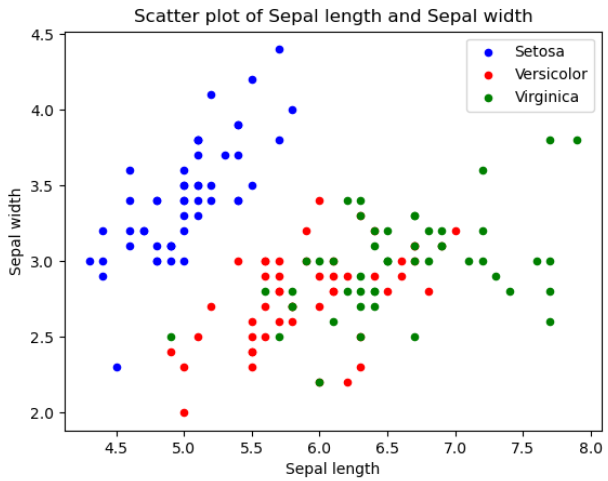
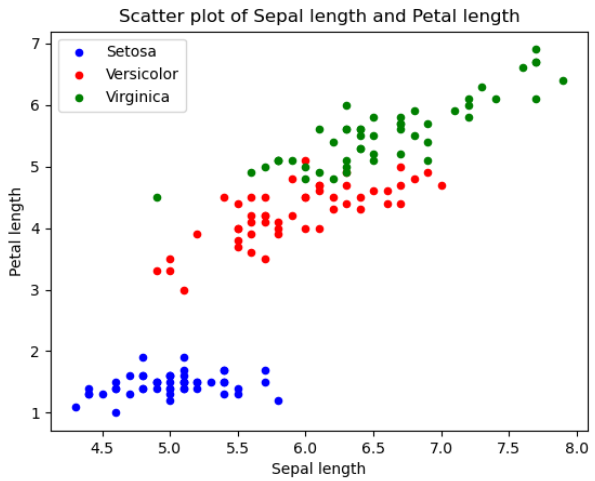
Machine Learning　HW1

312512005 黃名諄

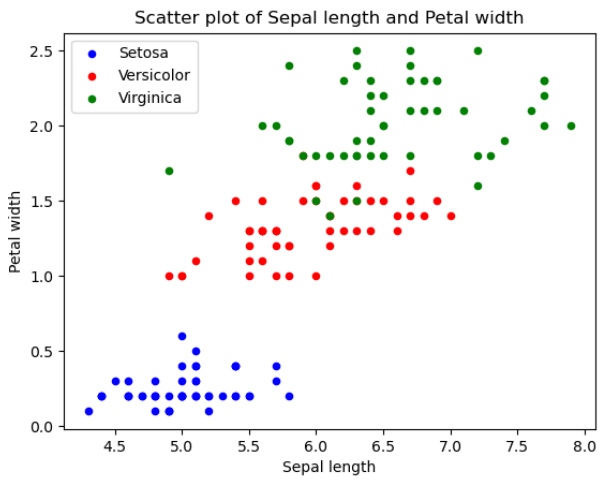
1. Scatter plot :



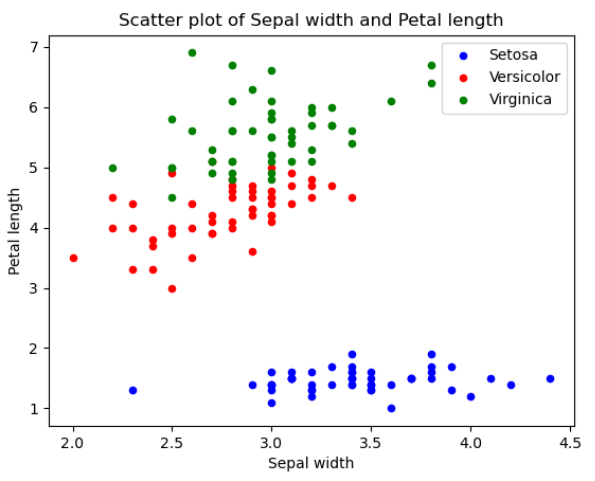
Sepal length v.s. Sepal width



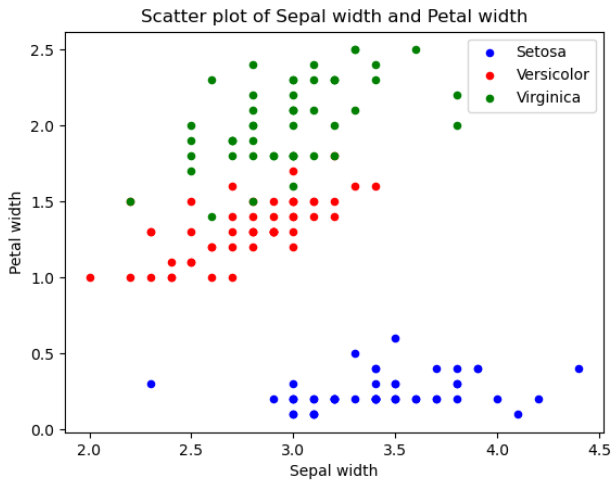
Sepal length v.s. Petal length



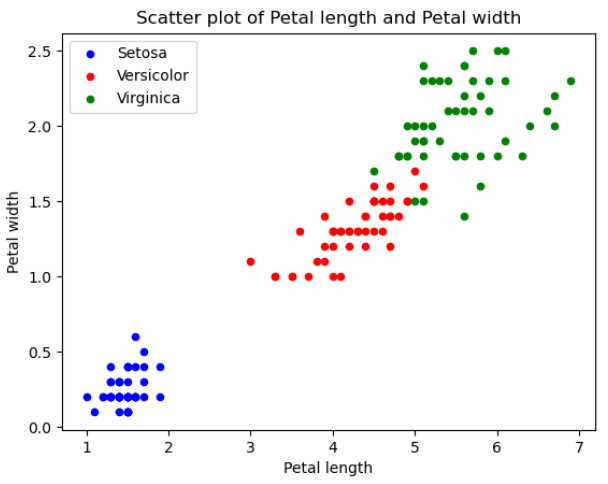
Sepal length v.s. Petal width



Sepal width v.s. Petal length



Sepal width v.s. Petal width



Petal length v.s. Petal width

2. Classification Rate(CR):

- K=1

k=1

Feature Combinations	CR(%)
Sepal length	59.33
Sepal width	48.67
Petal length	91.33
Petal width	91.33
Sepal length + Sepal width	70.67
Sepal length + Petal length	92.67
Sepal length + Petal width	88.67
Sepal width + Petal length	92.00
Sepal width + Petal width	93.33
Petal length + Petal width	95.33
Sepal length + Sepal width + Petal length	92.67
Sepal length + Sepal width + Petal width	92.67
Sepal length + Petal length + Petal width	94.67
Sepal width + Petal length + Petal width	96.67
Sepal length + Sepal width + Petal length + Petal width	94.67

- K=3

k=3

Feature Combinations	CR(%)
Sepal length	61.33
Sepal width	52.00
Petal length	92.67
Petal width	96.00
Sepal length + Sepal width	75.33
Sepal length + Petal length	92.67
Sepal length + Petal width	94.00
Sepal width + Petal length	92.00
Sepal width + Petal width	95.33
Petal length + Petal width	95.33
Sepal length + Sepal width + Petal length	92.67
Sepal length + Sepal width + Petal width	90.67
Sepal length + Petal length + Petal width	95.33
Sepal width + Petal length + Petal width	97.33
Sepal length + Sepal width + Petal length + Petal width	94.00

3. 結果討論:

- 散佈圖使以兩兩特徵組合繪製，因此可對應到 CR table 中使用兩種特徵去做 knn 分類的情況來做比較。
- 在 k=1 及 k=3 使用兩種特徵的情況下，Sepal length + Sepal width 的分類率都是 6 種中最低的，只有 70.67 %(k=1)及 75.33 %(k=3)，在分佈圖 Sepal length v.s. Sepal width 中也能看出，Versicolor 和 Virginica 的分佈是混雜在一起的，只有 Setosa 是明顯分開的，因此用此兩種特徵去做 knn 分類，只能有效分出 Setosa，對於其他兩類則難以區分，造成分類率低的問題
- 其他 5 種組合下，分類率都不錯，從散佈圖來看，也可發現 3 種類別的劃分區域較 Sepal length + Sepal width 來的明顯，和較高的分類率互相吻合，而當中又以 Setosa 分布區域最獨立，可預期此品種的分類預測會最正確，Versicolor 和 Virginica 的邊界還是有一些混雜，因此這兩類可預期會有預測錯誤發生，而這也是分類率無法達到 100%的主要原因
- 另一方面，因 k=3 使用了更多點去做分類上的預測，等於多了幾個能提供分類依據的點，比起 k=1 多了些許驗證性，其分類率可能較高的結果也是可預期的，在此 6 種情況中，k=3 的分類率都

大於等於 k=1 的情況。

- 在 k=1 或 k=3 中，可發現單獨使用 Sepal length 和 Sepal width 的分類率都非常低，因此此兩特徵各自在 3 品種中可能較沒有明顯差異能用來分類。
- 所有特徵組合中，不論在 k=1 或 k=3 中，都可發現用 Sepal width + Petal length +Petal width 的情況能有最高的分類率，可得到結論，使用此 3 種特徵可能能夠最有效的區分這 3 個品種。
- 在實作過程中，有發現在單獨使用 Sepal length 和 Sepal width 特徵時，點間距離會常有一樣的情況(也對應到低分類率的事實)，這種情況下，因浮點數精度問題，選最近點的結果依據不同的算法可能會有些許不同，和同學討論後，我認為這是導致大家的 CR 有些許細微差異的原因。我的邏輯是先依距離從小到大排列(使用 argsort)，遇到相同距離時也會依照原順序排列(照類別 1、2、3 的順序)，再根據 k 值選前 k 個項。

4. 心得:

這次作業讓我能實際去撰寫 knn 分類器，比起過去大學部的課上都只有使用開源套件來操作，此次實作能讓我更深刻的將上課學到的邏輯概念和自己的程式碼連結起來，knn 的邏輯在上課時聽起來不難，透過作業能更了解實作上架構的考量及不同情況或參數下分類能力的表現，也透過作業更熟悉像 numpy、pyplot 等函式功能使用，對於未來作業肯定有所幫助。