Report on IRIS Dataset

Snehlata Kumari
Student ID : 23080692
https://github.com/lata1207/ADS1-Clustering-and-Fitting

***Abstract*—In this report, we study the iris data set and explore the data set through K-means clustering and fitting (linear regression techniques) to analyse the future predictions and results.**

I. Introduction

This Iris dataset [1] is consisting of 150 samples from three different species (Setosa, Versicolor, and Virginica)of Iris flowers. Each sample has four features: sepal length, sepal width, petal length, and petal width. It has been selected to visualise the relationships between the features and the species of the flowers, to apply different techniques to cluster the data.

II. Exploring the Datasets

A. Exploratory Data Analysis (EDA)

The statistical summary of characteristics for sepal length [Mean(5.84), Median(5.8), Standard Deviation(0.828), Skewness(0.3140), Kurtosis(-0.552)] indicates that how each feature is distributed.

B. Bar Chart

The bar plot shows the three different species in the Iris dataset: Setosa(purple), Versicolor(green), and Virginica(blue).

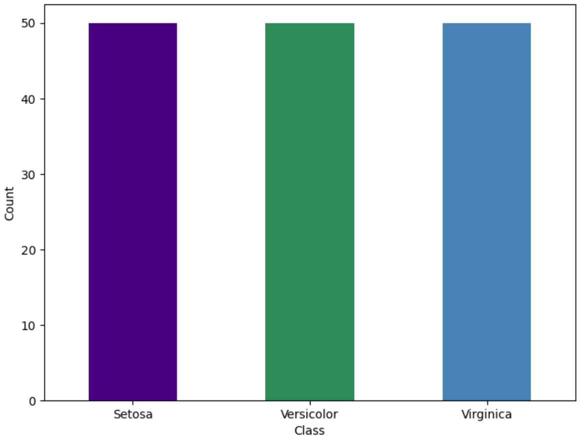

Fig. 1. Bar chart for the class distribution of target variable(species)

Each species has a count of 50 represents the analysis of number sample of each species.

C. Scatter Graph

The scatter graph exhibits all species with distinct clusters of data points, each represented by a different color: blue, orange, and green.
Setosa (blue): This cluster is located at the lower left of the plot, indicating that Setosa is separated from the other species.

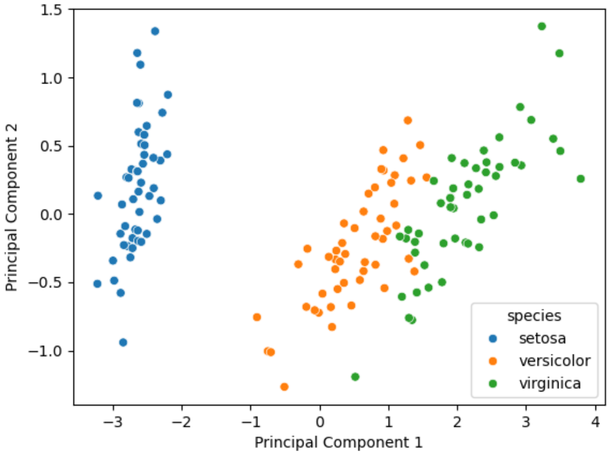

Fig. 2. Scatter plot : All species

Versicolor (orange): This cluster is central in the plot, showcasing some overlap with Virginica, though distinguishable.
Virginica (green): This cluster is positioned at the upper right of the plot, also slightly overlapping with Versicolor, but still distinct.

D. Heatmap

The heatmap visualises the correlation matrix among different features of the species and cluster. The color intensity shows the strength of the correlation, with darker shades indicating stronger relationships.

Sepal Length: There is strongly positively correlated with petal length (0.87) and petal width (0.82) leads to growth of the features, while negatively correlated with sepal width (-0.12) implies that petal length will be less if the sepal width is widely devloped.

Sepal Width: However,the sepal width is negatively correlated with petal length (-0.43) and petal width (-0.37) indicates that if one feature increases, sepal width tends to decrease.
Petal Length: It is strongly positively correlated with petal

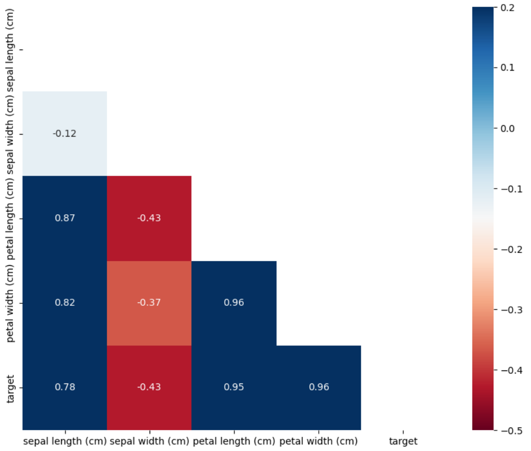

Fig. 3. Heatmap of Correlation matrix(features and cluster)

width (0.96) shows petal dimensions (length and width) are more critical for classification.

E. *Visualize distributions of features*

The distributions of all numerical and categorical features of the Iris dataset are analysed with respect to each feature sepal length and others like cluster and target(species) .

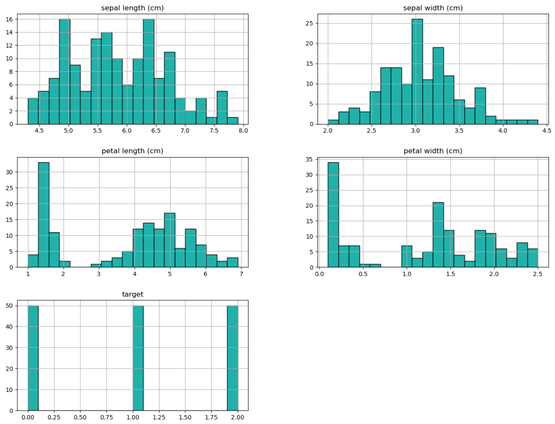

Fig. 4. Visualize distributions of features

The sepal length skews toward petal width, similarly sepal width follows the same trend with target(species).

III. Cluster Analysis

A. *Optimal Number of Clusters [Elbow Method]*

The Elbow Method is used to find the optimal number of clusters.to be viewed graphically,The sum of the distances of each sample to it's assigned cluster centre (called Within-Cluster Sum of Square- WCSS).

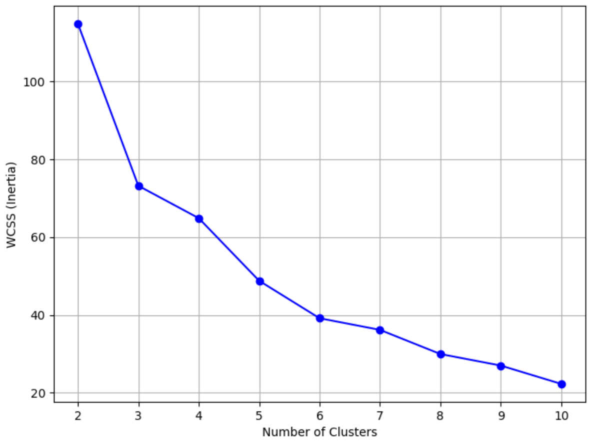

Fig. 5. Elbow Method for Optimal Number of Clusters

Here, we consider the number of cluster is four (k=4)

B. *Cluster Distribution [Pie Chart]*

A pie chart displays the proportion of data points in each cluster.

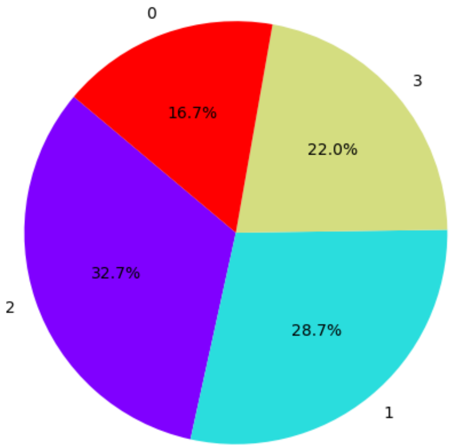

Fig. 6. Pie Chart : Proportion Data Points Cluster

Cluster 0 (red) makes up 14.0% of data points. It suggests that a subset of the data with distinct characteristics.
Cluster 1 (cyan) accounts for 26.7%, indicates a significant

portion of target, containing one of the main species.

Howover, Cluster 2 (purple) is Comprising 32.7%, with the largest cluster. It depicts a group with highly similar features. Lastly, Cluster 3 (Light Green) 26.7% of data points, it is significant in size similar to Cluster 1 and contain another species in the dataset.

C. *Visualising Clusters]*

Cluster 0 (Purple) covers a huge range of sepal lengths approximately 5.5 to 8.0 cm) as petal lengths 4.5 to 7.0 cm. It appears a mix of Versicolor and Virginica species. Cluster 1 (Teal) tightly pack in the lower left corner, with Sepal Lengths ranging from roughly 4.5 to 5.5 cm and Petal Lengths from 1.0 to 2.0 cm. Here in figure 8 we can analyse the Setosa species, known for its smaller sepal and petal measurements.

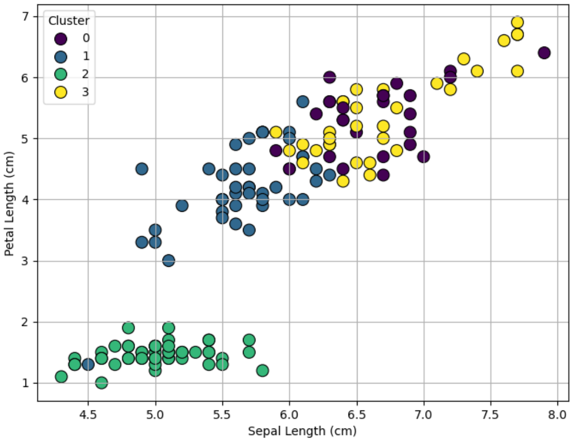

Fig. 7. Visualising Clusters : Clusters Based Sepal Length

Clusters 2 (Blue) with sepal Lengths around 4.5 to 6.5 cm and Petal Lengths from 3.0 to 5.0 cm and cluster 3 (Yellow) displaying overlap of data points, indicating that clusters might represent Versicolor and Virginica species, have similar sepal and petal measurements.

D. *Predicting future Clusters*

Future predictions through K-Means clustering is performed to predict the change in species(setosa, versi and virginica) satisfaction for the iris flower.

Predicted cluster for future data points to analyse the changes in the features of iris flower with all 3 different species.

IV. Regression Analysis

A. *Predicting fitting*

Regression analysis (fitting) is performed to predict the future species(setosa, versi and virginica) satisfaction for the iris flower.

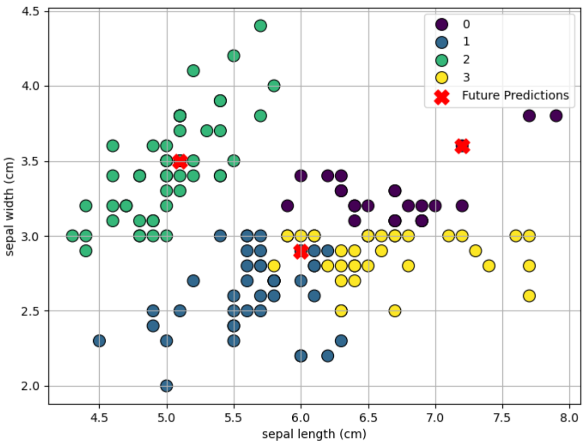

Fig. 8. K-Means Clustering with Future Predictions

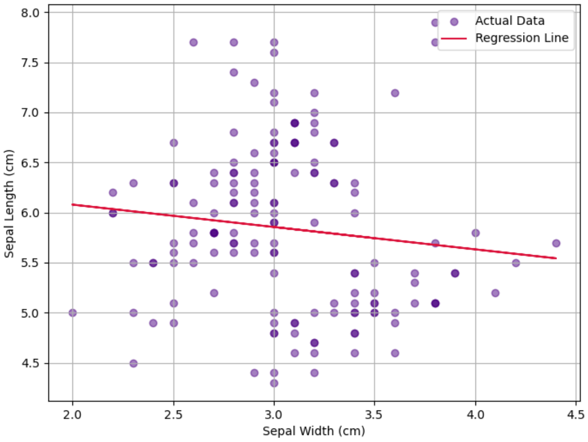

Fig. 9. Linear Regression Sepal Width and Sepal Length

The fitting line shows weak predictability due to stable feature of species.

V. Conclusion

In this work, I proposed this report with a clear view of clustering techniques such as K-means and identified distinct groups within the dataset. It is beneficial for statistical analysis and ML tasks, as it reduces bias and ensures that models do not favor one species over others.

Additionally, the visualisations through pie chart, scatter plot, and heatmap helps to figure out the trends and expectations of dataset. Overall, this analysis highlights the valuable insights of fundamental concepts in data analysis, visualization, and machine learning.

References

[1] Iris Dataset [Kaggle]. https://www.kaggle.com/datasets/saurabh00007/iriscsv