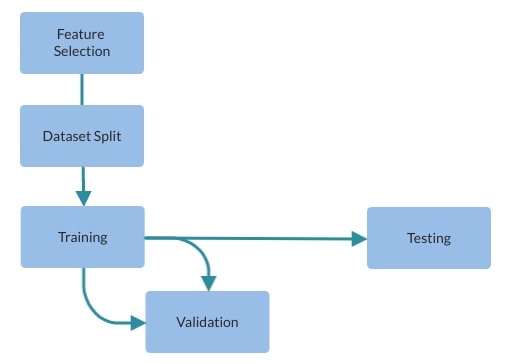
**使用K-Nearest Neighbors實現鳶尾花分類**

**邱品諺 B083040029**

**演算法說明：**

K-Nearest Neighbors演算法主要利用**距離計算**來找出與測試點最接近的k個鄰居，並在這k個鄰居中找出哪種種類最多，以決定測試點之種類為何。本程式中利用歐幾里德距離作為距離運算的方式，首先一筆一筆地將測試資料和所有訓練資料的各個維度做距離運算，並依照距離由小到大、由近到遠進行排序並記錄其對應之訓練資料類別，最後擷取前k筆距離並將數量最多的類別作為該測試資料的類別，以此方式將所有測試資料的類別算出，最後再和原始的ground truth進行比較，以計算準確率。

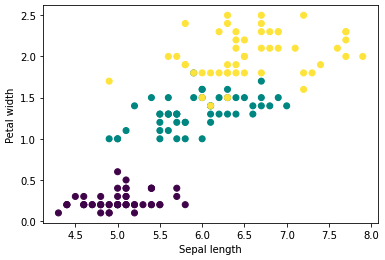
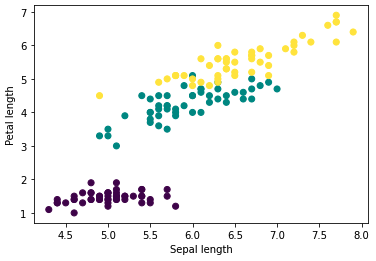
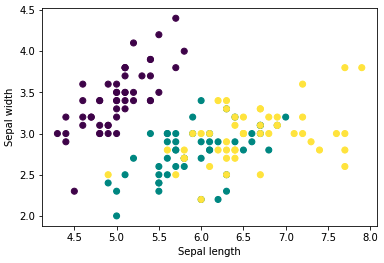
**程式執行：**

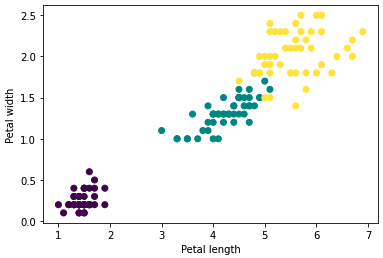
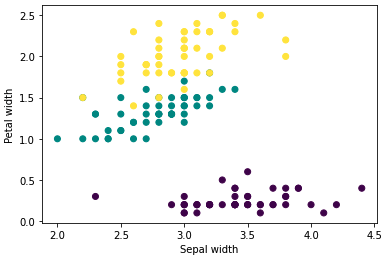
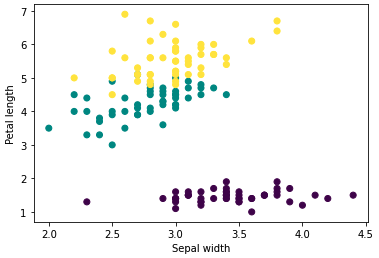


程式執行步驟

1. Feature Selection

選取方式為從鳶尾花資料集提供的4種特徵（Sepal length、Sepal width、Petal length、Petal width）中選取其中2種進行訓練及測試，因此有六種資料集組合。



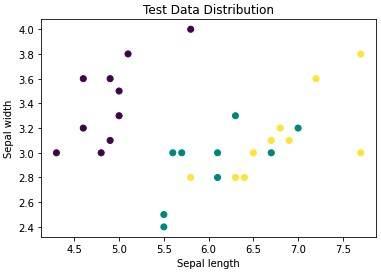


1. Dataset Split

先將資料集以隨機抽取的方式分割為Training Set及Testing Set（80%、20%），再將Training Set透過K-Fold Cross Validation的方式分為Training Set及Validation Set（分割為4個Fold）（60%、20%），分割比例如下圖：

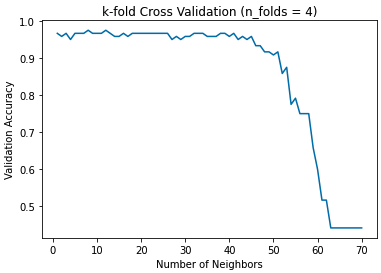
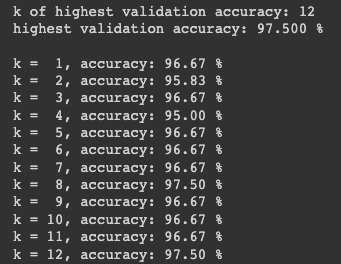


此外，由於資料集的分割是採取隨機分配，因此可能會有分配不均的問題，所以在每次分割時都會顯示測試資料的分佈，使用者可執行該區塊以取得理想的分割。



1. Training and Validation

透過K-Fold Cross Validation的方式每次抽取一個Fold作為Validation Set，將其餘k-1個Fold合併作為Training Set，來回訓練k次，並將k次準確率取平均即可得知在指定鄰居數量下之平均準確率。而本階段的目的為找出具有最高平均準確率的鄰居數，因此程式將根據使用者指定的鄰居數，尋找範圍內具有最高平均準確率之鄰居數，以作為進入Testing階段的鄰居數。另外，除了以數字顯示指定鄰居下的平均準確率，並以折線圖顯示各個鄰居數量下之準確率情形，如下圖，圖為鄰居數為12時有最高平均準確率。

1. Testing

最後利用上個階段所訓練得到的最佳鄰居數執行本階段，而且只能執行一次以獲得最終準確率。另外，本階段繪製了以測試資料為資料點的分佈分界圖並標示各個資料點的類別。如下圖，下圖為鄰居數為8時，以Sepal width及Petal width作為特徵的Testing結果。

