Data Mining:

EDA, Supervised, and Unsupervised Learning Techniques (KNN, K-means, and Hierarchical Clustering)

Introduction

The Iris flower dataset is a famous multivariate data set introduced by Ronald Fisher in 1936. It is also known as Anderson's Iris data set since the dataset was collected by Edgar Anderson.

This dataset includes 5 different variables including 4 quantitative and 1 qualitative variable. The dataset represents three flower species: Virginica, Versicolor, and Setosa, which is represented in the categorical variable called "Species". The other 4 quantitative variables are related to the morphologic details of these species including Sepal length, Sepal width, Petal length, and Petal width. The entire dataset consists of 150 observations which are divided into 50 observations for each type of flower.

The goal of this study is to explore the dataset via exploratory data analysis and the incorporation of some machine learning algorithms. Supervised (i.e., k nearest neighbor) and unsupervised learning (k-means clustering and hierarchical clustering) approaches are employed in the classification of the data.

1. Exploratory Data Analysis

In this section, the Iris data is explored and described using summary statistics for the entire dataset and the three flower species.

1.1. Summary Statistics

1.1.1 What Data Types are in the dataset?

The 5 column variables which describe the flower types have different variable types. Of all the five variables, four of them are numeric variable types. Species is a categorical variable with three levels, each representing the three species of Iris flower.

Table 1: Data Type of the Variables

Variable Names	Definition	Variable Type
Sepal.Length	Length of the flower's Sepal	Numeric
Sepal.Width	Width of the flower's Sepal	Numeric
Petal.Length	Length of the flower's Petal	Numeric
Petal.Width	Width of the flower's Petal	Numeric
Species	Type of flower	Categorical/Factor with 3 levels

1.1.2 Are the three flower types equally represented?

The three species of the Iris flower are equally represented in the dataset. That is, there are 50 observations for Setosa, Virginica, and Versicolor, totaling 150 observations in the data.

1.1.3 Five Number Summary for the entire dataset

We further attempt to describe the iris dataset by exploring the five-number summary. Table 2 shows the lowest value, highest value, median, mean, first quartile, and third quartile for all four numeric variables. None of the variables have equal mean and variance, but both statistics are very close for the Sepal variables.

The greatest difference in mean and median is observed in the two Petal variables. Their individual means are lower than their medians, suggesting that the distribution of the petal width is negatively skewed and not symmetric. We hope that the histogram of the variables will shed more light on the shape of the distributions.

Table 2: Five-number summary of the Variables (Entire dataset)

Summary	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Minimum	4.300	2.000	1.000	0.100
1st Quartile	5.100	2.800	1.600	0.300
Median	5.800	3.000	4.350	1.300
Mean	5.843	3.057	3.758	1.199
3rd Quartile	6.400	3.300	5.100	1.800
Maximum	7.900	4.400	6.900	2.500

1.1.4 Are there variations in the variables?

Given that the five-number summary doesn't show the variation in the dataset, we explore the variance across all five features. Petal.Length has the highest range and standard deviation indicating that this variable has the highest variation in the entire dataset, with a minimum petal length of 1cm and a maximum of 6.9 centimeters. On the other hand, the sepal width has the least spread, with a minimum of 2cm and a maximum of 4.4cm. It would be interesting to see the variation of these variables across the three species. This is explored in the next sub-section.

Table 3: Spread across the Variables (Entire dataset)

Summary	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
Standard deviation	0.828	0.436	1.765	0.762
Range	3.600	2.400	5.900	2.400

1.2 Summary Statistics by Species

The standard deviation of the four variables across the 3 species shows that the variation within the individual species is lower than the variation with all the species combined (combined dataset). Similar to the combined dataset, the median and mean of the variables within the three species are different, but very close in all cases. In other words, the difference in the mean and median of the variables are much lower within the flower species than in the entire iris data. This suggests that the distribution of the variables when the individual flower species are considered may be more symmetric than when the distribution of variables for the iris flower as a whole.

Table 4: Summary by Species

Species	*Statistic*	*Sepal.Length*	*Sepal.Width*	*Petal.Length*	*Petal.Width*
Setosa	Mean	5.006	3.428	1.462	0.246
	Median	5.000	3.400	1.500	0.200
	Standard Deviation	0.352	0.379	0.174	0.105
Virginica	Mean	6.588	2.974	5.552	2.026
	Median	6.500	3.000	5.550	2.000
	Standard Deviation	0.636	0.322	0.552	0.275
Versicolor	Mean	5.936	2.770	4.260	1.326
	Median	5.900	2.800	4.350	1.300
	Standard Deviation	0.516	0.314	0.470	0.198

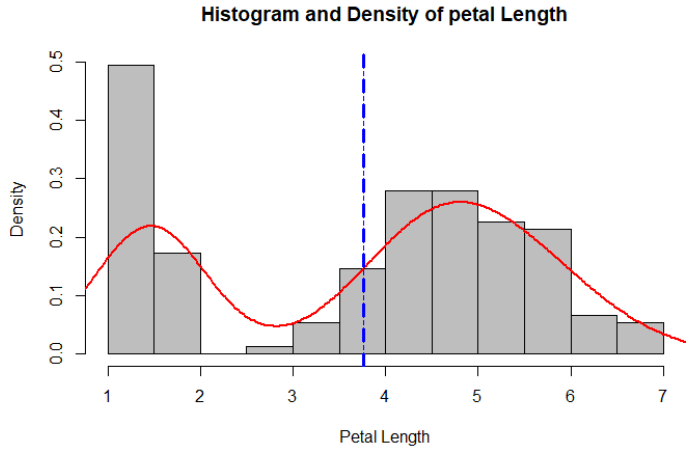
2. Iris Data Visualization

In this section, the Petal.Length variable is selected for further exploration due to the presence of higher variation compared to the other three numeric variables.

2.1 Histogram and Density Plots

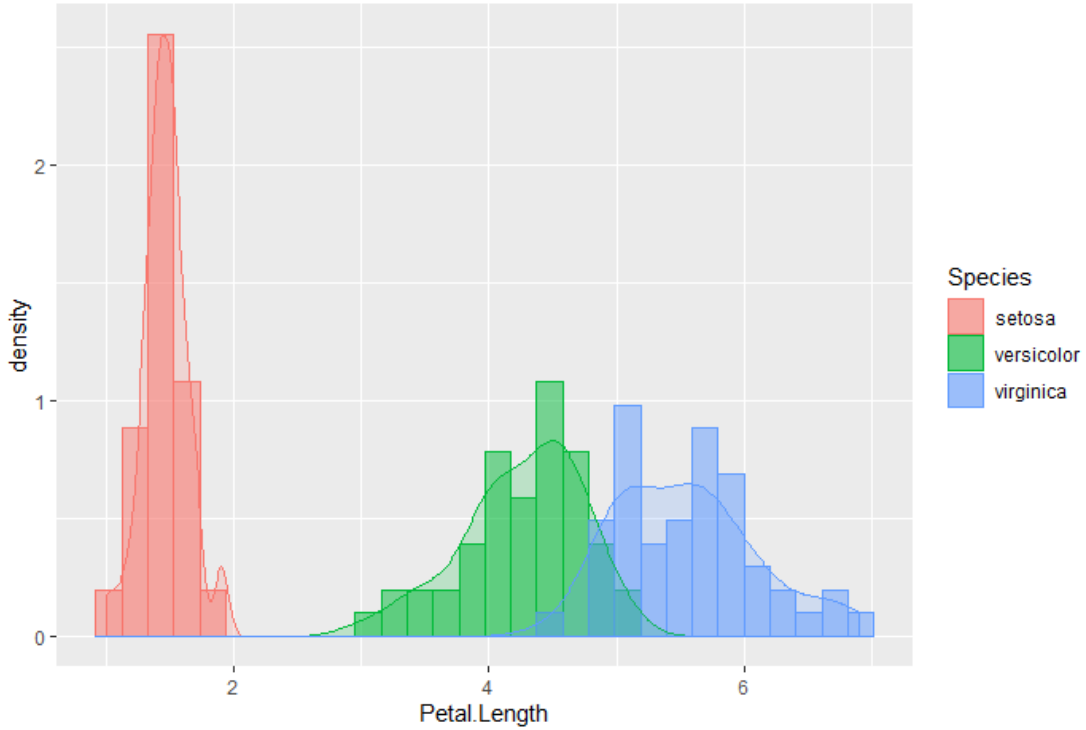
The histogram of the petal length is not symmetric, suggesting that the variable doesn't follow a normal distribution. The blue dotted line shows that the average petal length of the Iris flower is about 3.76 centimeters. As shown in Table 2, this mean is lower than the median indicating that the distribution is skewed. This is also confirmed in the histogram below.

Figure 1: Histogram and Density Plot of Petal.Length



The plot of the Petal_length by species is constructed to inspect the distribution and to ascertain if the species differ or are similar based on the length of their petals. Figure 2 suggests that based on petal length, Setosa is different from the two other species, while Versicolor and Virginica are more similar given that their petal lengths overlap.

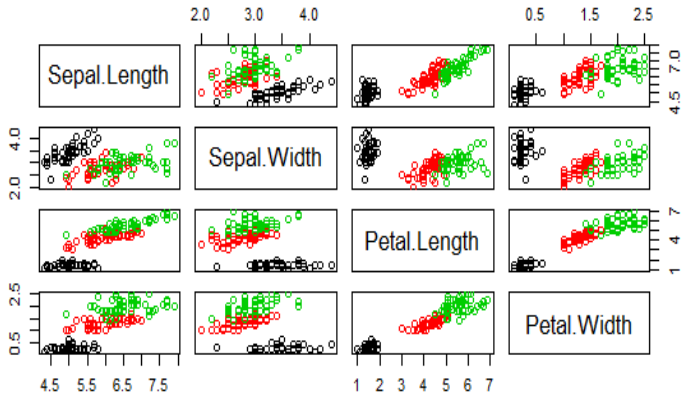
Figure 2: Histogram and Density Plot of Petal.Length by Species



2.2 Scatterplot

The scatterplot matrix is useful in examining the bivariate relationships among the variables. The scatterplot matrix below shows the pairwise correlation between the four numeric variables. Figure 3 shows that the strongest positive linear relationship appears between Petal length and the Petal width. A positive linear relationship is also observed in the following pairs: Sepal.Length and Petal.Width, Petal.Length and Sepal.Length. It is hard to make conclusions about the relationship between Sepal.Length and Sepal.Width, Sepal.Width and Petal.Length, and Sepal.Width and Petal.Width. The scatterplot also shows the presence of 3 species/groups in the dataset. In addition, the colors suggest that one group is quite distinct from the other two species.

Figure 3: Matrix Scatterplot of all the variables



3. Supervised Learning: KNN

The K nearest neighbor algorithm is used for the purpose of classification. The model is trained using 80 percent of the dataset and the prediction accuracy of the model is evaluated using the testing data which is the remaining 20 percent of the dataset. We also explore different values of k and assess the prediction accuracy.

When k=5, we observe a 6.7% misclassification rate and 2 misclassifications, whereas when k=10. The misclassification rate decreased to 3 with 1 misclassification. On the other hand, when k=100, the misclassification rate increased to 30% with 9 misclassifications. We suspect that this is because larger values of k make boundaries between classes less distinct.

Smaller values of K such as k=5 and k=10 result in more flexible fits, with low bias but high variance. On the other hand, larger values of K as k=100 will have smoother decision boundaries which means lower variance but increased bias. However, a higher K averages more data in each prediction and is more resilient to outliers.

Table 5: Prediction Accuracy for K=5

	Predicted		
True	Setosa	Versicolor	Virginica
Setosa	8	0	0
Versicolor	0	9	2
Virginica	0	1	11
No. of misclassification=2; Misclassification rate = 0.067			

Table 6: Prediction Accuracy for K=10

	Predicted		
True	Setosa	Versicolor	Virginica
Setosa	8	0	0
Versicolor	0	10	1
Virginica	0	0	11
No. of misclassification= 1 Misclassification rate = 0.033			

Table7: Prediction Accuracy for K=100

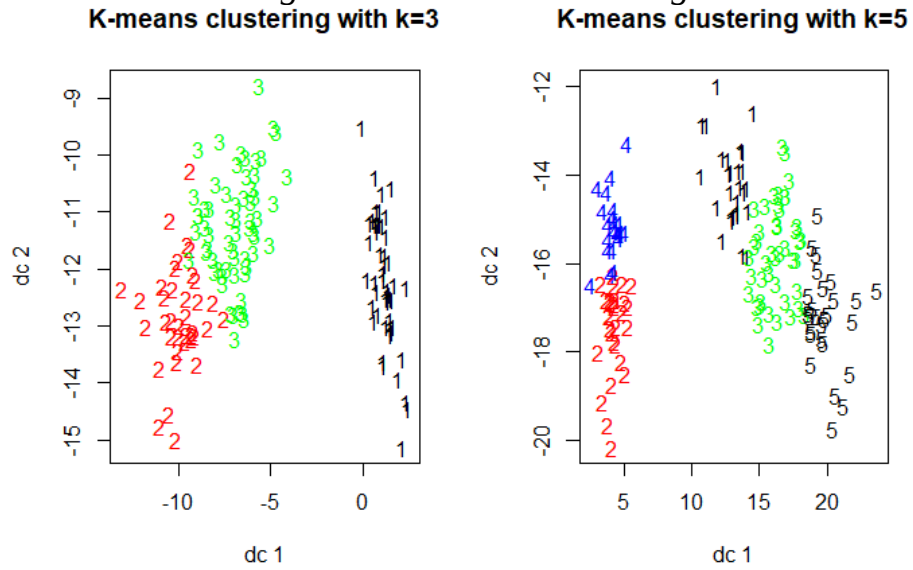
	Predicted		
TRUE	Setosa	Versicolor	Virginica
Setosa	8	0	0
Versicolor	1	6	4
Virginica	0	4	7
No. of misclassification=9; Misclassification rate = 0.3			

4. Unsupervised Learning

4.1. K-means clustering

The K-means algorithm is used to partition the Iris dataset into k clusters as shown in Figure 4. We explore different values of k. For k=3, cluster 1 is very distinct from the other two clusters. We suspect that cluster 1 represents Setosa which is different from the other two species as indicated in the data visualization. When k=5, there's more overlap between the clusters. But there's still an obvious distinction as the 5 clusters seem to be split into two groups. Clusters 1 and 4 are more similar, while clusters 2,3, and 5 are more alike.

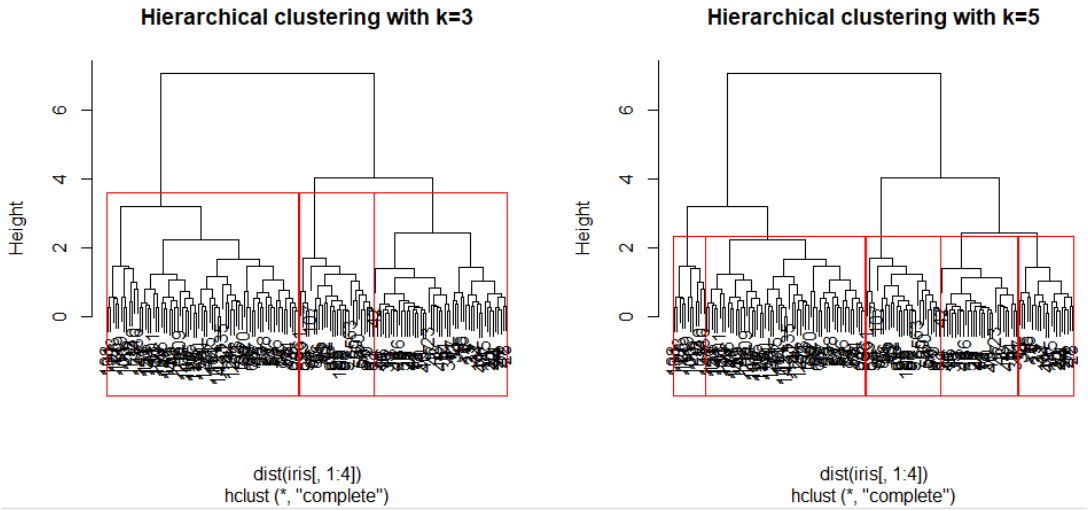
Figure 4: K-means clustering



4.2. Hierarchical Clustering

Hierarchical clustering, which is a method of performing clustering analysis is also considered, with the same values of k as in the K-means clustering. The dendrogram is consistent with the K-means clustering. It is seen that for k=3, one of the clusters is more distinct than the other two. More so, when k=5, the clusters are split into 2 groups based on similarity.

Figure 5: Hierarchical clustering (Dendrogram)



Conclusion

Of all the three flower species represented in the Iris flower dataset, Setosa is quite distinct, while Versicolor and Virginica are more similar. This explains why the variation of the variables within the individual species is lower than the variation with all the species combined (combined dataset).

The distinction of the Setosa from the other two species is also shown in the output from our clustering analysis using the K-means algorithm and hierarchical clustering technique.
In addition, our findings from the K-NN algorithm show that smaller values of k result in more flexible fits, with low bias but high variance, while larger values of k have lower variance but increased bias.