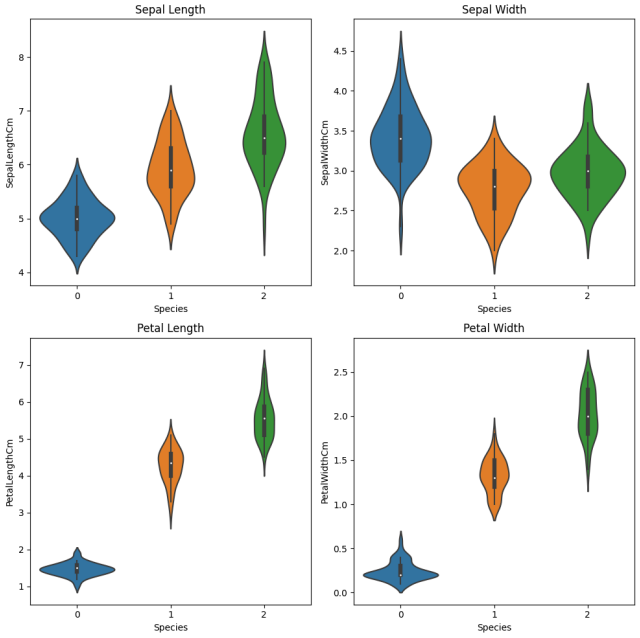
The Iris.csv data file was loaded into a Pandas dataframe. The 'Id' column was dropped as it was not needed for analysis. The 'Species' column containing the string labels was encoded into numerical values using LabelEncoder from scikit-learn.

Clustering Using K-Means

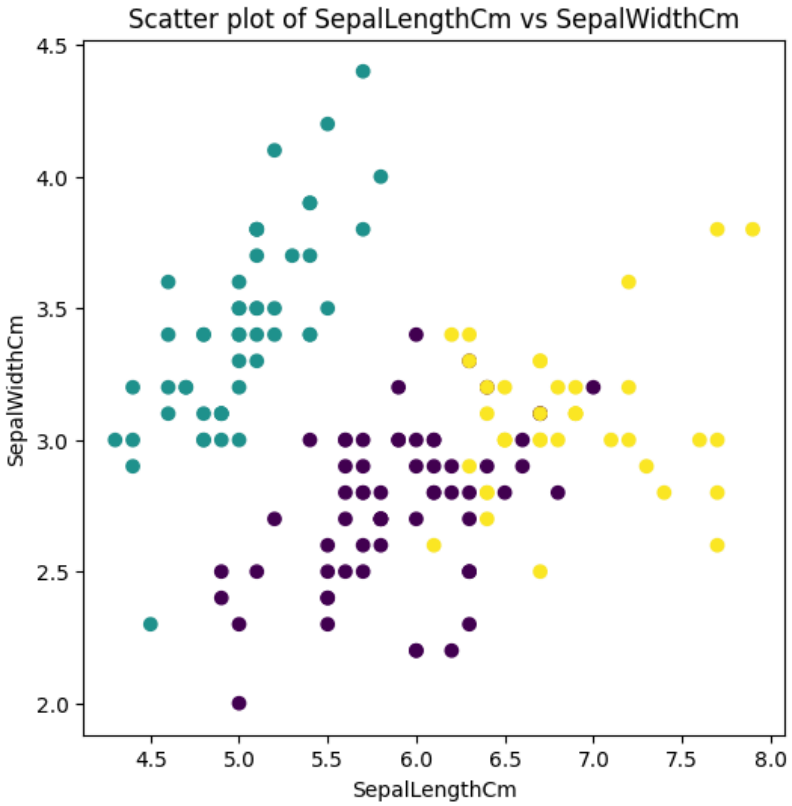
The k-means clustering algorithm from scikit-learn was applied to the numeric flower measurements, specifying 3 clusters since there are 3 known iris species. The cluster assignments were added to the dataframe.

Visual Analysis of Clusters

Violin plots were created using seaborn to visualize the distributions of each feature separated by species and colored by cluster assignment. This allows assessing how well the clusters separate the different species.

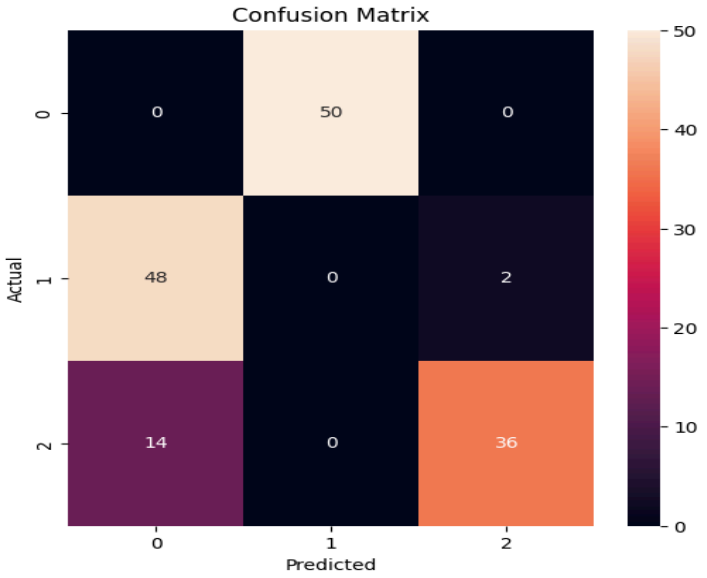


A scatter plot of sepal length vs sepal width was also created, with points colored by cluster, to inspect the spatial separation of the clusters.

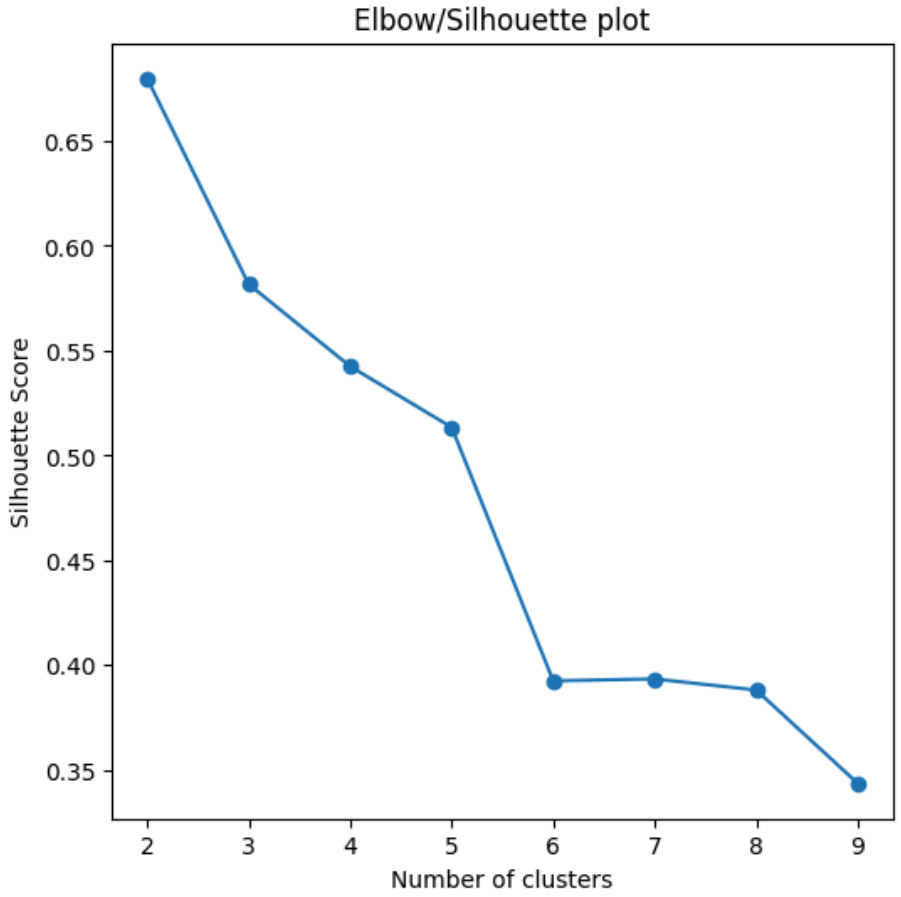


Quantitative Cluster Evaluation

A confusion matrix was computed by comparing the true species labels to the cluster assignments. This confusion matrix was visualized as a heatmap to see which species were commonly misclassified by the clustering algorithm.

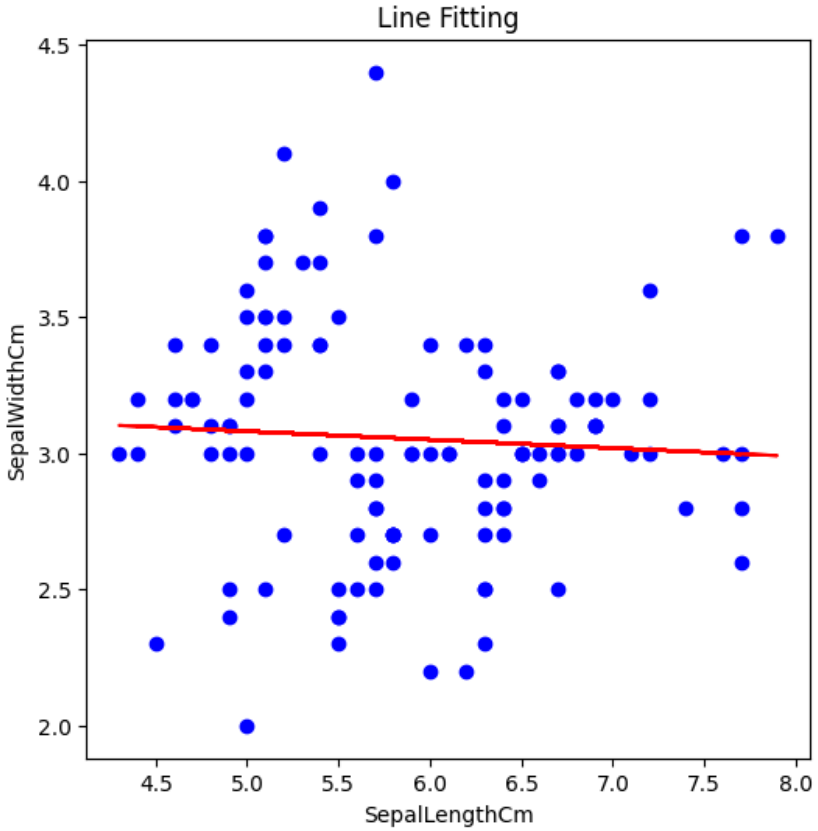


The silhouette score was calculated for different numbers of clusters (ranging from 2 to 9 clusters). This metric measures how well each data point fits into its assigned cluster compared to other clusters. The silhouette scores were plotted to find the optimal number of clusters for this dataset using the elbow method.



Line Fitting Analysis

As a separate analysis, simple linear regression was performed to fit a line to the sepal length and sepal width features. The data was split into training and test sets, a linear regression model was fit on the training data, and the fitted line was plotted against the training data points.



Closing Words

The violin plots showed that the clusters roughly corresponded to the different species, with some overlap between Iris-versicolor and Iris-virginica. The scatter plot also visually depicted the separation of the three clusters. The confusion matrix heatmap indicated that while Iris-setosa was distinct and clustered well, there was some mixing between the Iris-versicolor and Iris-virginica species. The silhouette analysis suggested that 3 clusters was a reasonable choice for this dataset, as adding more clusters did not significantly improve the silhouette score. The line fitting showed a positive linear relationship between sepal length and sepal width on the training data. This report demonstrated data preparation techniques like encoding and train-test splits, as well as applying clustering with k-means and simple regression analysis on the Iris dataset. Visualizations were used to interpret the clustering results and evaluate the appropriate number of clusters. While k-means was reasonably effective, the three species showed some overlap in their measurements.