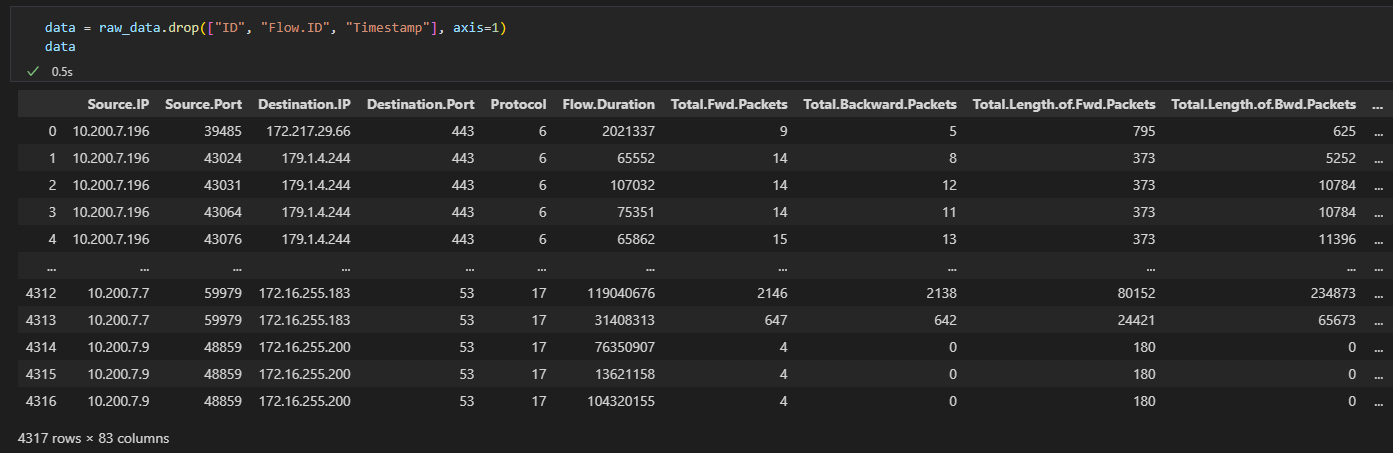
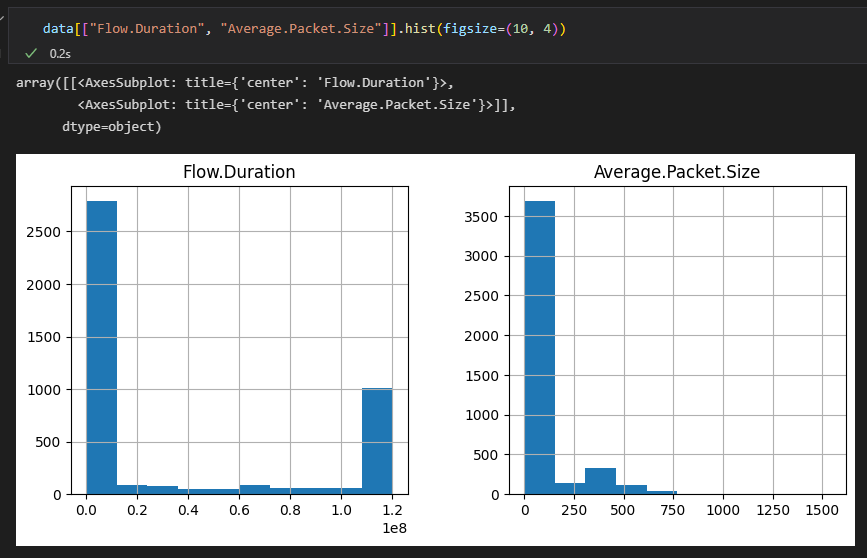
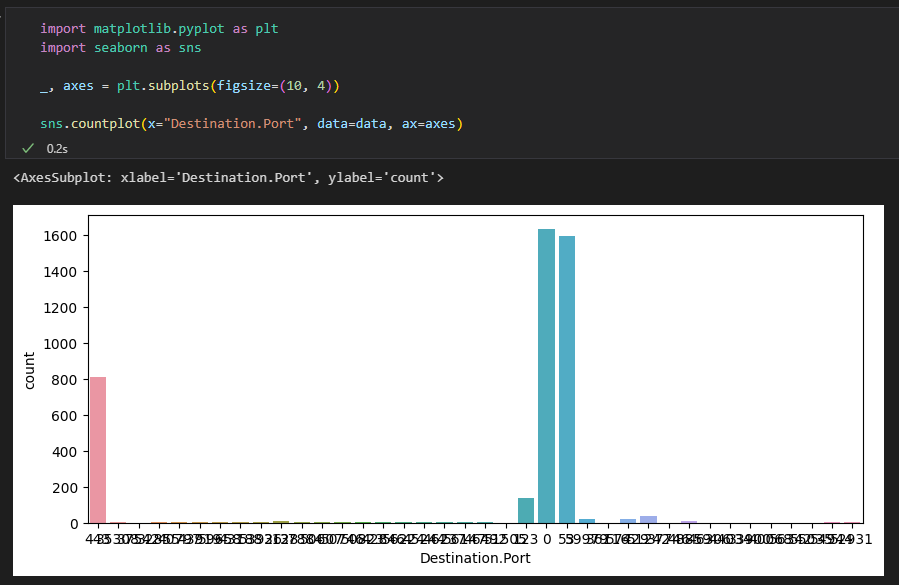
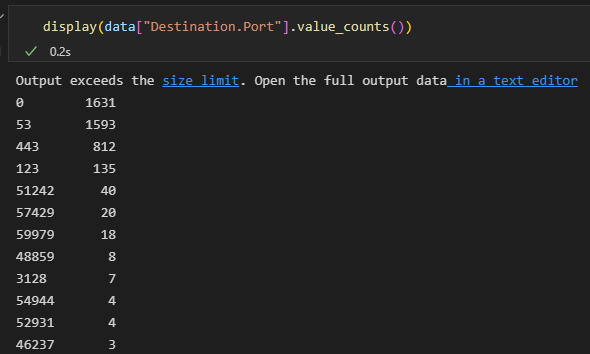
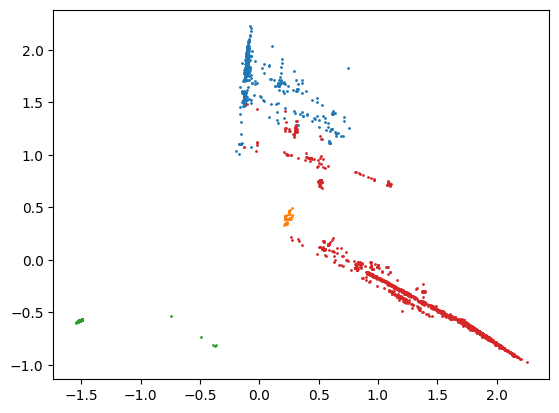
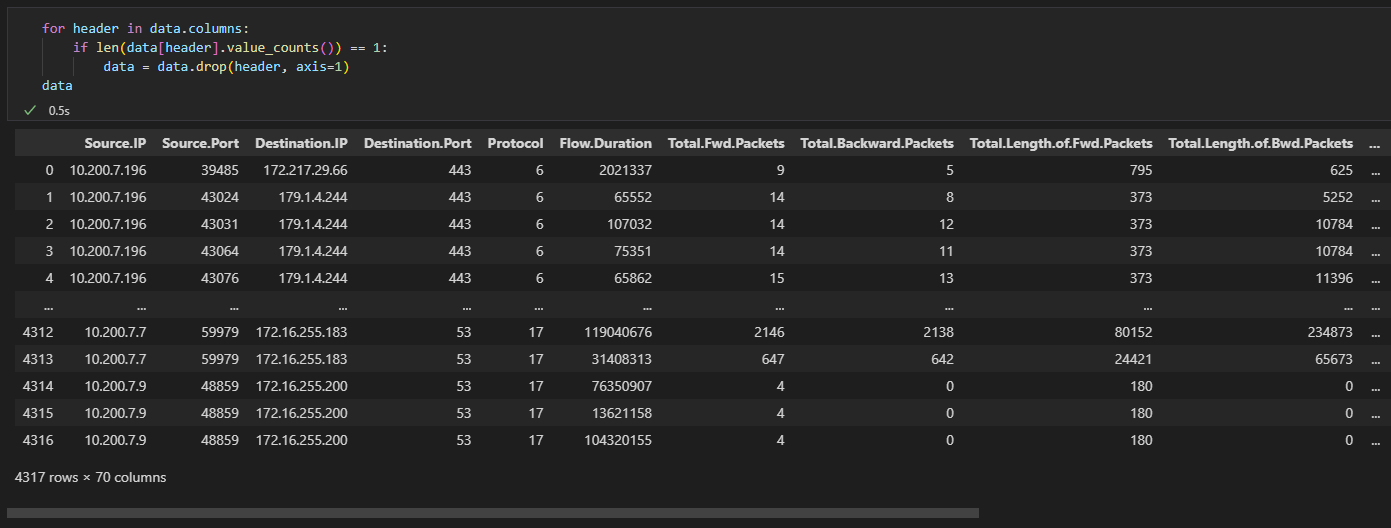
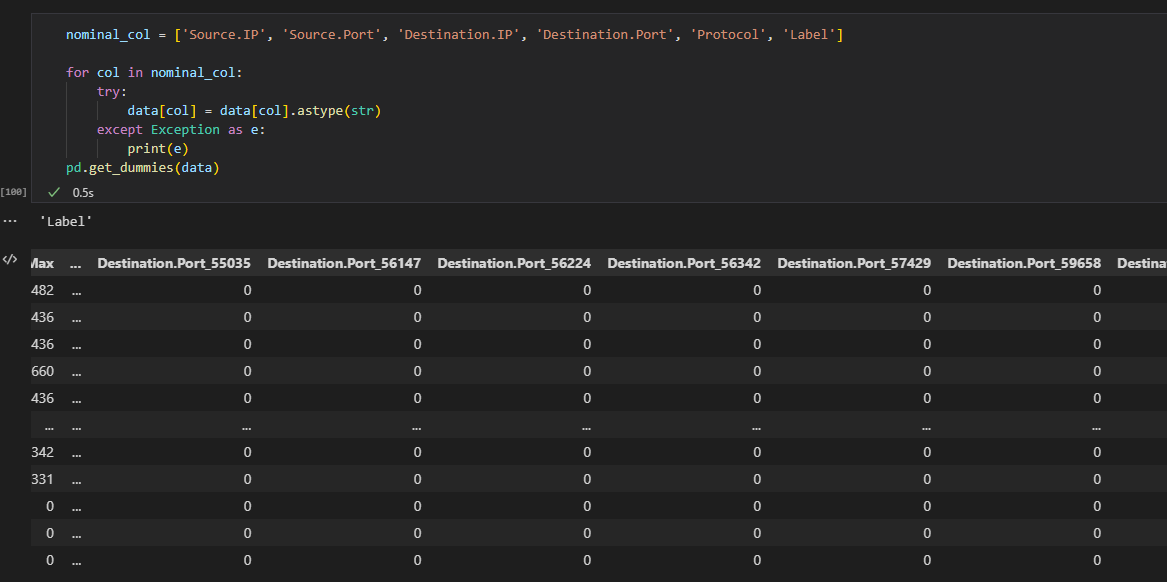
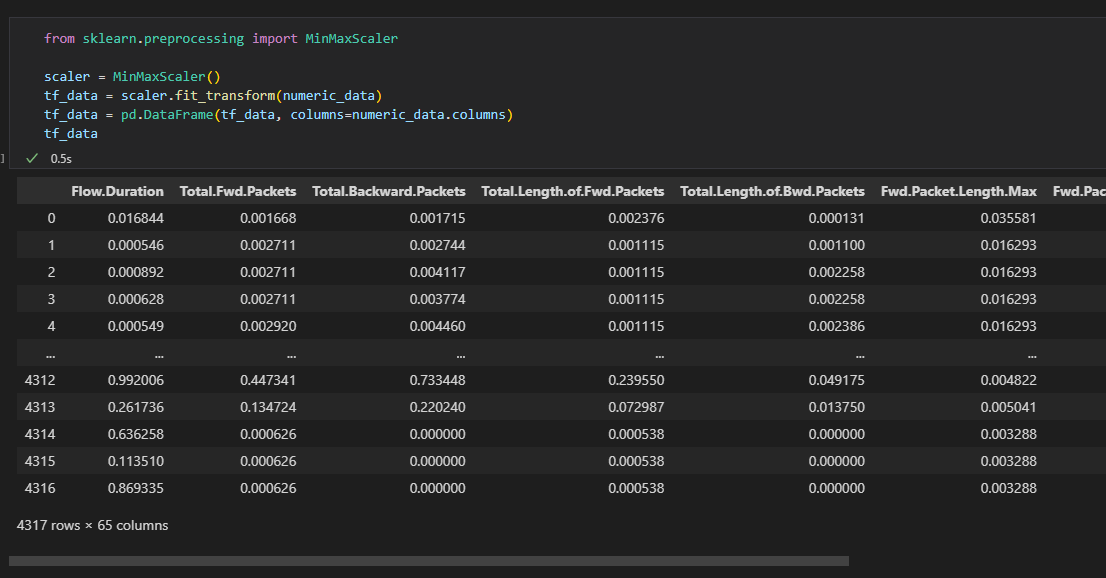
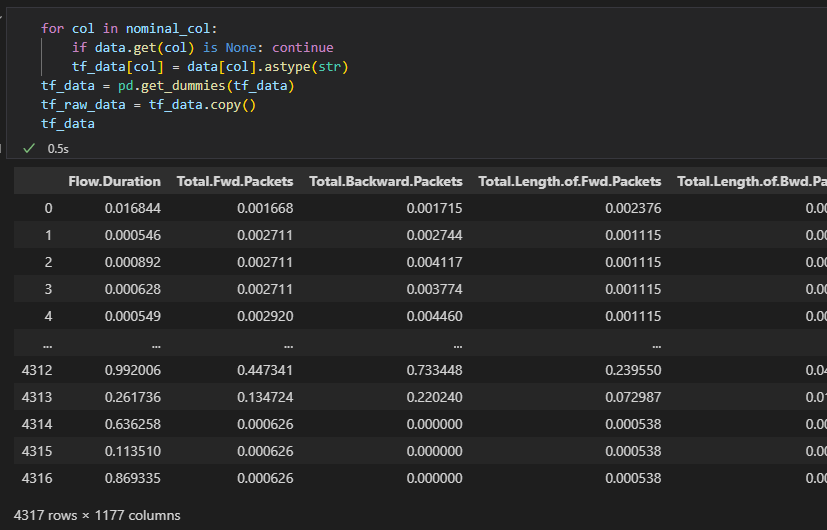
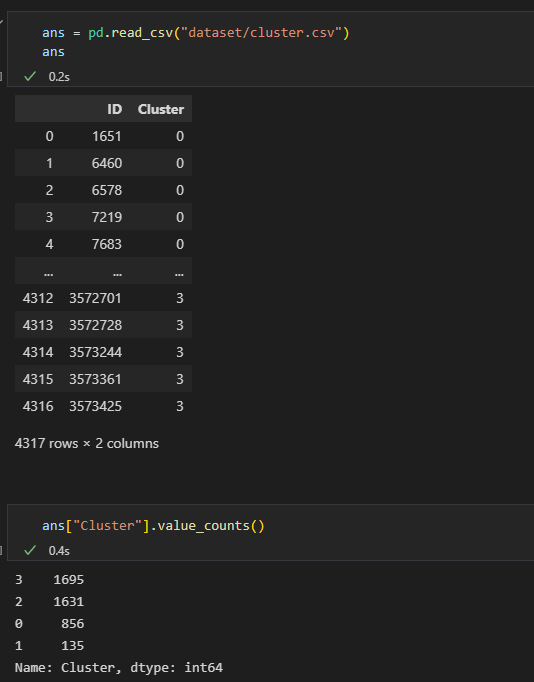
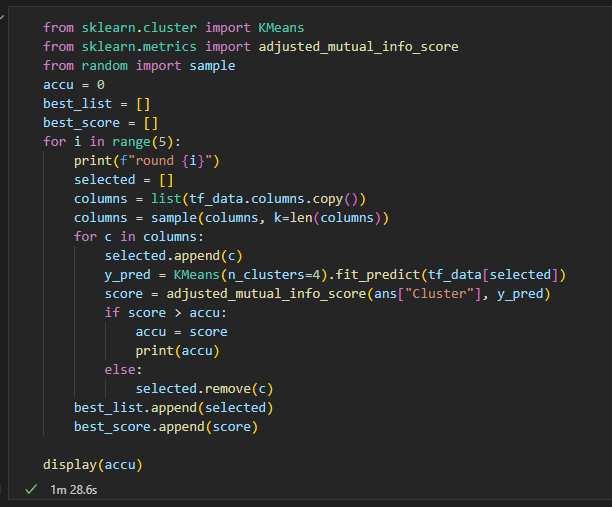
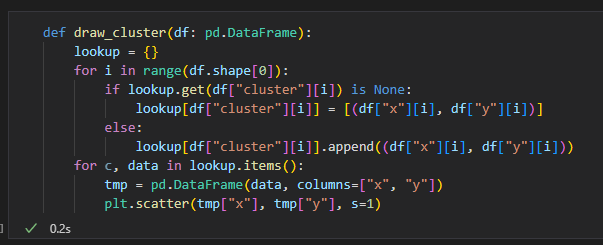
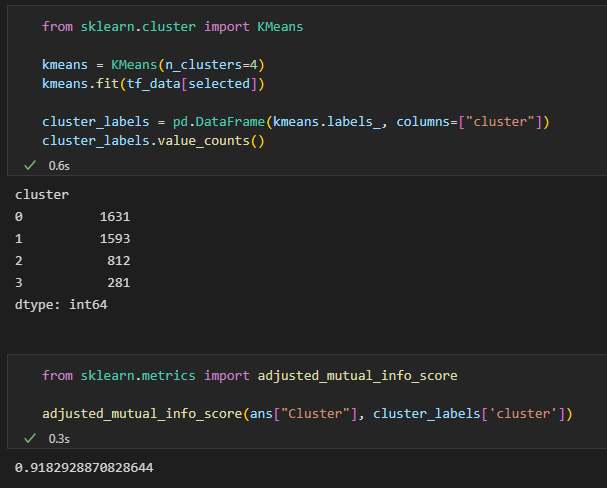
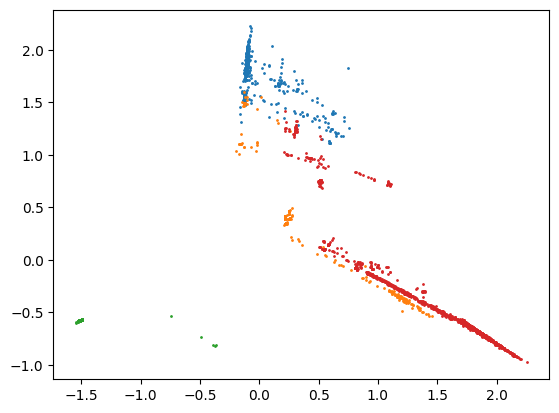
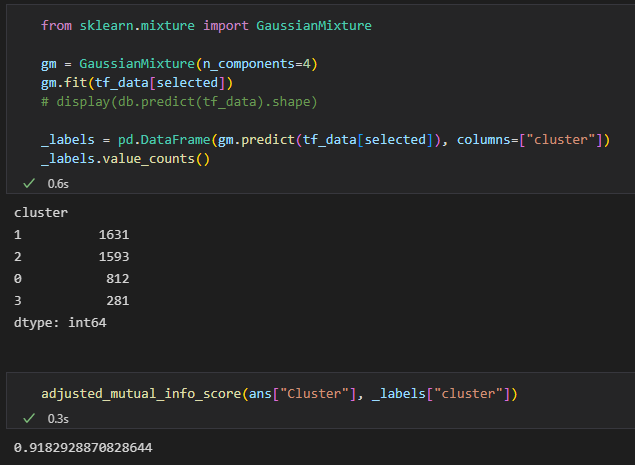
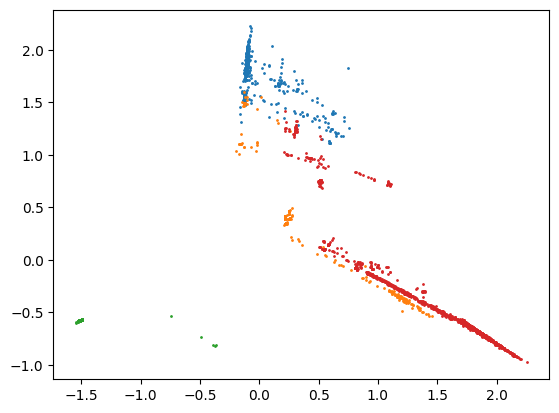
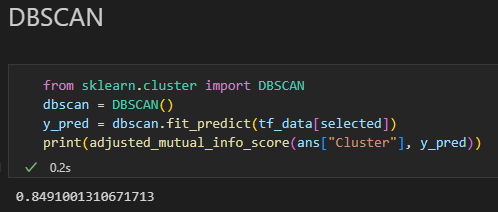
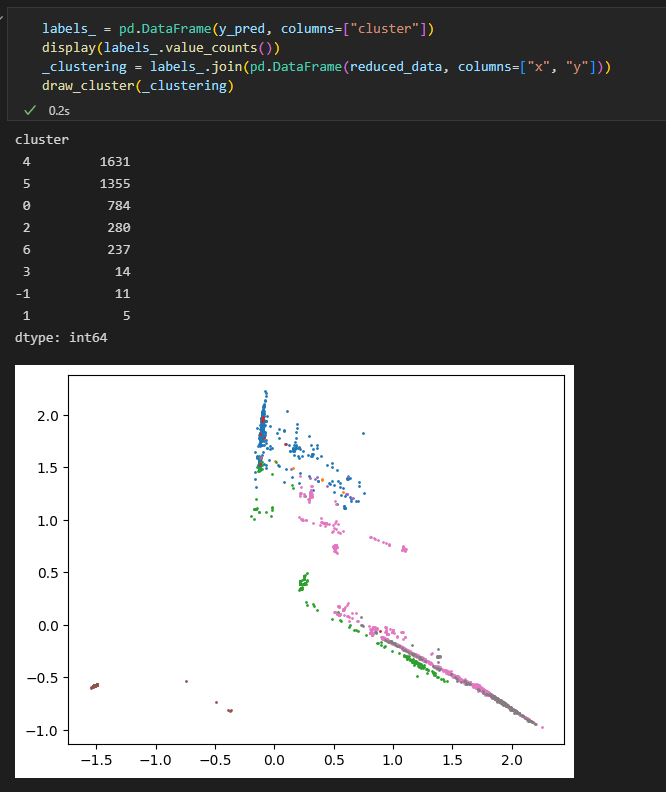
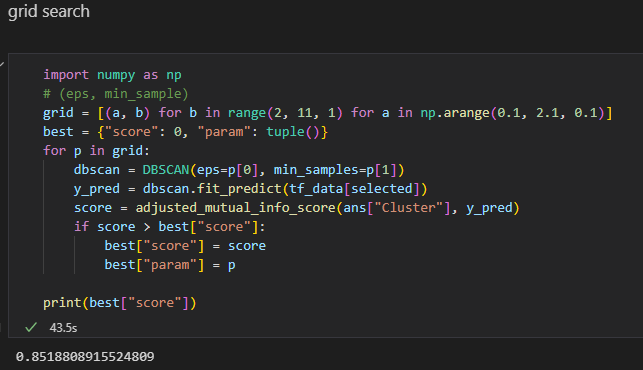
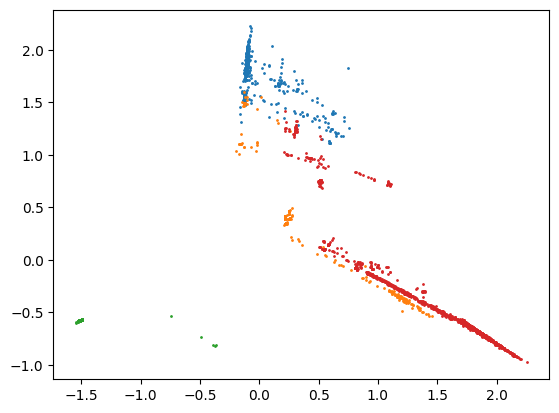
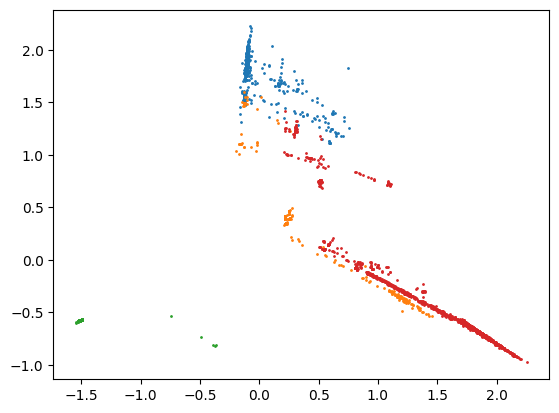
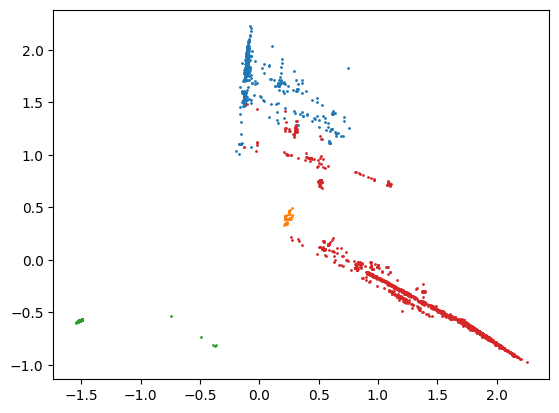
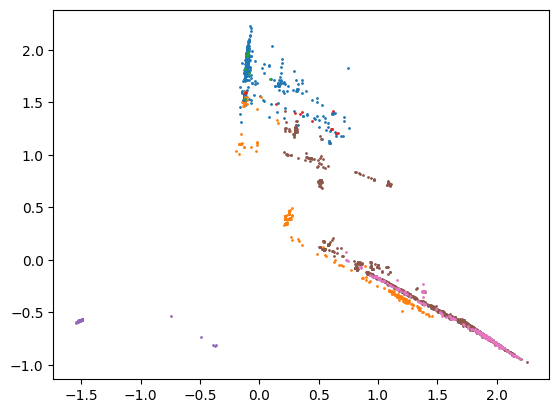
MLFN Lab2 Report

資科工一 游庭瑋

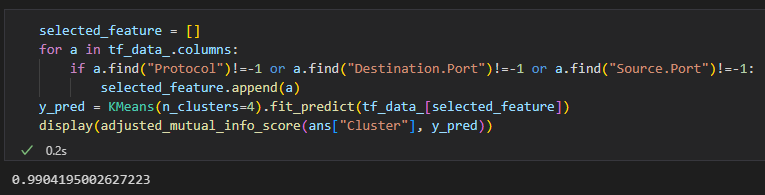
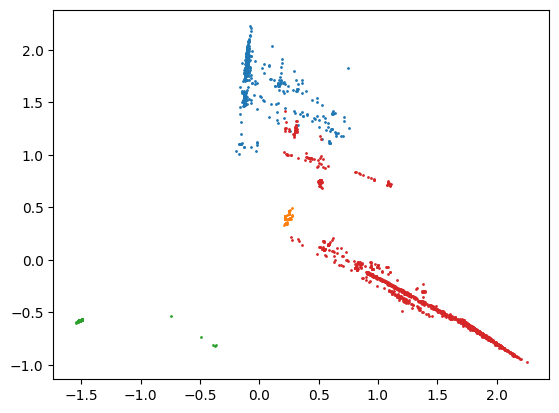
1. Data processing:  
   首先使用pandas讀取csv資料集，會發現有部分欄位為ID等不重複資料，因此必須先將其丟棄  
   
2. Data visualization:  
   這裡選用一些比較顯著的特徵來視覺化  
     
   從Flow Duration可以看出，絕大部分的Flow時間佔用都是比較少的，而從average packet size也可以看到，絕大多數的封包都小於250bytes  
     
   從Destination Port這邊則可以看出，大部分的Flow也都集中在特定的幾個Port上。並從value\_counts函數可以看出，這些Flow主要的port如下:  
     
   另外，就Protocol來統計則有以下結果  
     
   最後將所有numerical attribute正規化並將nominal attribute做one-hot-encoding後，使用PCA分析，將維度降至二維後比較cluster.csv中的正確答案可以得到下圖:  
     
   因此可以看出，此問題是可以被妥善分群的。
3. Feature engineering  
   在前面讀取資料時有發現到，有部分的attribute對於整個資料集的數據都相同，因此將有符合此特性的欄位皆刪除。留下共70個欄位。  
   移除IP、Label等跟分群無關的資料後，對所有非數值資料做one-hot-encding，目前共有1177個欄位。  
     
   接著對丟棄所有非數值資料做Min Max Scaling，方便後面分群的距離計算  
     
   將以上正規化後的數值欄位及one hot encoding完的非數值欄位合併，即得到能夠直接分群用的資料集  
   同時匯入cluster.csv作為參考  
     
   這裡使用Iterative feature selection，每次加入一個特徵，並重新測量分群的準確度。最後得出使用14個欄位即可達到91%的準確度。  
     
   
4. Clustering:  
   先定義好用來繪製分群結果的函數  
   
   1. KMeans  
        
      分群結果及視覺化  
      
   2. Gaussian Mixture  
        
      分析結果與視覺化  
      
   3. DBSCAN  
        
      分析結果與視覺化  
        
      可以看到，DBSCAN由於不須指定n\_cluster，會有多分群的狀況，造成結果不盡理想，在稍後會繼續做參數調整。
5. Parameter adjustment  
   這裡直接調整DBSCAN的參數，定義eps介於[0.1, 2.0]，min\_samples [2, 10]   
   搜尋最佳結果為0.85，參數為eps=0.7, min\_samples=4  
   DBSCAN的eps參數為資料點間的距離Threshold，因此若eps太小時，可能會造成分出的群數太多，出現分數較低的狀況。
6. 分群視覺化  
   這裡直接比較三種分群方法及原本的資料

Raw data

DBSCAN

GaussianMixture

KMeans

1. By domain knowledge  
   以Flow的分群來說，一般我們會最注重協定本身，這裡的協定泛指傳輸層協定及應用層協定。因為此資料集沒有提供應用層協定的資訊，所以這裡直接使用Source.Port及Destination.Port來判斷。  
   因此共有3種attribute可以使用，做完one-hot-encoding共923個欄位，並取得分群結果如下:  
     
     
   可以看到，直接以這三個attribute做分析即可得到0.99的分數，視覺化的結果也與原本的分群幾乎無異。
2. 結論與討論:  
   在這次的實驗中，我們發現:與分類一樣，分群依賴著大量的資料前處理過程，甚至因為沒有Label的關係，造成Feature selection的階段更難取得我們想要的特徵，這時如果就直接將所有資料丟到分群演算法中進行分群，就會造成不重要的特徵去影響重要的特徵，進而造成分群結果不理想。  
   由前面的Feature selection階段可以發現，在這次的分群的資料集中，只需要極少的特徵就可以達成較高的分群準度，並在後面的Domain knowledge階段發現，直接使用Protocol, Source Port, Destination Port就可以達到99%的準度，說明了如果有Domain knowledge的幫助，可以更加方便的直接以人工提取feature，並達到較高的準確度。