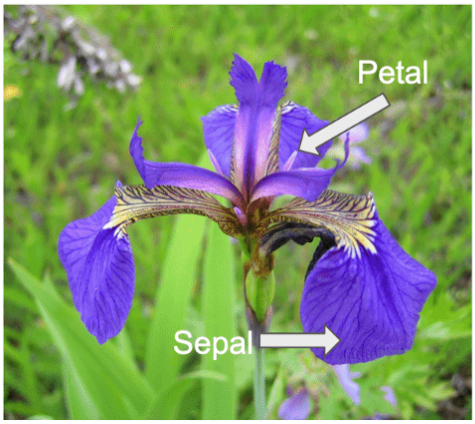
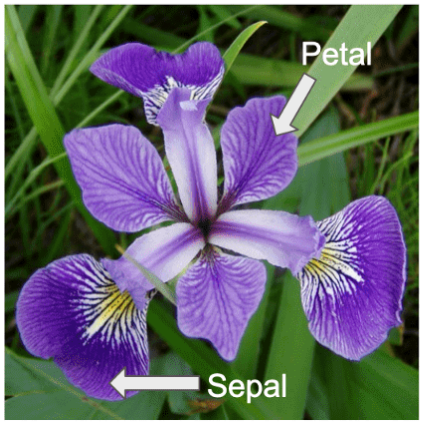
HW #0: Exploring the Iris Dataset

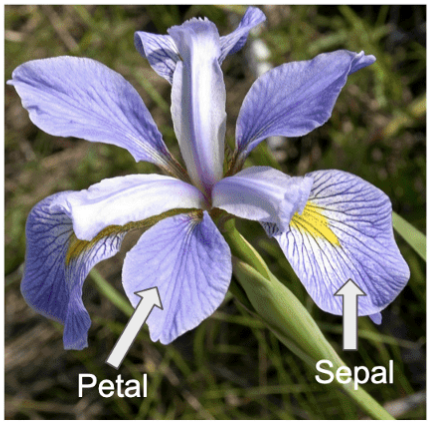
Iris setosa



Iris versicolor



Iris virginica



Fakharyar Khan
CS 5785: Applied Machine Learning
Professor Kuleshov

In this problem set, I explored the Iris dataset by generating scatterplots for each pair of features. First, I loaded the Iris dataset by reading the CSV file line by line and storing the features in a 150x4 NumPy array. I chose to store it in a NumPy array over a 2D list because it offers more efficient indexing and a cleaner approach when plotting pairwise scatterplots, as it allows for easy column indexing. The class labels for each sample were stored in a separate list.

To generate the scatterplots, I first mapped each class label to a distinct color. This helped distinguish the different Iris species in the scatterplots and allowed us to visually examine how the features relate to each other across classes. I then created a 4x4 grid of subplots and iterated through every pairwise combination of features, generating scatterplots for each pair. If a pair consisted of the same feature (i.e., along the diagonal of the grid), I displayed the feature name instead of generating a scatter plot. Otherwise, I produced a scatterplot comparing the two features.

The results can be seen in the figure below. One of the most notable observations is the clear separation of the Iris-setosa class from the other two classes in nearly every scatterplot. For example, in the scatterplot comparing petal length against petal width, the Iris-setosa samples are clustered near the origin, while the other two species have longer and wider petals. Although there is some distinction between Iris-versicolor and Iris-virginica, there is noticeable overlap between them in many of the scatterplots, making it harder to classify them perfectly.

However, based on the scatterplot, we can create a simple rule: flowers with petal widths less than 0.9 centimeters are classified as Iris-setosa, those with petal widths between 0.9 and 1.7 centimeters as Iris-versicolor, and the rest as Iris-virginica. This rule would misclassify only about six samples, giving a 96% accuracy. To verify this, I applied the rule to each sample and compared the predicted labels with the actual labels, confirming the 96% accuracy. While more complex rules could improve the accuracy by incorporating additional features, this risks overfitting the model to the specific dataset.

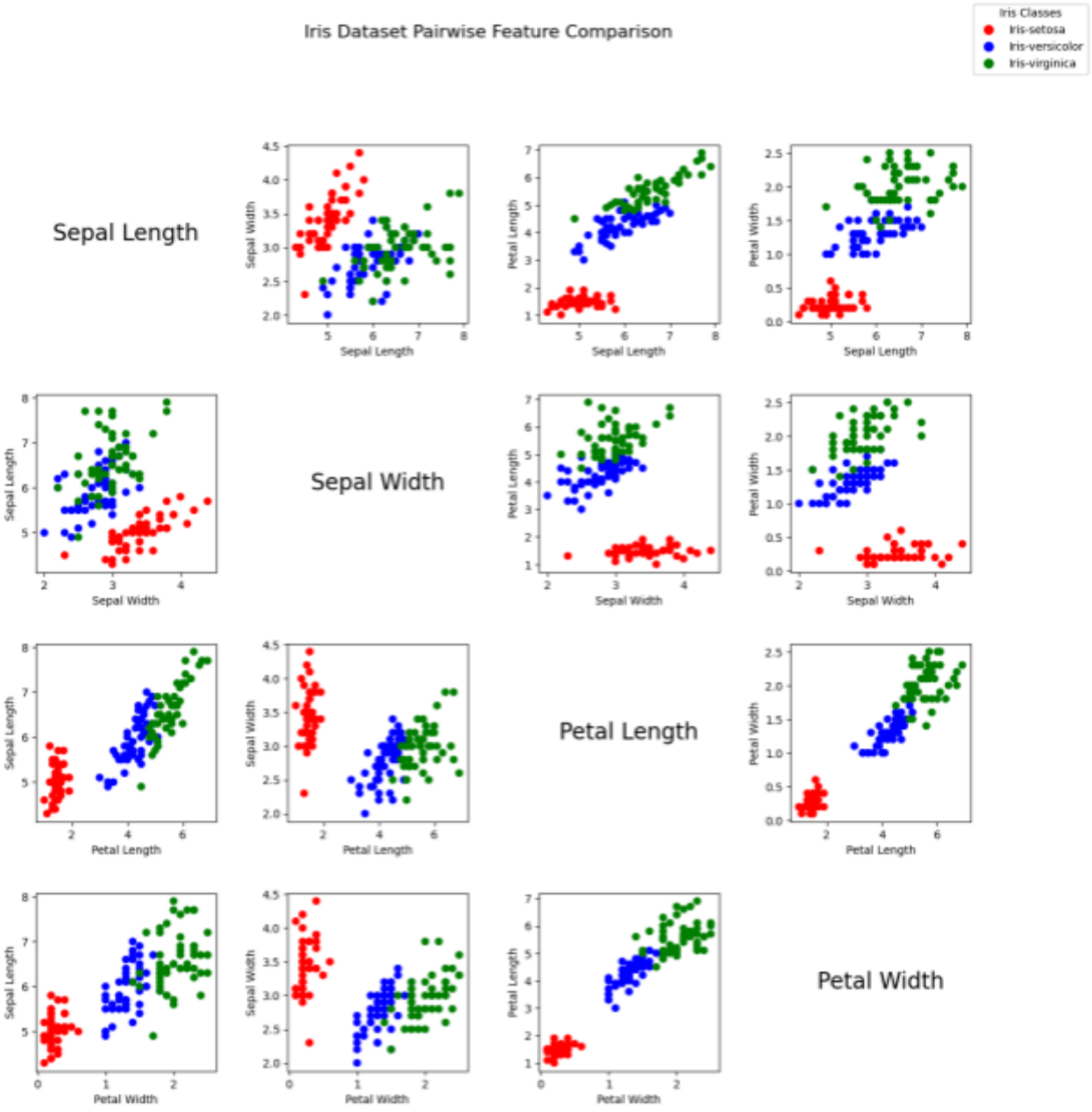


Figure 1: A pairwise feature scatterplot of each pair of features in the Iris dataset.