111356032 | 王楚雲 | 資管碩一

第一題

1. Meaning and range

Audio feature	Meaning	Range
Danceability	Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity.	Float: 0.0 to 1.0 A value of 0.0 is least danceable and 1.0 is most danceable.
Energy	Energy represents a perceptual measure of **intensity and activity.** Typically, energetic tracks feel fast, loud, and noisy. For example, death metal has high energy, while a Bach prelude scores low on the scale. Perceptual features contributing to this attribute include dynamic range, perceived loudness, timbre, onset rate, and general entropy.	Float: 0.0 to 1.0
Key(調性)	The key the track is in. Integers map to pitches using standard Pitch Class notation. E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on. If no key was detected, the value is -1.	Integer: >= -1 & <= 11
Loudness	The **overall loudness of a track in decibels (dB)**. Loudness values are **averaged across the entire track** and are useful for comparing relative loudness of tracks. Loudness is the quality of a sound that is the primary psychological correlate of physical strength (amplitude). Values typically range between -60 and 0 db.	Float: 0.0 to -60;
Mode(大/小調)	Mode indicates the **modality (major or minor)** of a track, the type of scale from which its melodic content is derived. Major is represented by 1 and minor is 0.	Integer: 0 or 1
Speechiness	Speechiness detects the **presence of spoken word**s in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value.	Float: 0.0 to 1.0; Values above 0.66 describe tracks that are probably made entirely of spoken words. Values between 0.33 and 0.66 describe tracks that may contain both

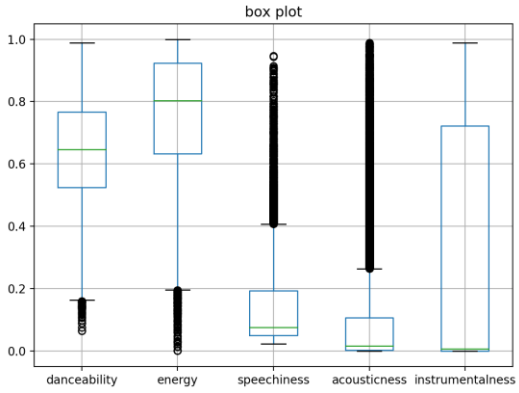
		music and speech, either in sections or layered, including such cases as rap music. Values below 0.33 most likely represent music and other non-speech-like tracks.
Acousticness	A confidence measure of whether the track is acoustic.	Float: 0.0 to 1.0; 1.0 represents high confidence the track is acoustic.
Instrumentalness (無人聲)	Predicts whether a track contains **no vocals**. "Ooh" and "aah" sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly "vocal".	Float: 0.0 to 1.0; The closer the instrumentalness value is to **1.0**, the greater likelihood the track contains **no vocal** content. Values above 0.5 are intended to represent instrumental tracks, but confidence is higher as the value approaches 1.0.
Liveness	Detects the **presence of an audience** in the recording. Higher liveness values represent an increased probability that the track was performed live.	Float: 0.0 to 1.0 A value above 0.8 provides strong likelihood that the track is live.
Valence	A measure from 0.0 to 1.0 describing the **musical positiveness** conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).	Float: 0.0 to 1.0
Tempo	The overall estimated tempo of a track in beats per minute (BPM). In musical terminology, tempo is the speed or pace of a given piece and derives directly from the average beat duration.	Float: 57.967 to 220.29
Time Signature(拍號)	An estimated time signature. The time signature (meter) is a notational convention to specify **how many beats are in each bar** (or measure). The time signature ranges from 3 to 7 indicating time signatures of "3/4", to "7/4".	Integer: 3 to 7

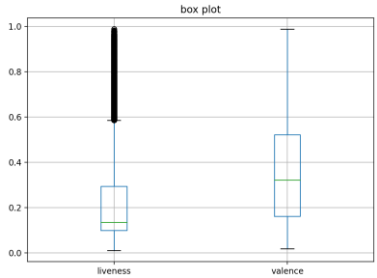
2. missing value and noise

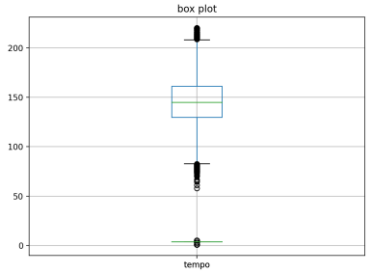
Audio feature	Missing value	Noise
Danceability	X	Y
Energy	X	Y
Key	X	X
Loudness	X	Y
Mode	X	X
Speechiness	X	Y
Acousticness	X	Y
Instrumentalness	X	X
Liveness	X	Y
Valence	X	X
Tempo	X	Y
Time Signature	X	X

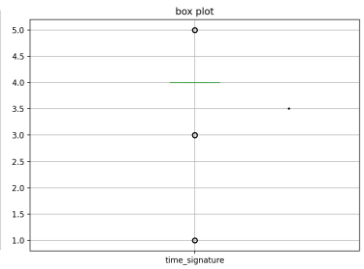
- Missing value: 根據下圖執行結果，總共有 42305 筆資料，所有 audio feature 欄位的非空值皆有 42305 筆，代表沒有空值
- Noise: 利用 box plot(下圖)，發現除了 key, mode, Instrumentalness, valence 以及 time signature，其他欄位皆有值大於第三四分位距 + 1.5 倍四分位距(Q3 + 1.5IQR) 或小於 第一四分位距 - 1.5 倍四分位距(Q1 - 1.5IQR)，這些超過的值極為 noise

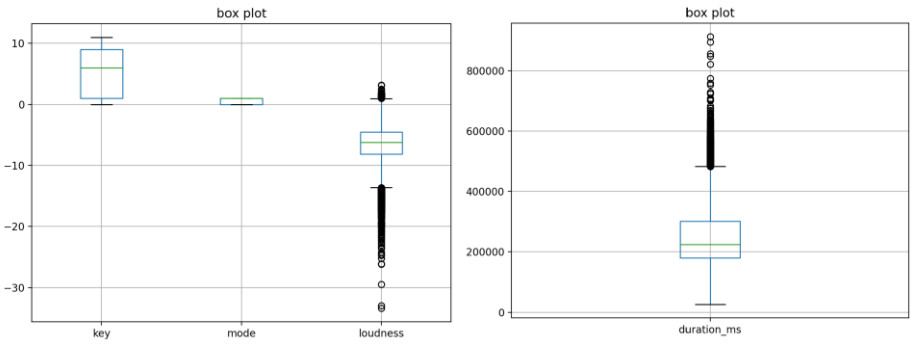
```
RangeIndex: 42305 entries, 0 to 42304
Data columns (total 22 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   danceability      42305 non-null  float64
 1   energy            42305 non-null  float64
 2   key               42305 non-null  int64
 3   loudness          42305 non-null  float64
 4   mode              42305 non-null  int64
 5   speechiness       42305 non-null  float64
 6   acousticness      42305 non-null  float64
 7   instrumentalness  42305 non-null  float64
 8   liveness          42305 non-null  float64
 9   valence           42305 non-null  float64
 10  tempo             42305 non-null  float64
 11  type              42305 non-null  object
 12  id                42305 non-null  object
 13  uri               42305 non-null  object
 14  track_href        42305 non-null  object
 15  analysis_url      42305 non-null  object
 16  duration_ms       42305 non-null  int64
 17  time_signature    42305 non-null  int64
 18  genre             42305 non-null  object
 19  song_name         21519 non-null  object
 20  Unnamed: 0        20780 non-null  float64
 21  title             20780 non-null  object
dtypes: float64(10), int64(4), object(8)
memory usage: 7.1+ MB
```









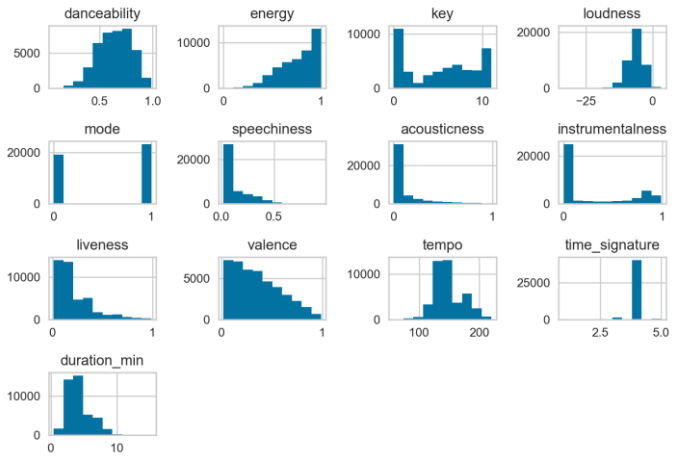


第二題 資料前處理

1. 去除不必要欄位

   甲、經過 .value_counts() 或 .unique() 觀察，發現 id, song name, uri, track_href, analysis_url, Unnamed: 0, title 欄位只是為了辨識各個歌曲的名字、type 欄位全部的值都是 audio feature，皆對於歌曲的分群沒有幫助，因此先去除，最後留下 12 種 audio feature

2. Null value：經過觀察 (df.isnull().sum()) 在各個欄位都是 0，因此不需要處理空值問題
3. Duration_ms 轉換成分鐘，經過觀察，如果用毫秒為單位，每筆資料的時長就會相差很多，很難找出相似性，也不符合人聽歌的單位習慣(我們聽歌都是以分鐘和秒為單位)，因此轉換為分鐘
4. 去除 outlier：經過前一題 box plot 的觀察，發現有些值落在大於 3 個標準差外，因此把這些值去除
5. Feature 標準化

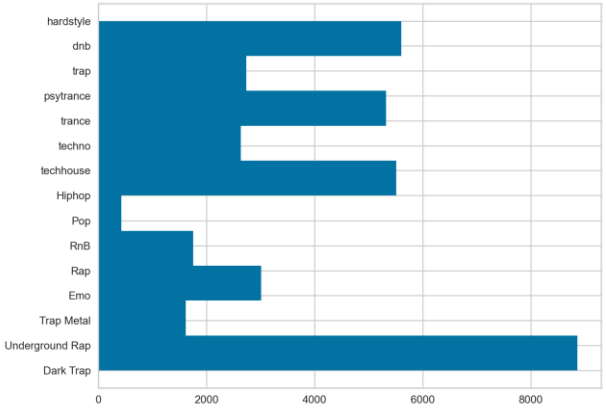


   甲、在 describe data 的時候(上圖)，發現每個 columns 之間的平均數、變異數相差很大，因此利用 StandardScalar()也就是 **z-score** 的方法 把資料標準化，避免資料因為值域的不同而產生干擾

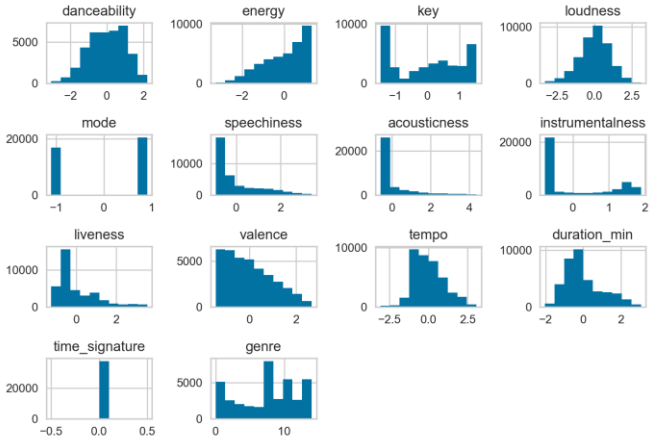
   乙、在使用 python 進行標準化時，發現轉化完各個欄位的平均數並不完全

=0，而是接近 0 的數字，像是-0.00000000000000004736951571，推測原因可能是電腦浮點數的表示限制導致

6. Encoding：因為 dependent variable genre 是 nominal variable，需要利用 LabelEncoder()事先先轉換成數字表示。以下是各種歌曲資料分布狀況。



下圖是經過前處理的資料分布狀況



第三題-Random Forest

1. Train test split：首先，我們先把 data 分割成 75% training data 和 0.25% testing data 去評估較好的模型參數。
2. 使用 Random Search 和 Grid search 去嘗試不同的模型參數

首先先利用 random search 從比較大的參數範圍隨機選取參數組合找出 accuracy 較高的參數組合以縮小 grid search 的參數範圍，以幫助我們去更好地設定 random forest classifier 的參數，尤其是其中的 n_estimator (幾顆 decision tree)和 max_feature(要考慮幾個 feature)

a. Random search

- **參數範圍：**

{'bootstrap': [True, False],

'max_depth': [10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None],
'max_features': ['auto', 'sqrt'],
'min_samples_leaf': [1, 2, 4],
'min_samples_split': [2, 5, 10],
'n_estimators': [200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000]}

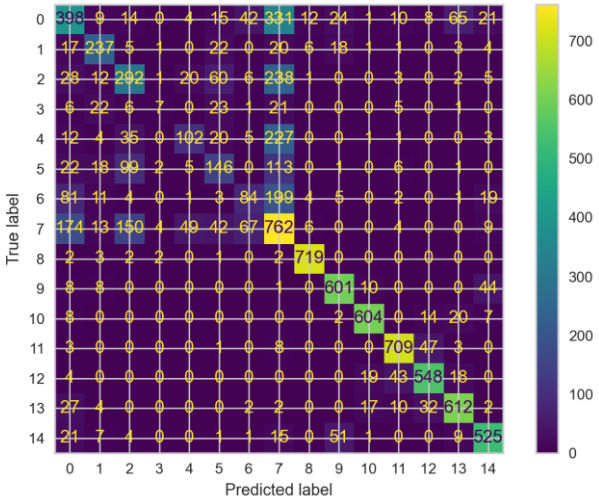
- 由 Random search 找到的參數

{'n_estimators': 1600, 'min_samples_split': 10, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': 20, 'bootstrap': True}

- 此組參數在 testing data 上的 Performance:

  Accuracy: 0.6772678762006403

  Confusion matrix:



b. Grid search

由上一個 random search 步驟縮小參數範圍後，我們針對上述表現較佳的參數組合，找出數字範圍接近的參數組合去更小範圍地找出較好的參數

- Grid search 參數範圍

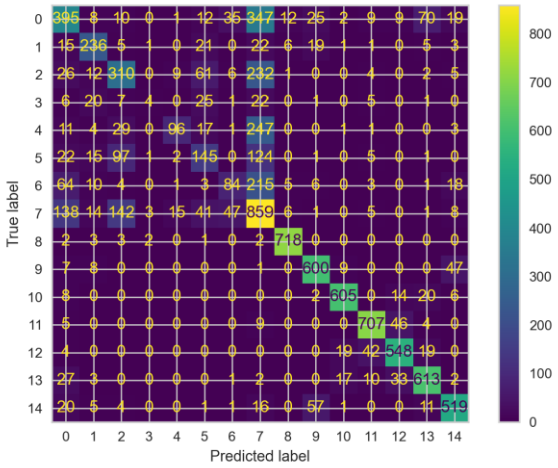
{'n_estimators': [1500, 1600, 1700], 'min_samples_split': [10], 'min_samples_leaf': [1], 'max_features': ['sqrt'], 'max_depth': [15, 20, 25], 'bootstrap': [True]}

- Grid search 找到的較佳參數

{'bootstrap': True, 'max_depth': 15, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 1600}

- 此組參數在 testing data 上的表現

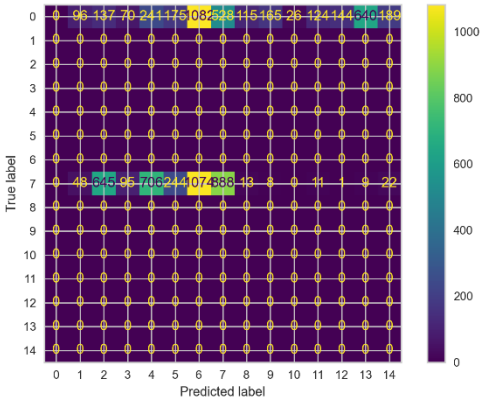
Accuracy: 0.6871931696905016



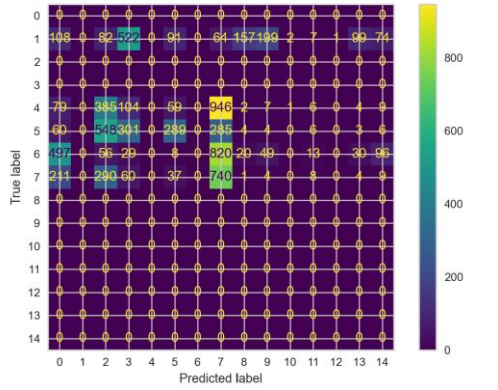
c. k-fold(k=5), SMOTE (Oversampling) 和 Model Evaluation

由 Grid Search 得到較佳參數組後，我們進行 k-fold validation 把資料切割成 5 份，輪流當 testing data，以下是 5 組各自的 Accuracy 和 Confusion matrix(Fold 0 ~ Fold 4)：

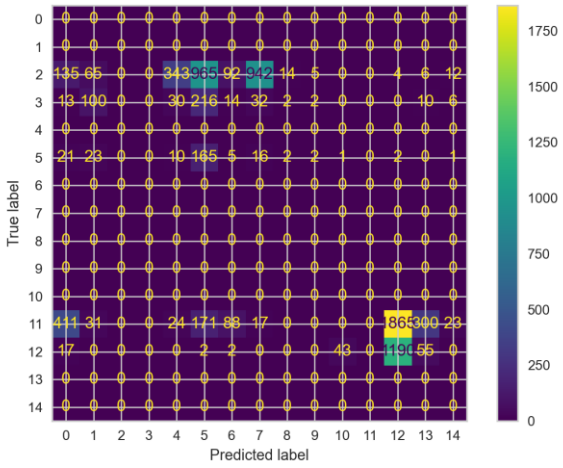
```
fold: 0 Accuracy: 0.11846318036286019
```



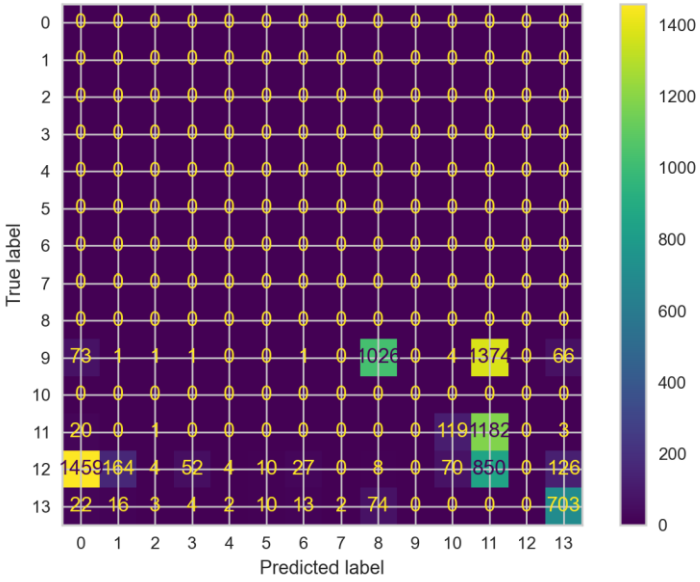

```
fold: 1 Accuracy: 0.13727321237993598
```



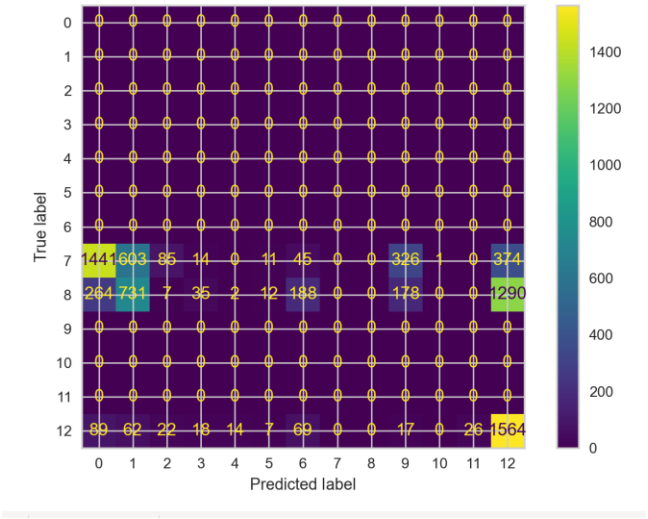

```
fold: 2 Accuracy: 0.180787191460974
```




```
fold: 3 Accuracy: 0.2515010006671114
```




```
fold: 4 Accuracy: 0.20867244829886591
```



d. Overall Accuracy 和 Feature importance

- 最後將 5 次的結果取平均，得到最後的 accuracy score 是 0.18

```
Overall accuracy Score: 0.17933940663394948
```

- Feature importance:以下是 5 次 validation 的各自的 feature importance，我們可以發現較為重要的 feature 是 0, 3, 7, 10, 12 (超過 0.1)，分別是 dancibility, loudness, Instrumentalness, tempo, time_signature

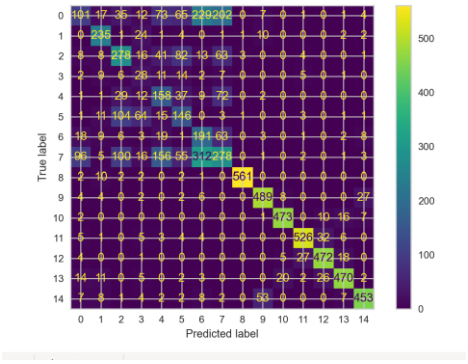
```
Fold 0
Features:
[0.10278045 0.06547822 0.01854125 0.07899017 0.00696397 0.06303512
 0.05460445 0.1123722  0.03558831 0.06492006 0.25929378 0.
 0.13743203]
Fold 1
Features:
[0.09961947 0.07378295 0.02050389 0.08595194 0.00598569 0.06228366
 0.05649785 0.10104943 0.03548962 0.06473867 0.25974088 0.
 0.13435595]
Fold 2
Features:
[0.09132307 0.07158752 0.0179021  0.08500607 0.0064854  0.0573115
 0.05085894 0.11748754 0.02904505 0.0471965  0.27001136 0.
 0.15578495]
Fold 3
Features:
[0.09774448 0.07046623 0.02263833 0.07803088 0.00796695 0.07006547
 0.05420776 0.1169013  0.03619127 0.0729153  0.23177111 0.
 0.14110093]
Fold 4
Features:
[0.10517229 0.07446692 0.02137658 0.08709869 0.00599394 0.06456431
 0.05605139 0.12470126 0.03718211 0.07012649 0.19872213 0.
 0.1545439 ]
```

e. 優化方式: **Stratified k fold**

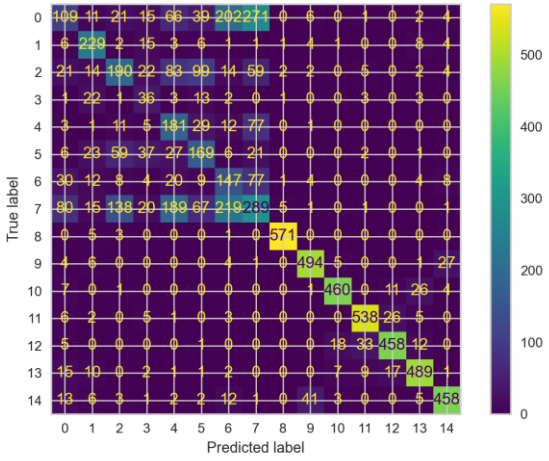
因為發現如果單純只是把 training data 和 testing data 切分，正確率可以達到 0.68，但進行 k fold 時，正確率掉到 0.2，加上資料相當不平衡，

因此就在思考是不是在 k fold 也要進行**分層抽樣**再做 oversampling(比較符合原本資料的分布)，依照以下結果發現結果確實進步許多：

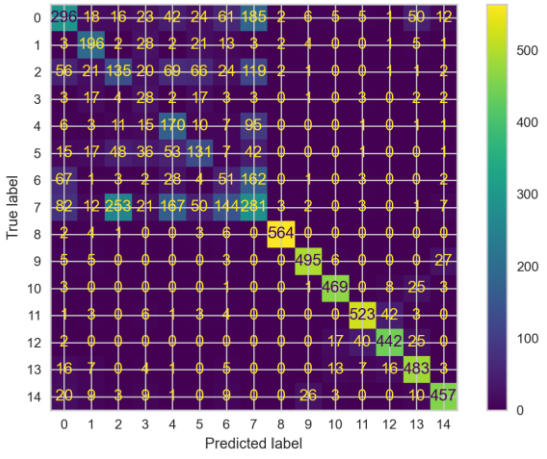
```
fold: 0 Accuracy: 0.6482123799359658
```



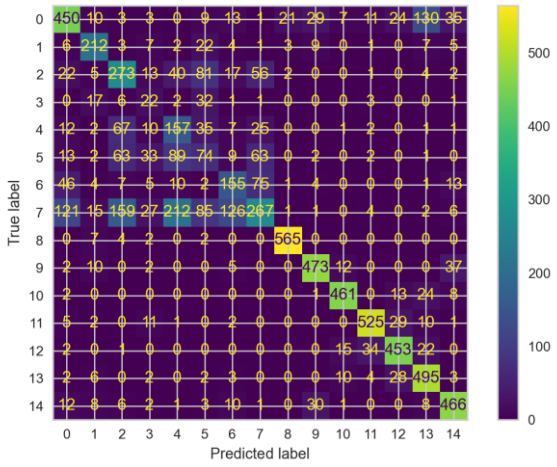

```
fold: 1 Accuracy: 0.6427427961579509
```



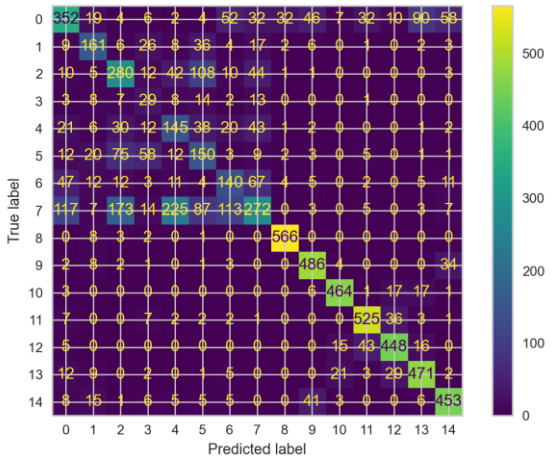

```
fold: 2 Accuracy: 0.6298865910607071
```




```
fold: 3 Accuracy: 0.6735156771180787
```




```
fold: 4 Accuracy: 0.6593729152768513
```




```
Overall accuracy Score: 0.6507460719099107
```

f. **哪些種類之間難以分別**：由上述結果發現 Dark trap 和 trap (0 和 13)、Dark trap 和 underground rap (0 和 7)、hiphop 和 RnB (2 和 5)、Hiphop 和 underground rap (2 和 7)、Rap 和 underground rap (4 和 7)

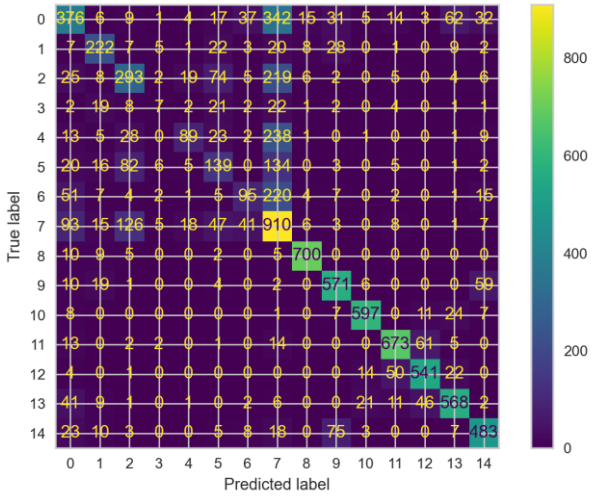
第四題 SVM

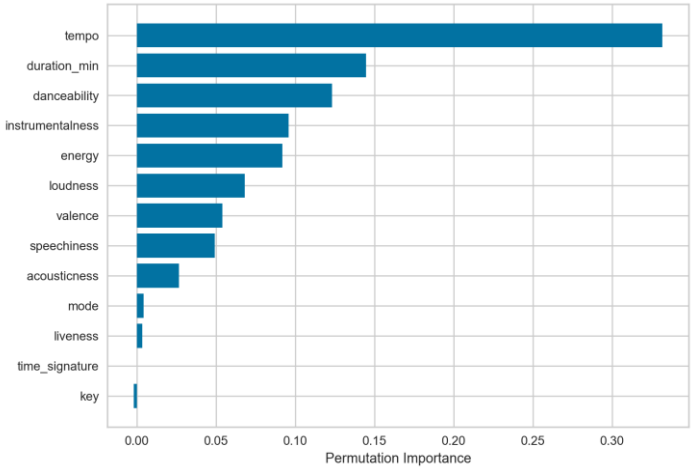
1. Train test split
2. Grid search 選擇參數
   - 參數選擇原因：

     首先先利用 Grid search，在{'C': [0.1, 1, 10, 100, 1000], 'gamma': [1, 0.1, 0.01, 0.001, 0.0001], 'kernel': ['rbf'], 'decision_function_shape':['ovo']}，找出較佳的參數 gamma 和 c。而因為我們的任務是要做 multi-class classification，因此 decision_function_shape 選擇一對一(one versus one)，kernel function 選擇 RBF 使我們可以使用非線性的分界線。
   - Grid search 結果

{'C': 1000, 'decision_function_shape': 'ovo', 'gamma': 0.01, 'kernel': 'rbf'}

- 此組參數在 testing data 中的表現：Accuracy 大約是 0.67



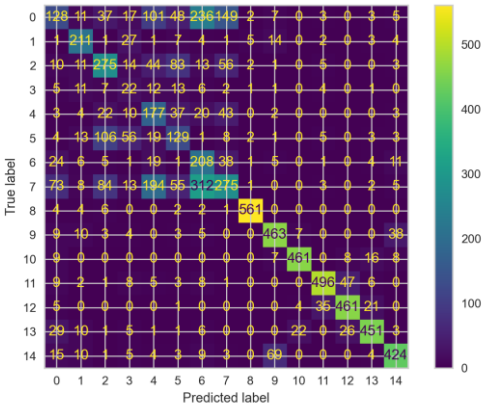


	precision	recall	f1-score	support
0	0.54	0.39	0.46	954
1	0.64	0.66	0.65	335
2	0.51	0.44	0.47	668
3	0.23	0.08	0.11	92
4	0.64	0.22	0.32	410
5	0.39	0.34	0.36	413
6	0.49	0.23	0.31	414
7	0.42	0.71	0.53	1280
8	0.94	0.96	0.95	731
9	0.78	0.85	0.82	672
10	0.92	0.91	0.92	655
11	0.87	0.87	0.87	771
12	0.82	0.86	0.84	632
13	0.80	0.80	0.80	708
14	0.77	0.76	0.77	635
accuracy			0.67	9370
macro avg	0.65	0.61	0.61	9370
weighted avg	0.68	0.67	0.66	9370

3. Stratified K fold, oversampling, evaluation
   a. k-fold(k=5), SMOTE (Oversampling) 和 Model Evaluation

      由 Grid Search 得到較佳參數組後，我們進行 stratified k-fold validation 把資料切割成 5 份，輪流當 testing data，並針對 training data 進行 oversampling，以下是 5 組各自的 Accuracy 和 Confusion matrix(Fold 0 ~ Fold 4)

```
fold: 0 Accuracy: 0.6326040554962646
```

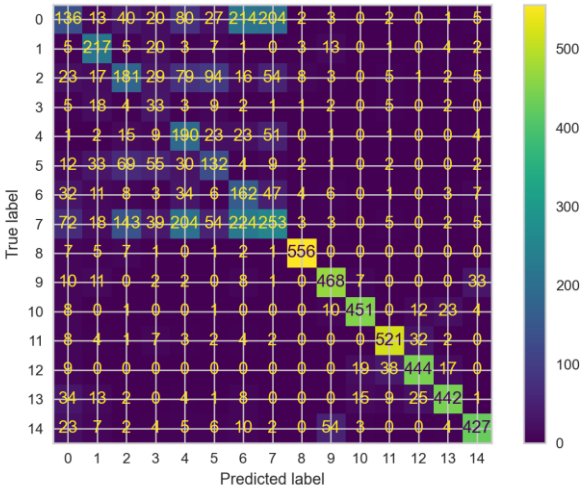



```
Fold 0
Performance:
              precision    recall  f1-score   support

           0       0.39      0.17      0.24       747
           1       0.68      0.75      0.71       281
           2       0.50      0.53      0.52       517
           3       0.12      0.26      0.16        85
           4       0.31      0.55      0.39       321
           5       0.33      0.37      0.35       350
           6       0.25      0.64      0.36       324
           7       0.48      0.27      0.34      1025
           8       0.98      0.97      0.97       580
           9       0.81      0.85      0.83       542
          10       0.93      0.91      0.92       509
          11       0.90      0.85      0.87       586
          12       0.85      0.87      0.86       527
          13       0.88      0.81      0.84       555
          14       0.84      0.78      0.80       547

    accuracy                           0.63      7496
   macro avg       0.62      0.64      0.61      7496
weighted avg       0.66      0.63      0.63      7496
```

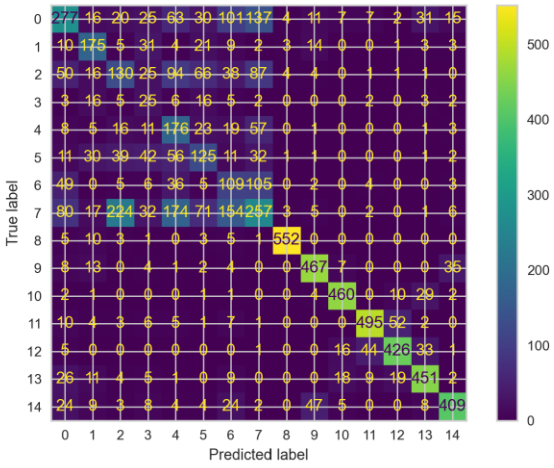
```
fold: 1 Accuracy: 0.6153948772678762
```



Fold 1
Performance:

	precision	recall	f1-score	support
0	0.35	0.18	0.24	747
1	0.59	0.77	0.67	281
2	0.38	0.35	0.36	517
3	0.15	0.39	0.21	85
4	0.30	0.59	0.40	320
5	0.36	0.38	0.37	351
6	0.24	0.50	0.32	324
7	0.40	0.25	0.31	1025
8	0.96	0.96	0.96	580
9	0.83	0.86	0.85	542
10	0.91	0.88	0.90	510
11	0.88	0.89	0.89	586
12	0.86	0.84	0.85	527
13	0.88	0.80	0.84	554
14	0.86	0.78	0.82	547
accuracy			0.62	7496
macro avg	0.60	0.63	0.60	7496
weighted avg	0.63	0.62	0.61	7496

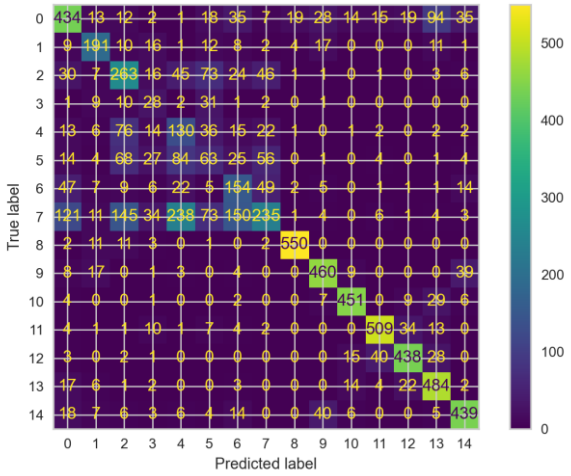
```
fold: 2 Accuracy: 0.6049366244162775
```



Fold 2
Performance:

	precision	recall	f1-score	support
0	0.49	0.37	0.42	746
1	0.54	0.62	0.58	281
2	0.28	0.25	0.27	517
3	0.11	0.29	0.16	85
4	0.28	0.55	0.37	320
5	0.34	0.36	0.35	351
6	0.22	0.34	0.27	324
7	0.38	0.25	0.30	1026
8	0.97	0.95	0.96	580
9	0.84	0.86	0.85	541
10	0.90	0.90	0.90	510
11	0.88	0.84	0.86	586
12	0.83	0.81	0.82	526
13	0.80	0.81	0.81	555
14	0.85	0.75	0.79	547
accuracy			0.60	7495
macro avg	0.58	0.60	0.58	7495
weighted avg	0.62	0.60	0.61	7495

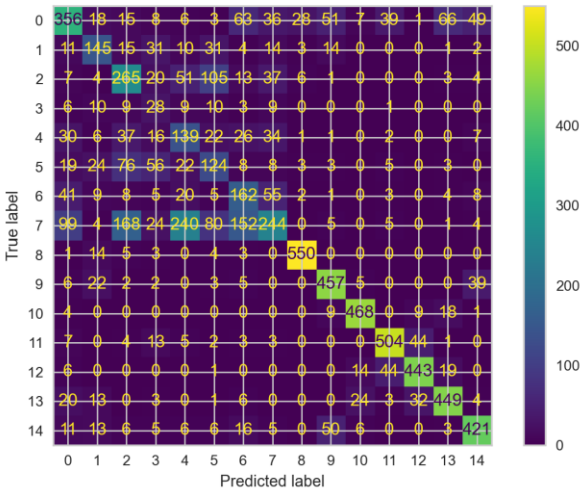
fold: 3 Accuracy: 0.6442961974649767



Fold 3
Performance:

	precision	recall	f1-score	support
0	0.60	0.58	0.59	746
1	0.66	0.68	0.67	282
2	0.43	0.51	0.47	516
3	0.17	0.33	0.22	85
4	0.24	0.41	0.30	320
5	0.20	0.18	0.19	351
6	0.35	0.48	0.40	323
7	0.56	0.23	0.32	1026
8	0.95	0.95	0.95	580
9	0.82	0.85	0.83	541
10	0.88	0.89	0.89	509
11	0.87	0.87	0.87	586
12	0.84	0.83	0.83	527
13	0.72	0.87	0.79	555
14	0.80	0.80	0.80	548
accuracy			0.64	7495
macro avg	0.61	0.63	0.61	7495
weighted avg	0.66	0.64	0.64	7495

fold: 4 Accuracy: 0.6316210807204803



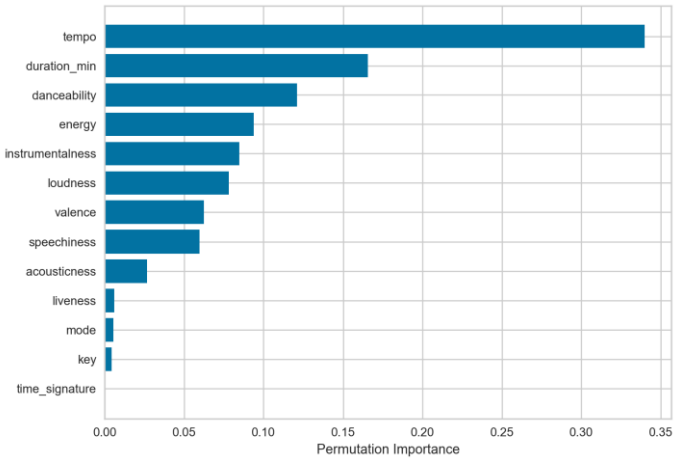
Fold 4

	precision	recall	f1-score	support
0	0.56	0.48	0.52	746
1	0.53	0.52	0.53	281
2	0.44	0.51	0.47	516
3	0.12	0.29	0.17	85
4	0.28	0.45	0.34	321
5	0.30	0.34	0.32	351
6	0.33	0.48	0.39	323
7	0.56	0.24	0.33	1026
8	0.92	0.95	0.94	580
9	0.78	0.84	0.81	541
10	0.89	0.92	0.90	509
11	0.82	0.86	0.84	586
12	0.83	0.83	0.83	527
13	0.79	0.80	0.80	555
14	0.77	0.76	0.77	548
accuracy			0.63	7495
macro avg	0.59	0.62	0.60	7495
weighted avg	0.65	0.63	0.63	7495

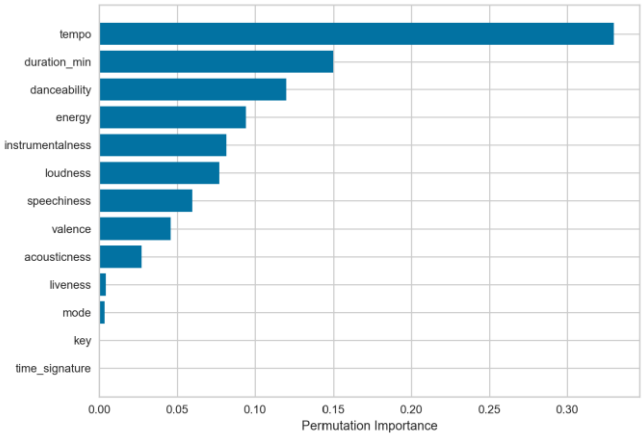
b. Overall accuracy, feature importance

- Overall accuracy
  i. `Overall accuracy Score: 0.6267846654083868`
- Feature importance：因為前面的 kernel function 使用 rbf，feature importance 計算的時候是使用 permutation importance 得到以下結果：

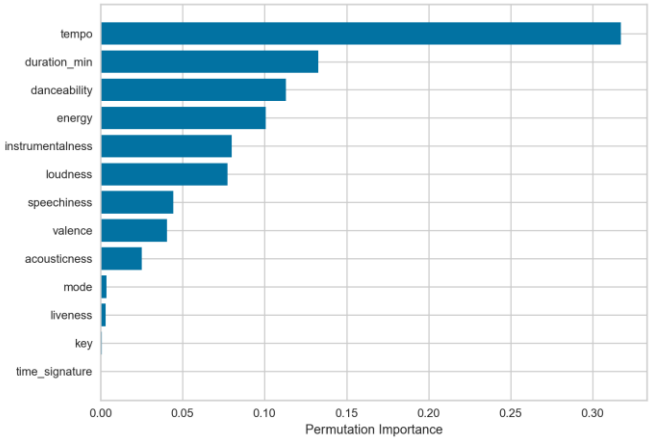
Fold 0



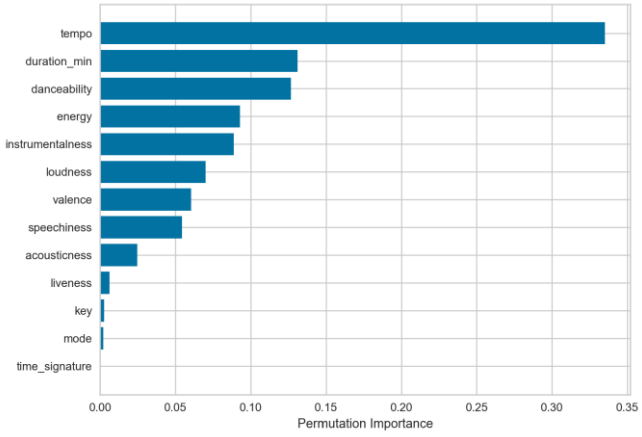
Fold 1



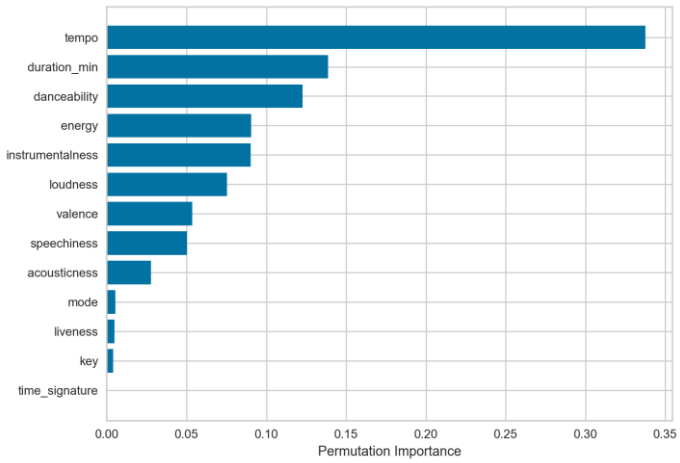
Fold 2



Fold 3



Fold 4



c. **難以區分的類別：**從上述結果我們發現， Dark trap、underground rap 和 trap metal (0、6 和 7)、hiphop 和 RnB (2 和 5)難以區分。

4. Improvement：stratified kfold、Adaboost
   a. **Stratified k fold:** 在以上建模的過程中，我們一開始採用的 k fold 並沒有分層抽樣，導致 Accuracy 最高只有到 0.28，其他大多在 0.1~0.2 之間徘徊，而後來採用 stratified k fold 就大幅提升 Accuracy，推測原因是因為資料真的太不平衡。
   b. **Adaboost:** 除了 stratified k fold 之外，我們後來還有跑 adaboost 去看效果是否有比較好，結果發現只有比原本稍微好一點到 0.64

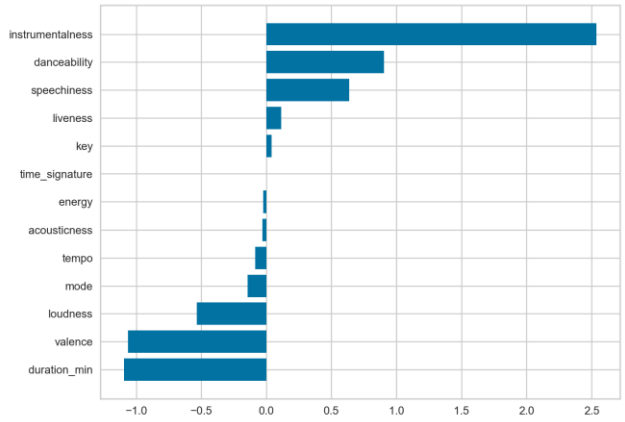
	precision	recall	f1-score	support
0	0.51	0.39	0.44	954
1	0.47	0.61	0.53	335
2	0.51	0.39	0.44	668
3	0.00	0.00	0.00	92
4	0.55	0.21	0.31	410
5	0.36	0.31	0.33	413
6	0.49	0.20	0.29	414
7	0.42	0.73	0.53	1280
8	0.87	0.91	0.89	731
9	0.78	0.74	0.76	672
10	0.91	0.90	0.91	655
11	0.80	0.86	0.83	771
12	0.83	0.84	0.84	632
13	0.77	0.77	0.77	708
14	0.79	0.68	0.73	635
accuracy			0.64	9370
macro avg	0.60	0.57	0.57	9370
weighted avg	0.64	0.64	0.63	9370

第五題 Clustering

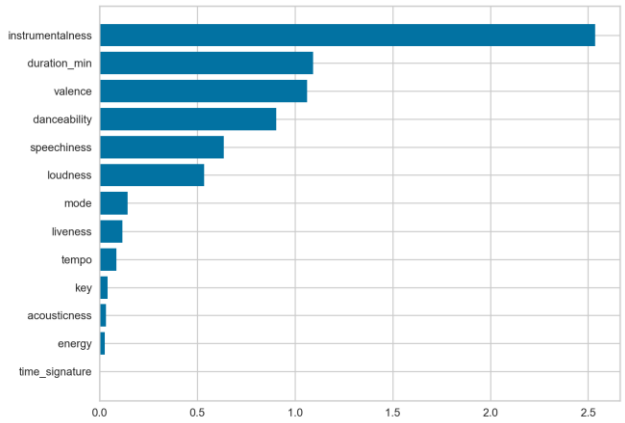
1. Linear SVM 的 feature importance

   甲、Grid search: 透過 grid search 找出的 linear SVM 較好參數為{'C': 10, 'decision_function_shape': 'ovo', 'gamma': 1, 'kernel': 'linear'}

   乙、 Feature importance: (x 軸是係數 coef)



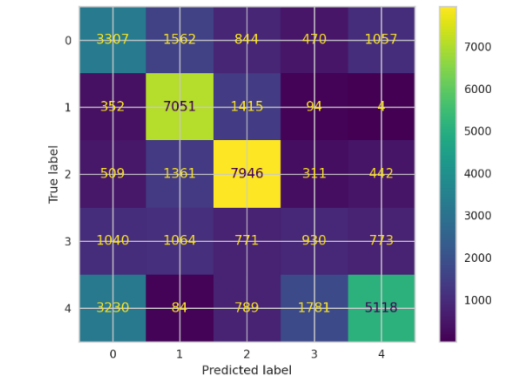
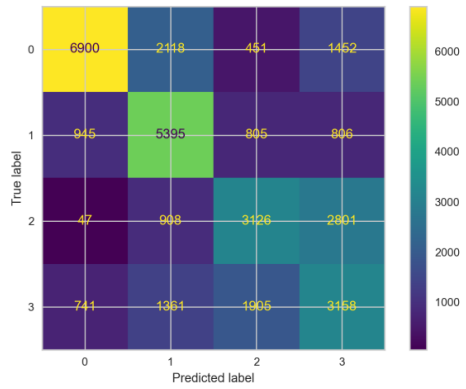
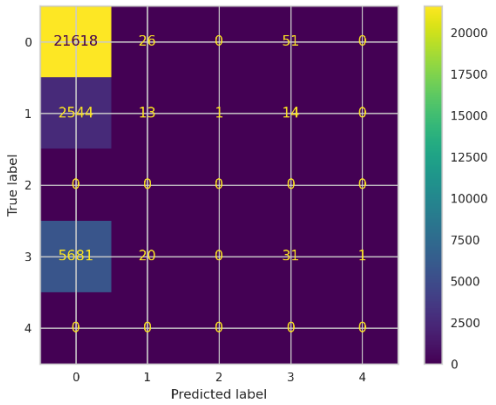
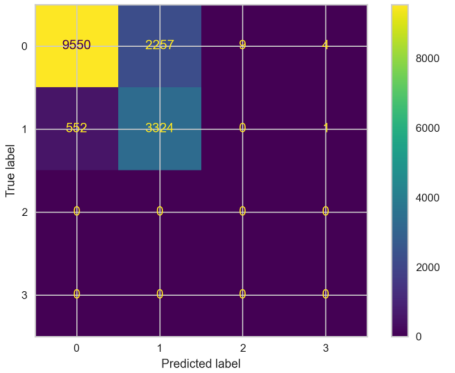
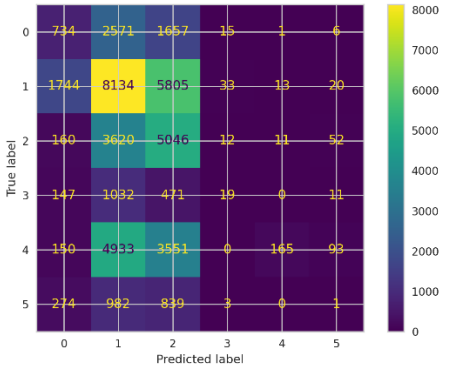
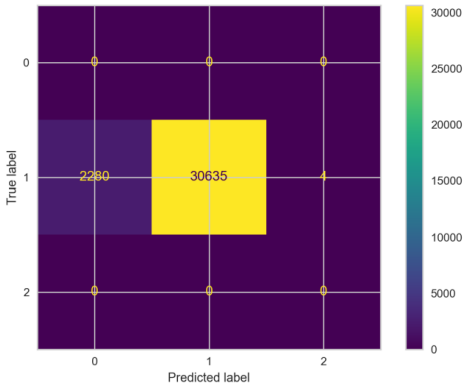
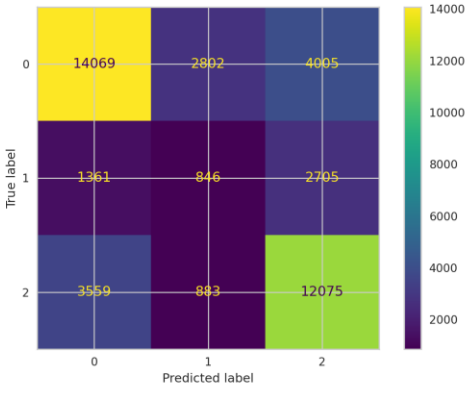
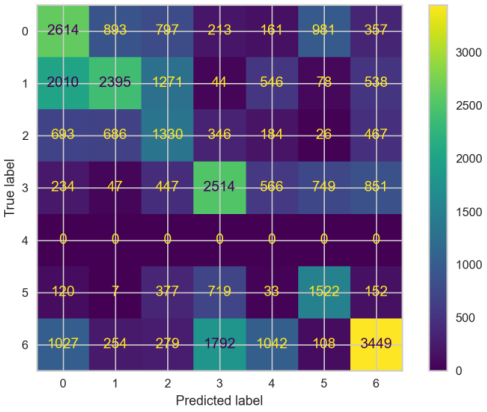
取絕對值(如果不在意正負號只在乎影響程度)

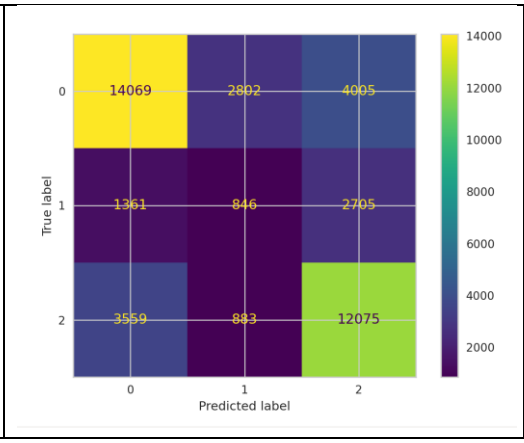
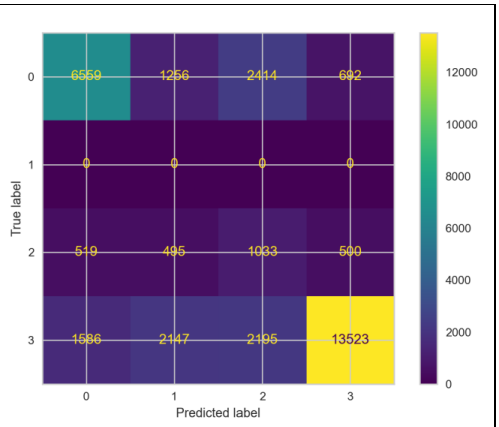


丙、Feature：最後選擇 top 6 features 來跑 clustering

2. Clustering 結果與作業二比較

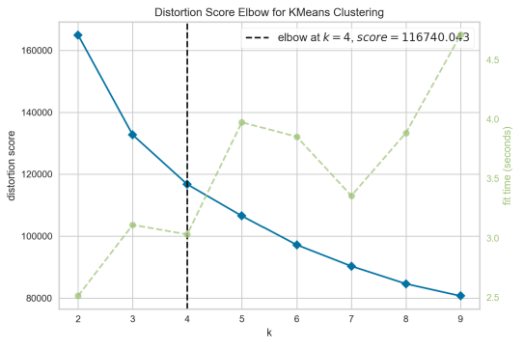
甲、Confusion matrix

model	HW2	Hw3(use top 6 features)
k-means		
Hierarchical		
DBSCAN		
BIRCH		

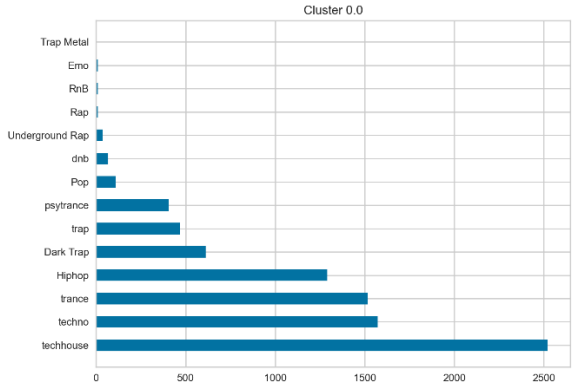
GMM		

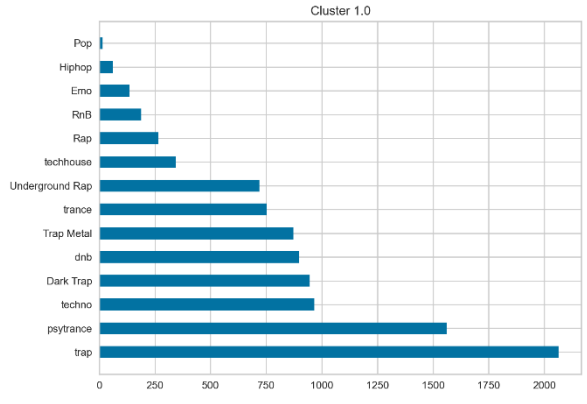
乙、 分析：由上表的 confusion matrix 以及 genre 在各個 cluster 的分布可以發現經過 feature importance 後的效果較佳。首先是資料在各個 cluster 的分布較為平均，在作業二的結果中 cluster 的資料分布相當不平均，此外，我們也可以發現就算在群數多的情況，多數情況而言資料還是能被分到正確的群。

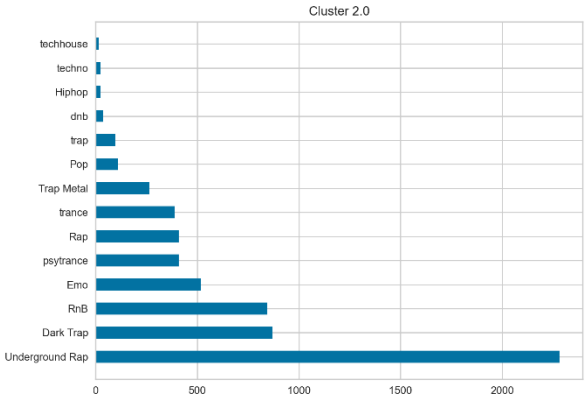
3. **各個 cluster 模型的執行結果：**包含參數設定(elbow)、silhouette coefficient、各個 cluster genre 分布、confusion matrix、Rand Index、Normalized Mutual Information、Adjusted Mutual Information、V Measure、Fowlkes-Mallows Scores
   a. 使用 Top 6 features
      甲、K means
         i. Elbow: k=4

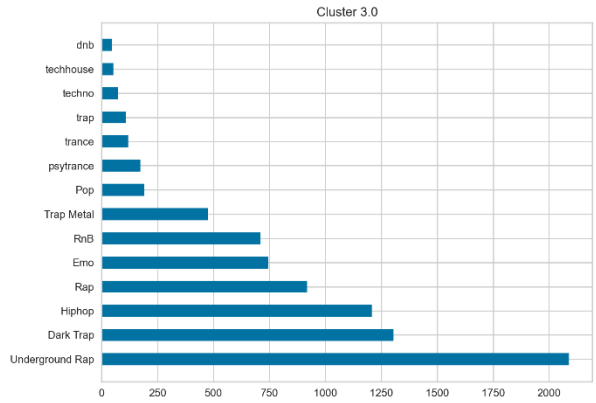


         ii. 各個 cluster 的 genre 分布



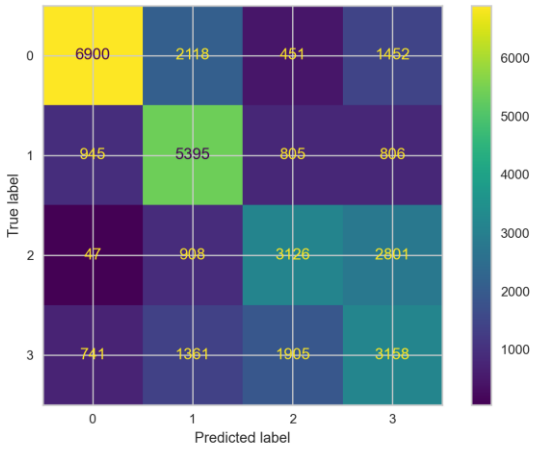







```
1.0    9782
0.0    8633
3.0    8217
2.0    6287
Name: cluster, dtype: int64
```

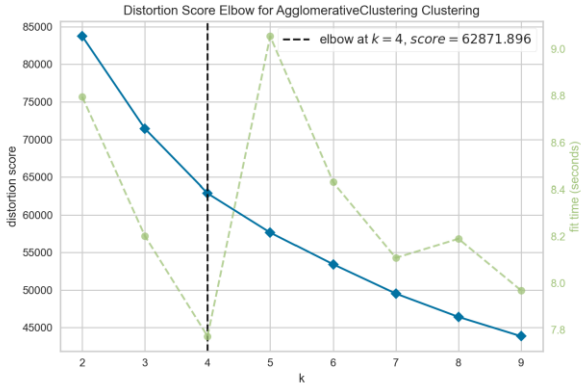
iii. Performance




```
Rand Index: 0.7064086521336634
Normalized Mutual Information: 0.22870802101472246
Adjusted Mutual Information: 0.22863111376598855
V Measure: 0.22870802101472243
Fowlkes-Mallows Scores: 0.43025029180963253
```

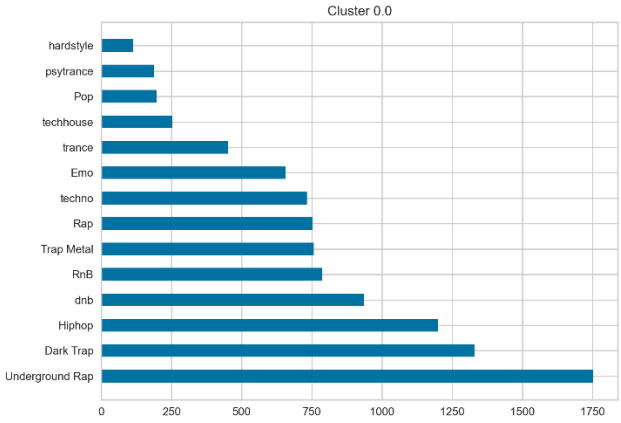
乙、Hierarchical clustering

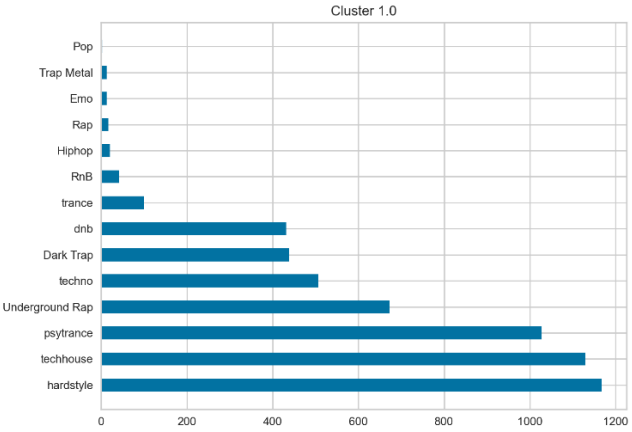
i. Elbow

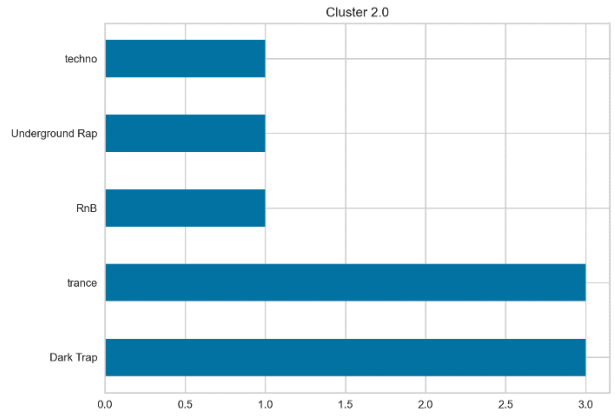


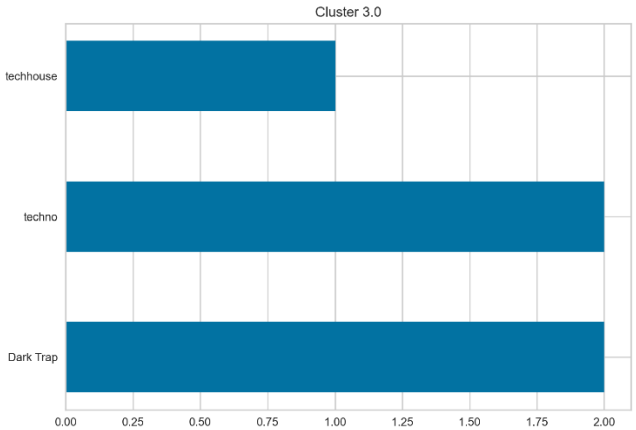

```
Silhoutte Coefficient: 0.12577474429391736
```

ii. 各個 genre 的歌曲分布


Cluster 0.0
hardstyle
psytrance
Pop
techhouse
trance
Emo
techno
Rap
Trap Metal
RnB
dnb
Hiphop
Dark Trap
Underground Rap
0
250
500
750
1000
1250
1500
1750

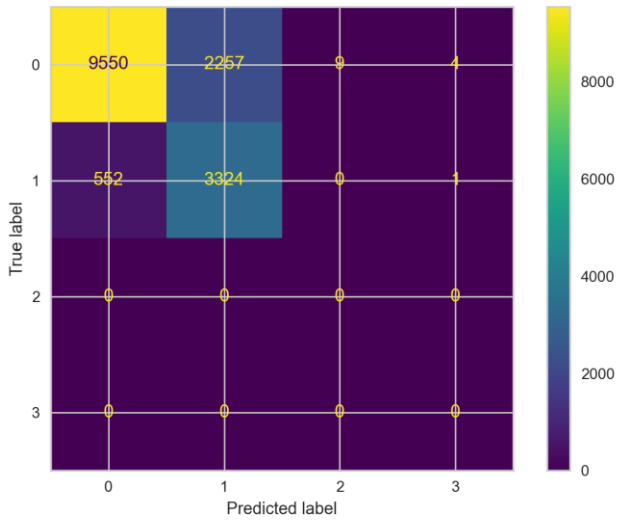

Cluster 1.0
Pop
Trap Metal
Emo
Rap
Hiphop
RnB
trance
dnb
Dark Trap
techno
Underground Rap
psytrance
techhouse
hardstyle
0
200
400
600
800
1000
1200


Cluster 2.0
techno
Underground Rap
RnB
trance
Dark Trap
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Cluster 3.0
techhouse
techno
Dark Trap
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

```
0.0    10102
1.0     5581
2.0        9
3.0        5
Name: cluster, dtype: int64
```

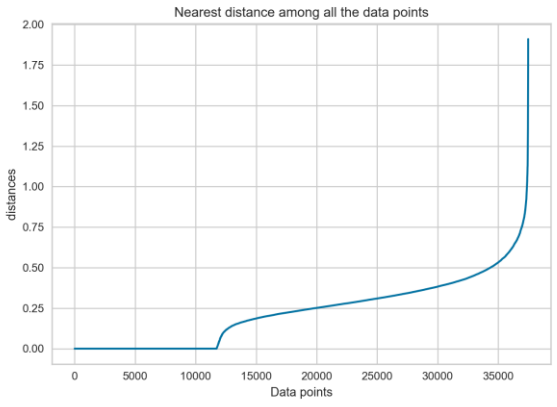
iii. Performance




```
Rand Index: 0.7051671686877065
Normalized Mutual Information: 0.29991886752255
Adjusted Mutual Information: 0.29979656169460966
V Measure: 0.29991886752255004
Fowlkes-Mallows Scores: 0.7497982562042084
```

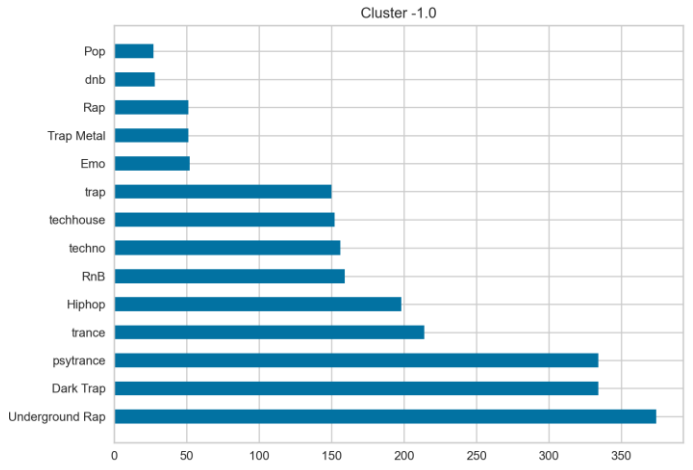
丙、DB Scan

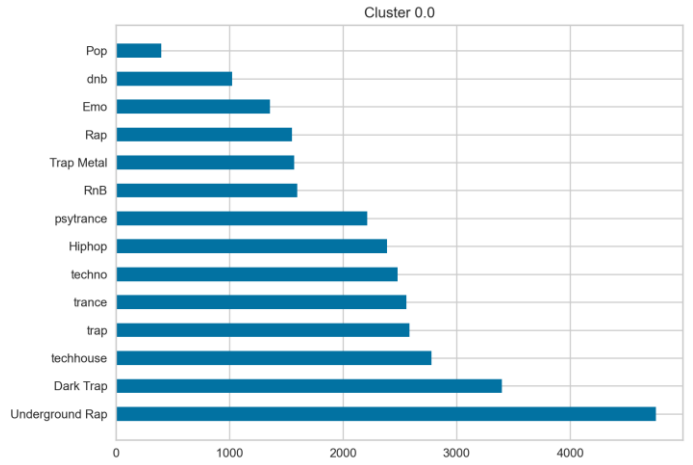
i. elbow using k nearest neighbor

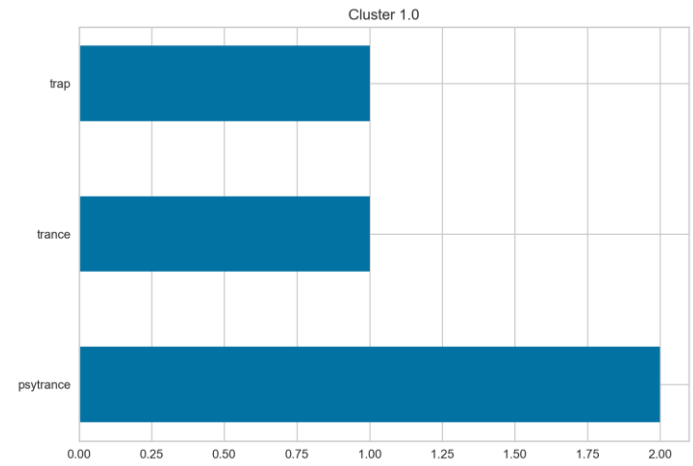



```
Silhoutte Coefficient: -0.17190793029211074
```

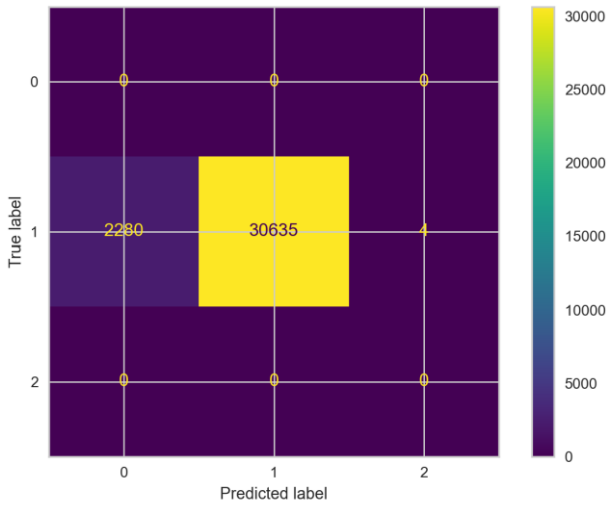
ii. Genre of each cluster







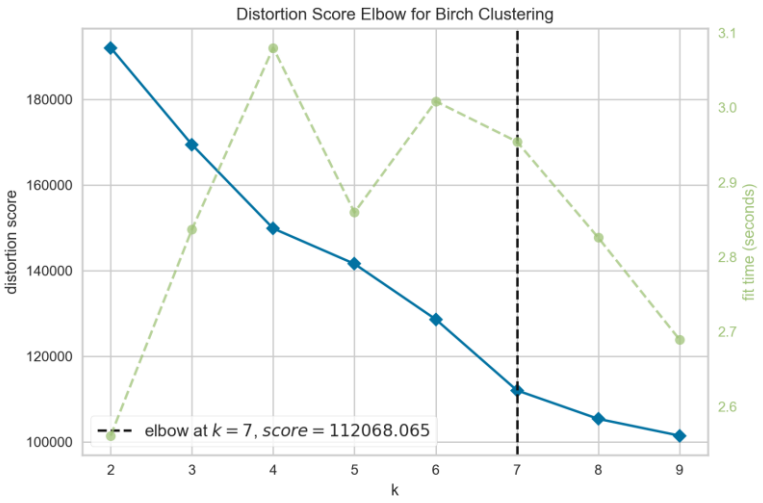
iii. Performance




```
Rand Index: 0.870422389985507
Normalized Mutual Information: 0.0
Adjusted Mutual Information: 0.0
V Measure: 0.0
Fowlkes-Mallows Scores: 0.9331892835853564
```

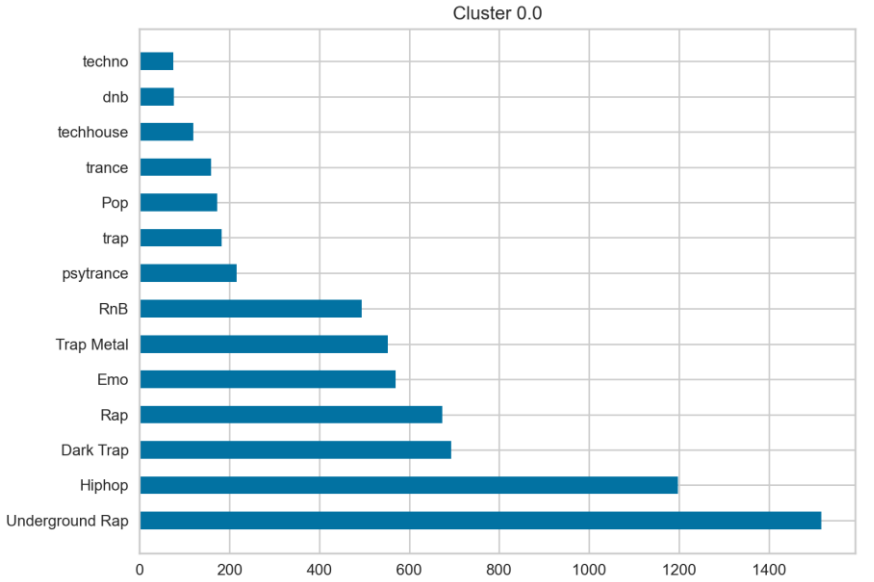
丁、Birch

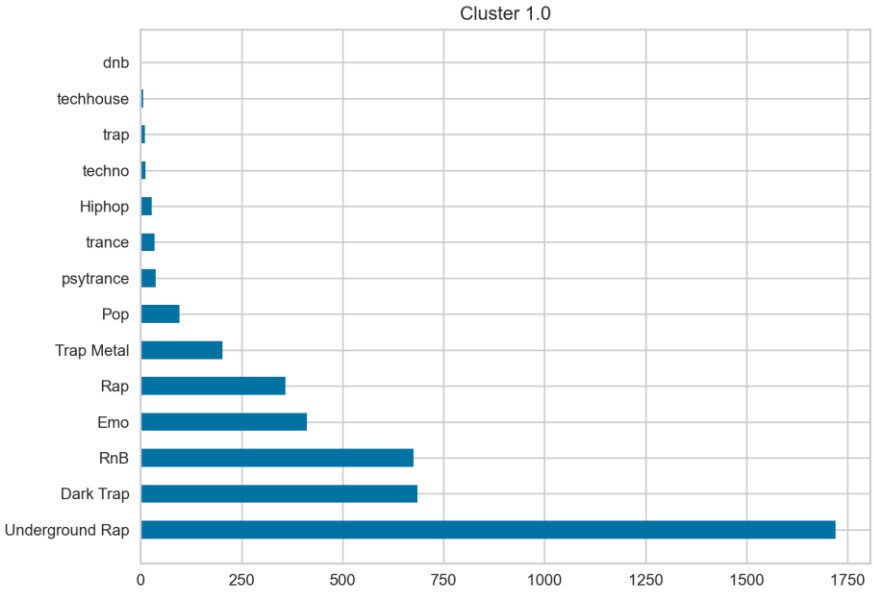
i. Elbow k=7

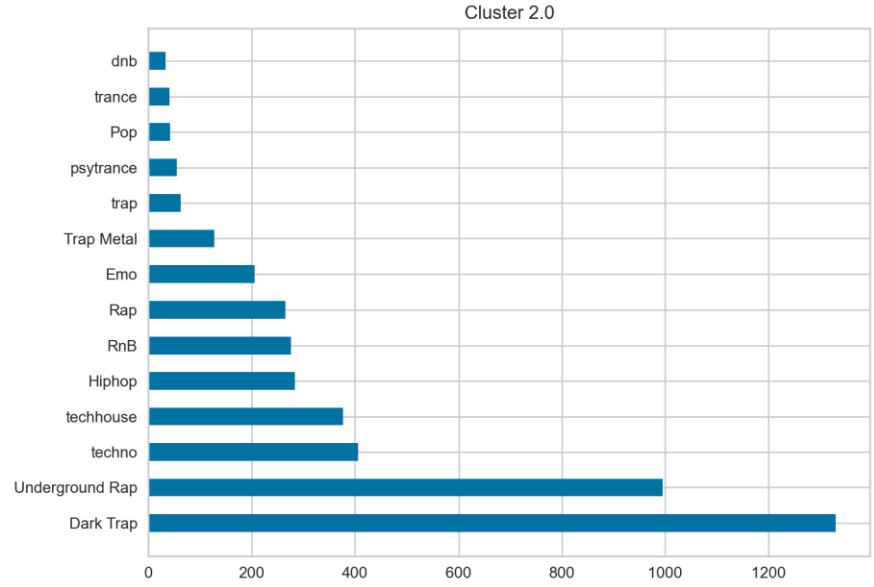


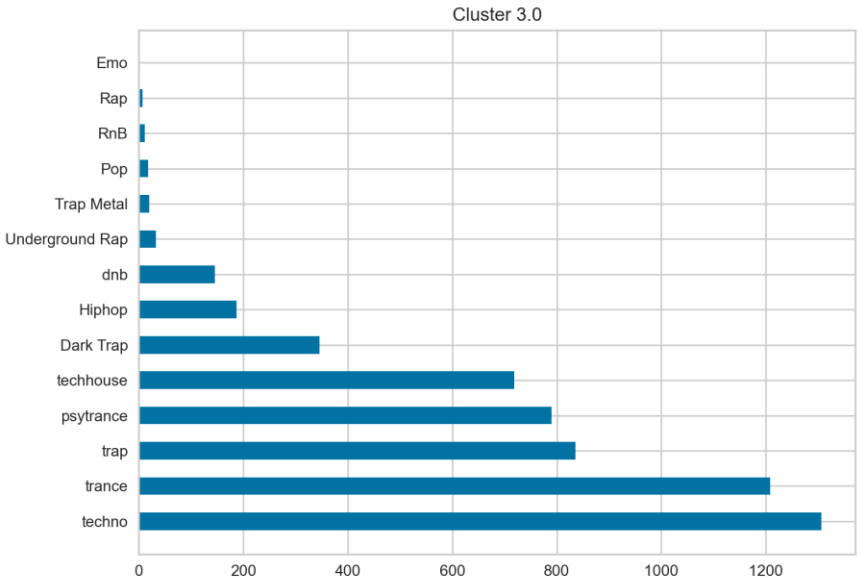

```
Silhoutte Coefficient: -0.0331352578531546
```

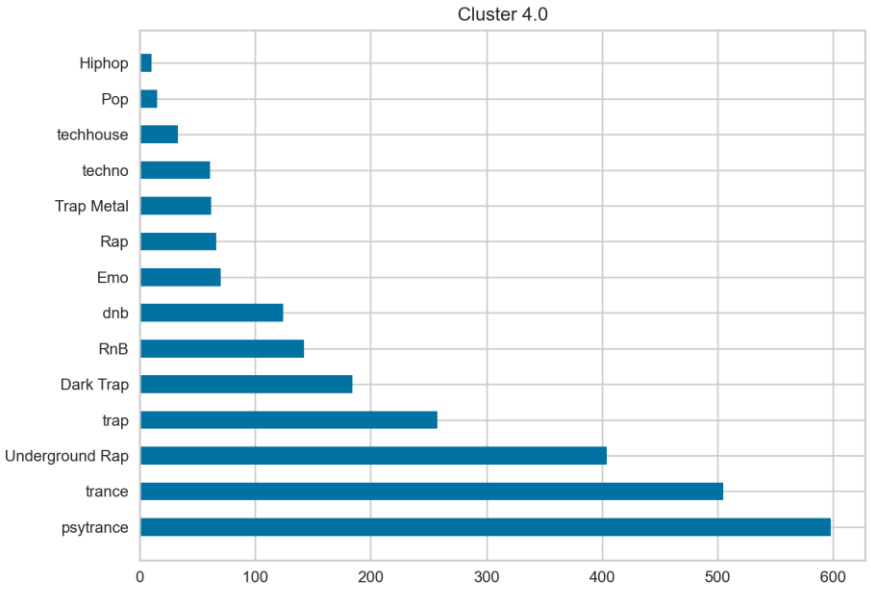
ii. Genre

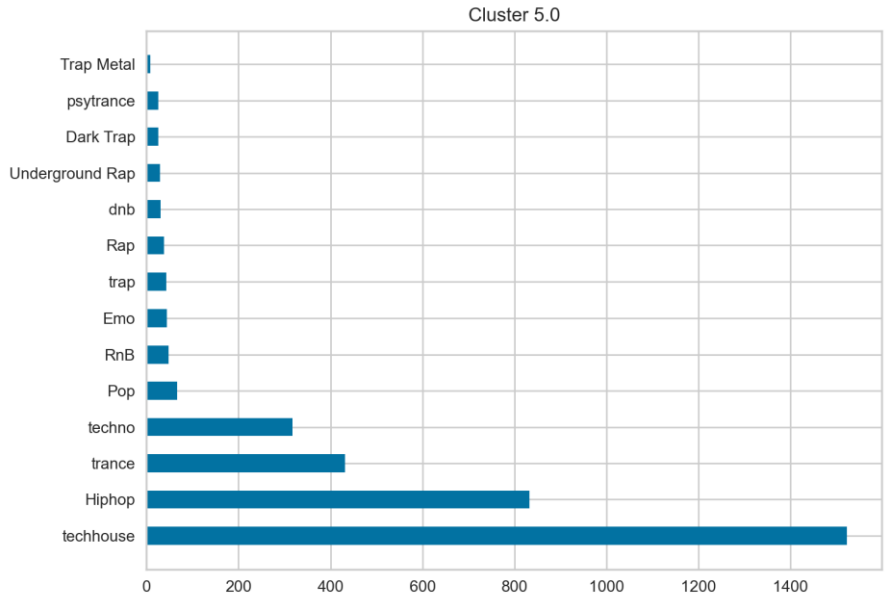

Cluster 0.0
techno
dnb
techhouse
trance
Pop
trap
psytrance
RnB
Trap Metal
Emo
Rap
Dark Trap
Hiphop
Underground Rap
0
200
400
600
800
1000
1200
1400

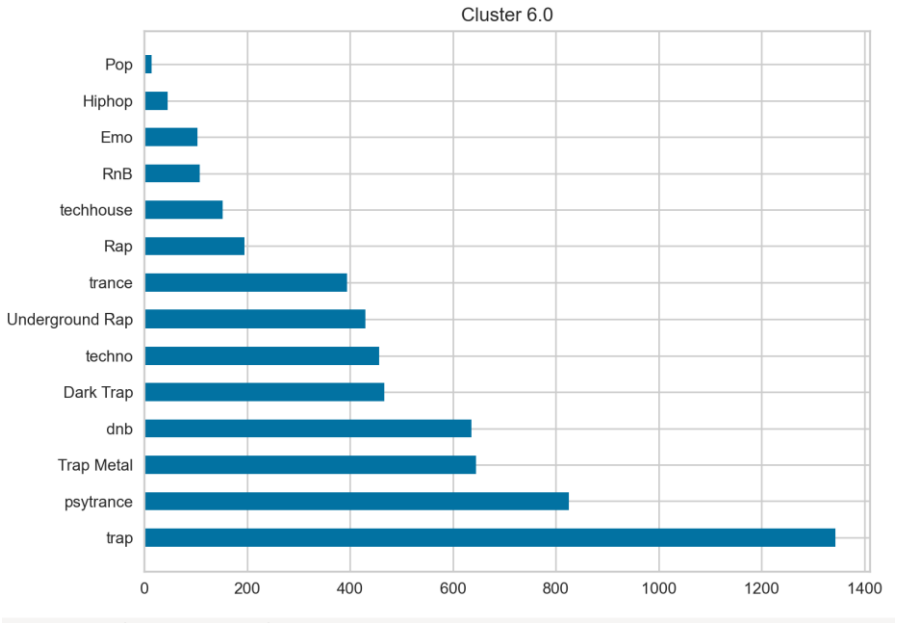

Cluster 1.0
dnb
techhouse
trap
techno
Hiphop
trance
psytrance
Pop
Trap Metal
Rap
Emo
RnB
Dark Trap
Underground Rap
0
250
500
750
1000
1250
1500
1750


Cluster 2.0
dnb
trance
Pop
psytrance
trap
Trap Metal
Emo
Rap
RnB
Hiphop
techhouse
techno
Underground Rap
Dark Trap
0
200
400
600
800
1000
1200


Cluster 3.0
Emo
Rap
RnB
Pop
Trap Metal
Underground Rap
dnb
Hiphop
Dark Trap
techhouse
psytrance
trap
trance
techno
0
200
400
600
800
1000
1200

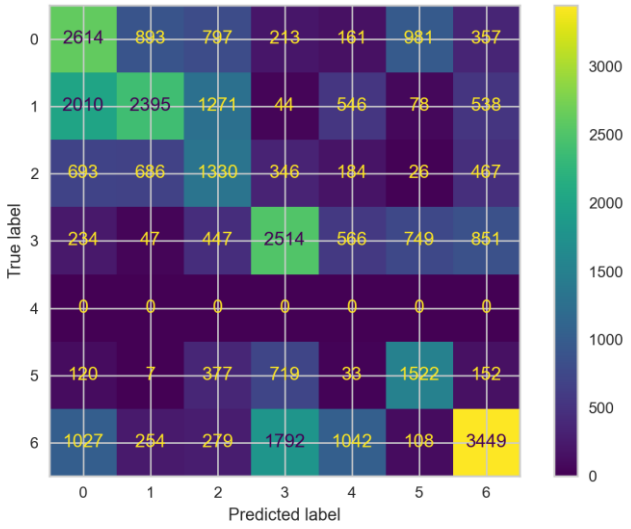

Cluster 4.0
Hiphop
Pop
techhouse
techno
Trap Metal
Rap
Emo
dnb
RnB
Dark Trap
trap
Underground Rap
trance
psytrance
0
100
200
300
400
500
600


Cluster 5.0
Trap Metal
psytrance
Dark Trap
Underground Rap
dnb
Rap
trap
Emo
RnB
Pop
techno
trance
Hiphop
techhouse
0
200
400
600
800
1000
1200
1400




```
0.0    6698
6.0    5814
3.0    5628
2.0    4501
1.0    4282
5.0    3464
4.0    2532
Name: cluster, dtype: int64
```

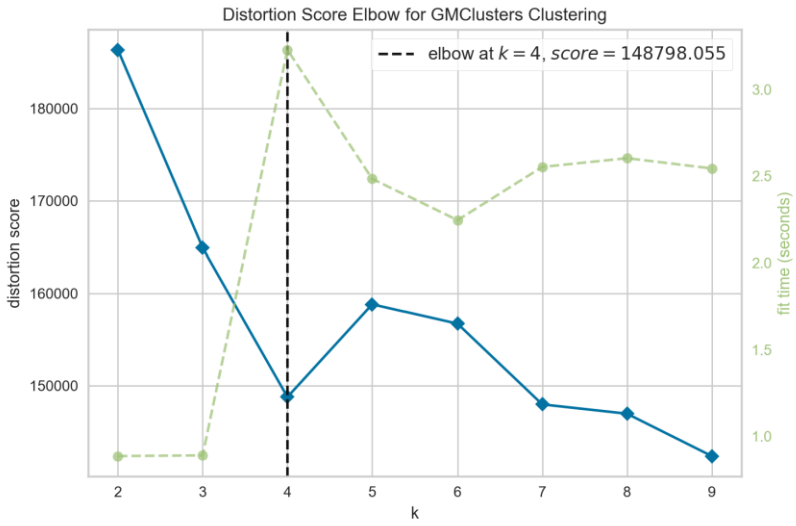
iii. Performance



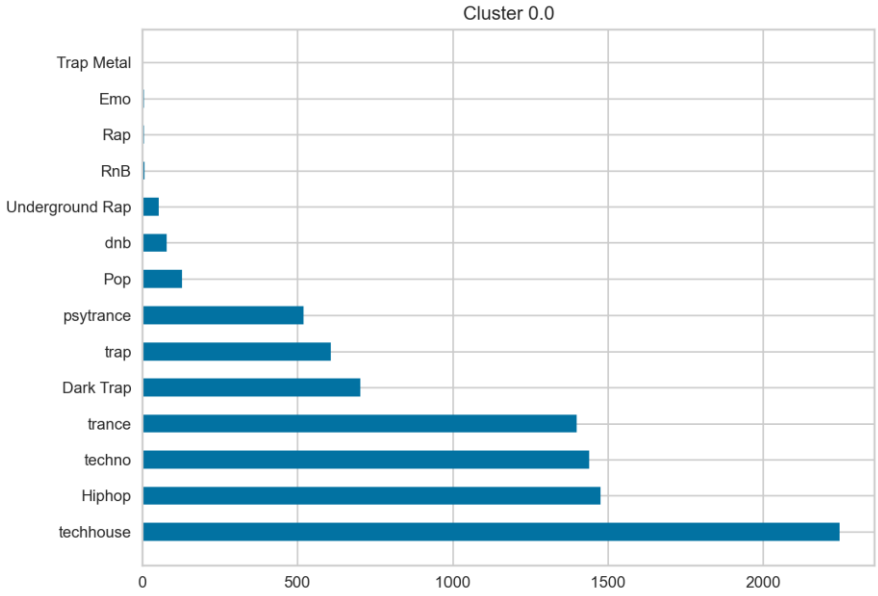

```
Rand Index: 0.7603710186639924
Normalized Mutual Information: 0.20450537002149577
Adjusted Mutual Information: 0.20430628135519327
V Measure: 0.20450537002149577
Fowlkes-Mallows Scores: 0.29121658681879475
```

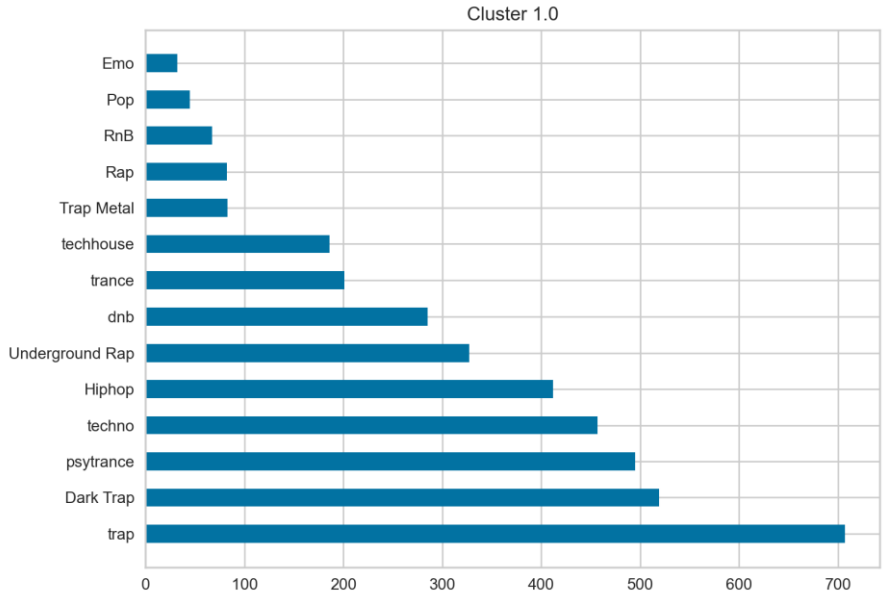
戊、GMM

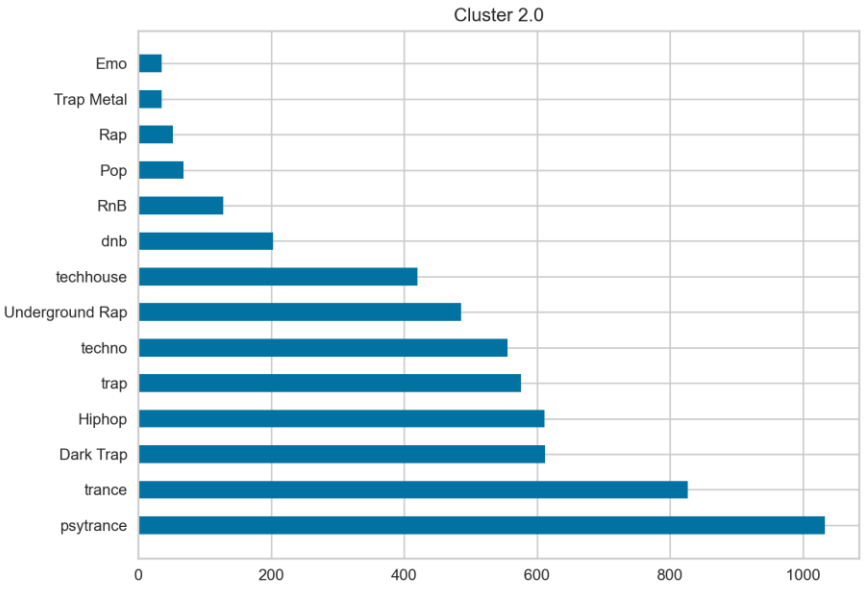
i. Elbow k=4

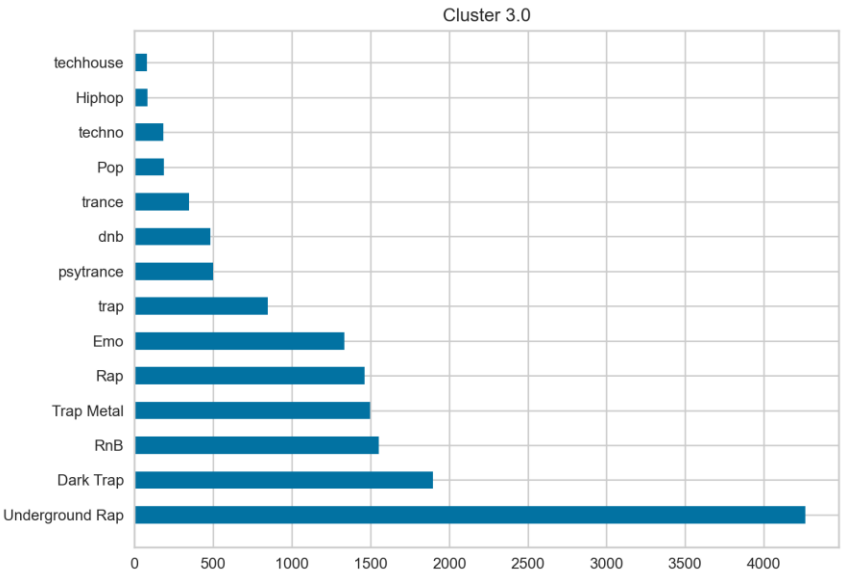

Distortion Score Elbow for GMClusters Clustering
elbow at $k = 4$, $score = 148798.055$
distortion score
fit time (seconds)
k
180000
170000
160000
150000
3.0
2.5
2.0
1.5
1.0
2
3
4
5
6
7
8
9
Silhoutte Coefficient: 0.07102915763499011

ii. Genre


Cluster 0.0
Trap Metal
Emo
Rap
RnB
Underground Rap
dnb
Pop
psytrance
trap
Dark Trap
trance
techno
Hiphop
techhouse
0
500
1000
1500
2000

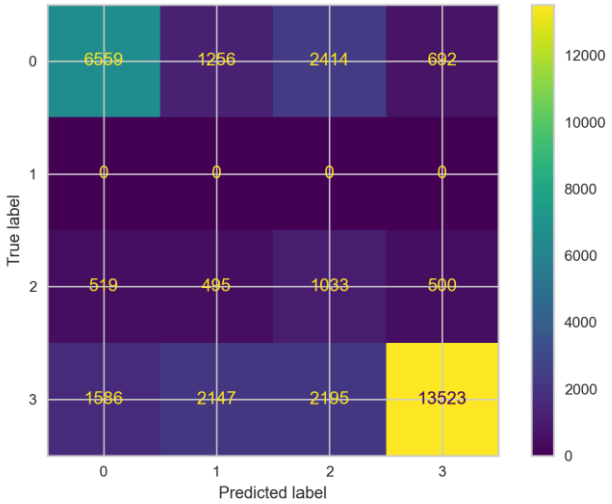

Cluster 1.0
Emo
Pop
RnB
Rap
Trap Metal
techhouse
trance
dnb
Underground Rap
Hiphop
techno
psytrance
Dark Trap
trap
0
100
200
300
400
500
600
700


Cluster 2.0
Emo
Trap Metal
Rap
Pop
RnB
dnb
techhouse
Underground Rap
techno
trap
Hiphop
Dark Trap
trance
psytrance
0
200
400
600
800
1000


Cluster 3.0
techhouse
Hiphop
techno
Pop
trance
dnb
psytrance
trap
Emo
Rap
Trap Metal
RnB
Dark Trap
Underground Rap
0
500
1000
1500
2000
2500
3000
3500
4000

```
3.0    14715
0.0     8664
2.0     5642
1.0     3898
Name: cluster, dtype: int64
```

iii. Performance




```
Rand Index: 0.6792081481435668
Normalized Mutual Information: 0.23477379131757034
Adjusted Mutual Information: 0.2347086275325618
V Measure: 0.23477379131757034
Fowlkes-Mallows Scores: 0.5991367961873573
```