Professor: Claudia Mosquera

Iris Species Clustering Analysis

NRC 407 MACHINE LEARNING
WEEK 4
JOSÉ IGNACIO GUTIÉRREZ GUZMÁN

Introduction

► In this presentation, we will analyze the Iris Species dataset using clustering techniques to identify possible groupings of samples based on their characteristics.

The data set selected was:
https://www.kaggle.com/datasets/uciml/iris

► It includes three Iris species with 50 samples each as well as some properties about each flower. One flower species is linearly separable from the other two, but the other two are not linearly separable from each other.

Introduction

▶ An example of data is:

Id	Sepal Length	Sepal Width	Petal Length	Petal Width	Species
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3.0	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa

Problem

▶ The main objective is to determine if it's possible to cluster the Iris Species samples into distinct groups based on their characteristics using the K-means clustering algorithm.

Development

▶ 1. Data loading and exploration

▶ 2. Data preprocessing

▶ 3. Selection of a clustering algorithm

▶ 4. Application of the clustering algorithm

▶ 5. Visualization of the results

Please check the note book attached to see the complete process if is necessary.

Results

▶ Below is the visualization of the results of the Iris Species clustering using the K-means algorithm.
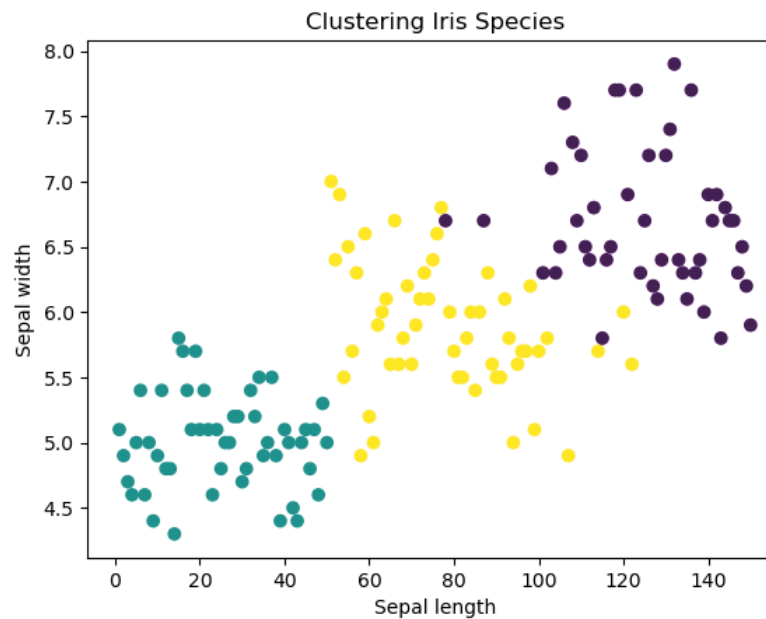
Conclusions

▶ 1. Distinct groups of Iris Species samples were identified using the K-means algorithm.

▶ 2. The inclusion of the 'Species' column in the analysis affected the clustering results.

▶ 3. Further analysis and consideration of alternative clustering algorithms are recommended for a better understanding of the data structure.

Bibliography

▶ UCI Machine Learning Repository. (n.d.). Iris Data Set. Retrieved from https://www.kaggle.com/uciml/iris

▶ "OpenAI. (2021). ChatGPT [Generative language model]. Retrieved from https://openai.com"