# Fashion -dataset analysis

# 탐색적 데이터 분석

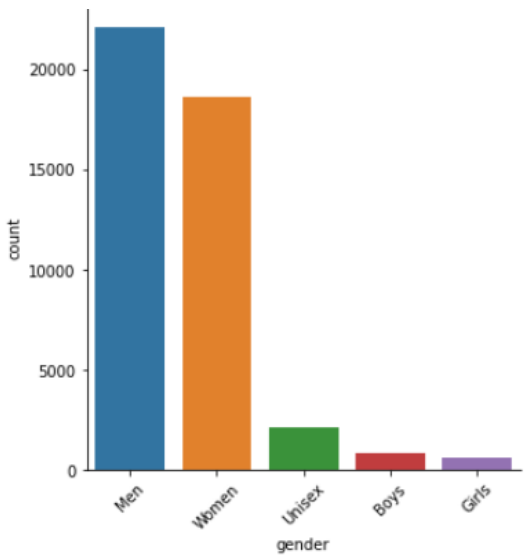


데이터는 Kaggle의 Fashion Product Images Dataset[1]으로 styles.csv, images.csv, image(jpg) 로 구성

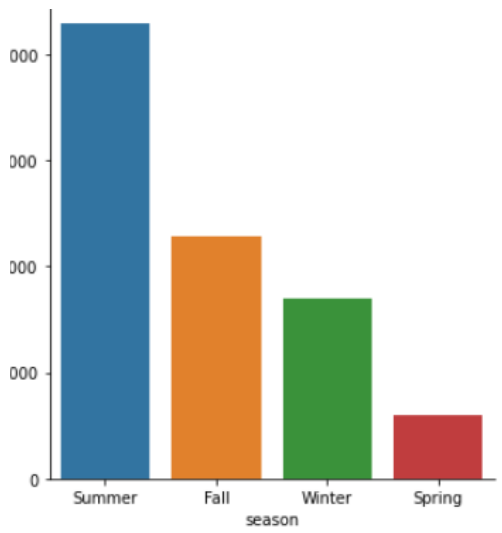
styles.csv 에는 각 image에 대한 메타 정보 (gender, season, usage, color, masterCategory, subcategory)가 있어, 이를 시각화하여 분포 확인

- 성별 분포: Men, Women의 수가 많고, 이외 Unisex, Boys, Girls의 수는 상대적으로 적은 편이다.
- 계절별 분포: Summer, Fall, Winter, spring 순으로 빈도가 높다.
- 용도별 분포: Casual이 압도적으로 빈도가 높으며, 이후 Sports, Ethnic, Formal 순으로 빈도가 높다.
- 대분류별 분포: Apparel, Accessories, Footwear, Personal Care 순으로 빈도가 높다.

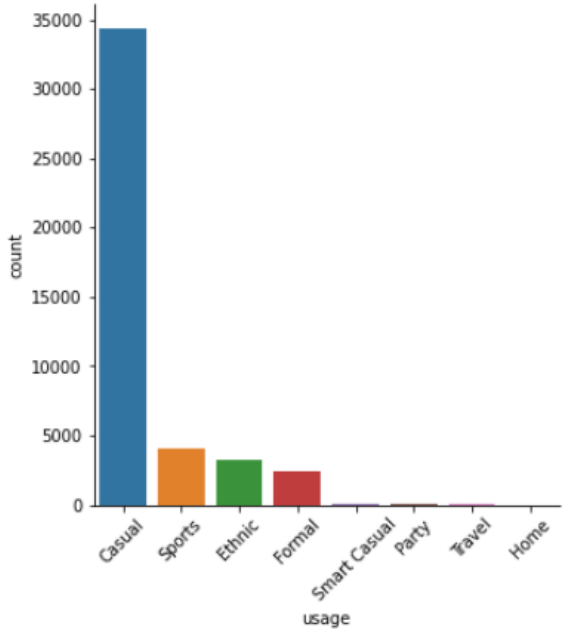
성별 분포



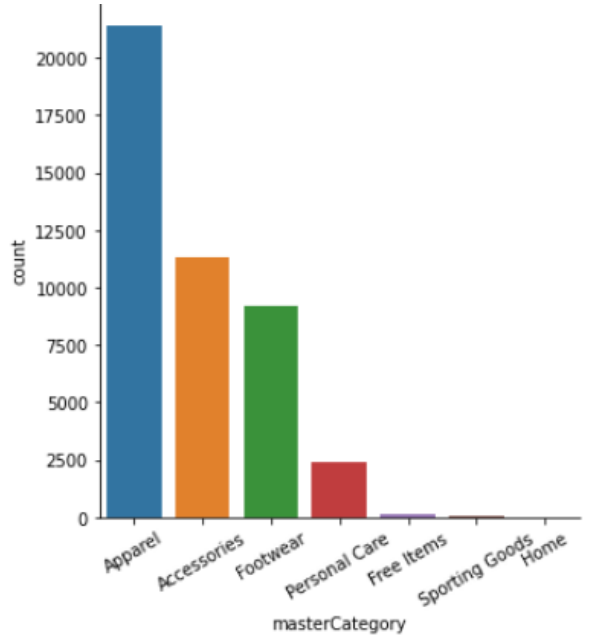

계절별 분포




용도별 분포




대분류별 분포




[1] https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-dataset.

# 탐색적 데이터 분석

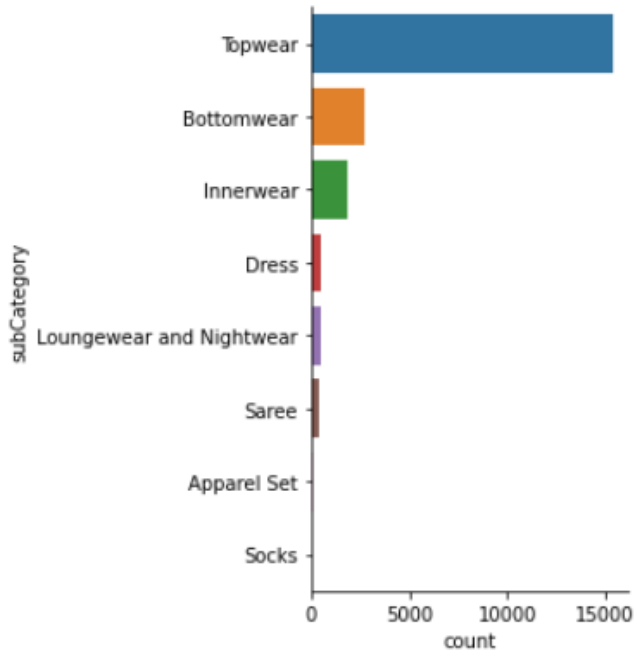


이미지의 카테고리 정보로 대분류(masterCategory)와 소분류(subcategory) 존재

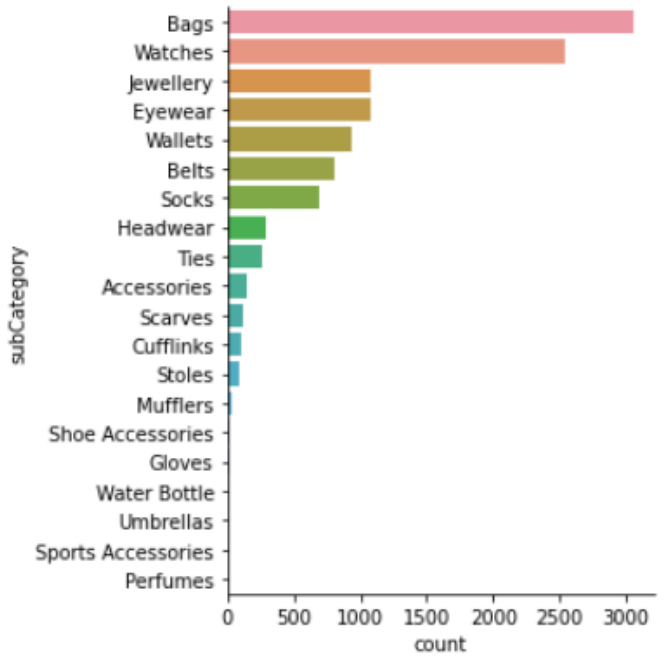
대분류와 소분류는 계층적 관계가 있어, 대분류에 따른 소분류 항목의 빈도를 시각화

- 대분류가 Apparel인 경우, 소분류 항목은 총 8개이며 Topwear, Bottomwear, Innerwear, Dress 순으로 빈도가 높다.
- 대분류가 Accessories 인 경우, 소분류 항목은 총 20개이며 Bags, Watches, Jewellery, Eyewear 순으로 빈도가 높다.
- 대분류가 Footwear 인 경우, 소분류 항목은 총 3개이며 Shoes, Sandal, Flip Flops 순으로 빈도가 높다.

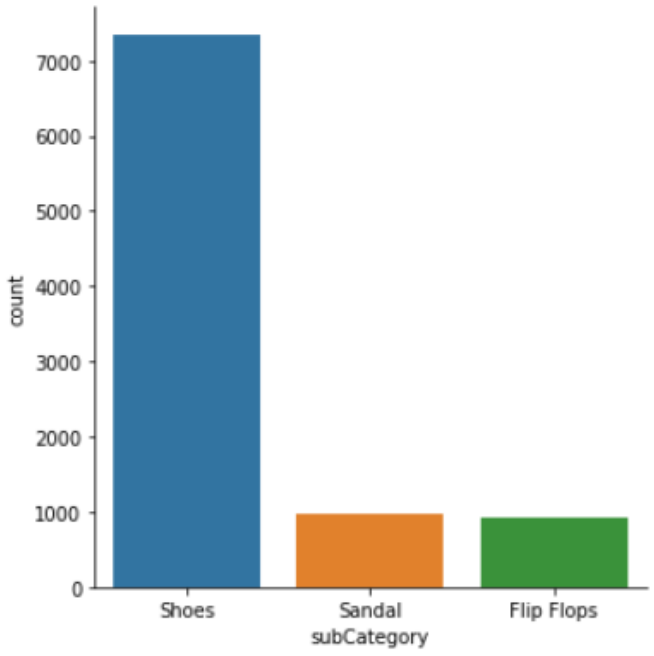
대분류 Appare의 소분류




대분류 Accessories의 소분류




대분류 Footwear의 소분류



# 탐색적 데이터 분석

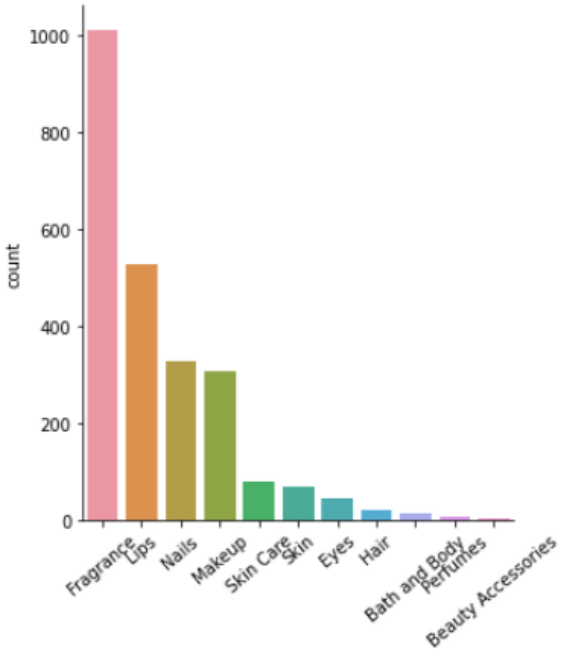


이미지의 카테고리 정보로 대분류(masterCategory)와 소분류(subcategory) 존재

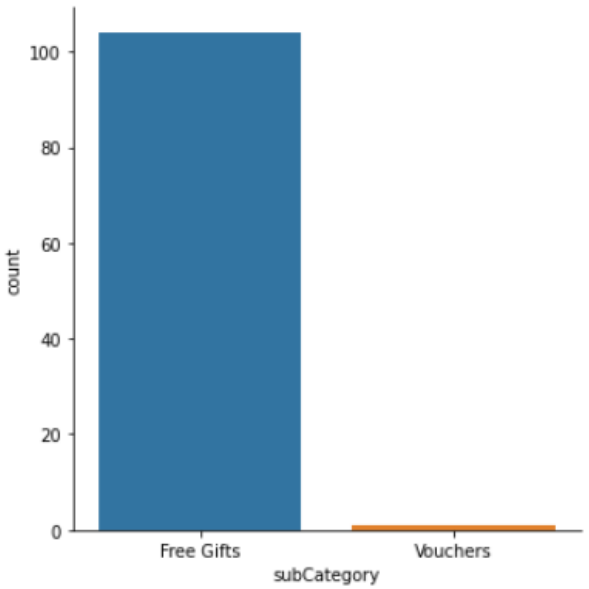
대분류와 소분류는 계층적 관계가 있어, 대분류(7가지)에 따른 소분류 항목의 빈도를 시각화

- 대분류가 Personal Care인 경우, 소분류 항목은 총 11개이며 Fragrance, Lips, Nails, Makeup 순으로 빈도가 높다.
- 대분류가 Free Items인 경우, 소분류 항목은 총 2개이며 Free Gifts의 빈도가 Vouchers 보다 압도적으로 많다.
- 대분류가 Sporting goods인 경우, 소분류 항목은 총 2개이며 Sports Equipment, Wirstbands 순으로 빈도가 높다.
- 대분류가 Home인 경우, 소분류 항목은 1개이며 이 1개 항목의 빈도 역시 1 뿐이다.

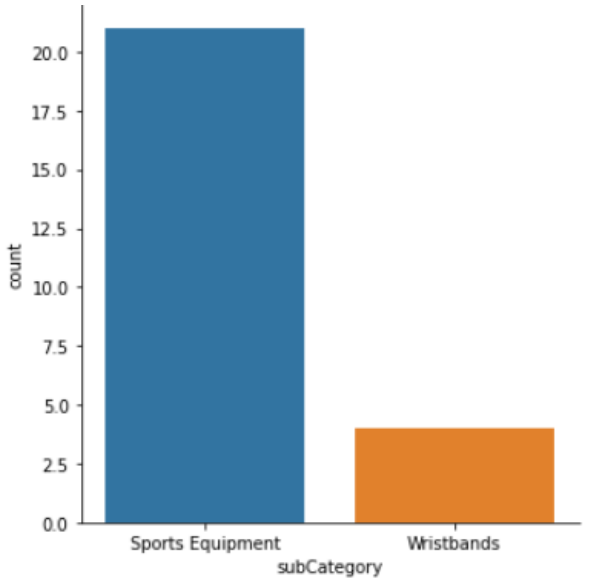
대분류 Personal Care의 소분류



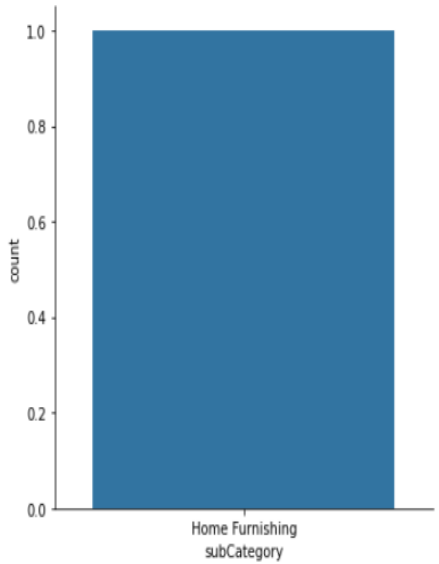

대분류 Free Items의 소분류




대분류 Sporting goods의 소분류




대분류 Home의 소분류



# 탐색적 데이터 분석



## 성별(gender)에 따른 옷의 주 색상 분포 차이가 있을 것이라는 가정하에 데이터 탐색 진행

- 색상은 총 46가지 존재. 각 색상 별로 성별의 비율을 계산한 후, 비율의 최대값이 0.8 이상인 색상을 아래 표와 같이 정리
- Charcoal과 Khaki 색상의 경우 Men이 80% 이상 비중을 차지하고 있다.
- 이외 Bronze, Gold, Lime Green, Magenta 등의 색상에서는 Women의 높은 비율을 갖는다.
- 성별(gender)에서 Men과 Women의 빈도가 상대적으로 많기 때문에 높은 비율 역시 해당 성별에서 관측이 된다.
- 따라서 baseColour 정보는 gender를 분류하는데 의미 있는 정보를 담고 있을 가능성이 높다.

| | baseColour | baseColour_Boys | baseColour_Girls | baseColour_Men | baseColour_Unisex | baseColour_Women |
|---|---|---|---|---|---|---|
| 3 | Bronze | 0.000000 | 0.000000 | 0.105263 | 0.000000 | 0.894737 |
| 6 | Charcoal | 0.004386 | 0.000000 | 0.807018 | 0.013158 | 0.175439 |
| 11 | Gold | 0.000000 | 0.001590 | 0.135135 | 0.020668 | 0.842607 |
| 15 | Khaki | 0.035971 | 0.000000 | 0.812950 | 0.064748 | 0.086331 |
| 17 | Lime Green | 0.000000 | 0.166667 | 0.000000 | 0.000000 | 0.833333 |
| 18 | Magenta | 0.000000 | 0.116279 | 0.023256 | 0.000000 | 0.860465 |
| 20 | Mauve | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 1.000000 |
| 22 | Multi | 0.010152 | 0.002538 | 0.137056 | 0.012690 | 0.837563 |
| 26 | Nude | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 1.000000 |
| 30 | Peach | 0.000000 | 0.061538 | 0.061538 | 0.005128 | 0.871795 |
| 34 | Rose | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 1.000000 |
| 36 | Sea Green | 0.000000 | 0.045455 | 0.045455 | 0.000000 | 0.909091 |
| 38 | Skin | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 1.000000 |
| 41 | Taupe | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 1.000000 |

**문제 정의 1.**

Fashion product image 데이터를 이용하여 gender, masterCategory, subCategory를 분류하는 개별 모델 개발

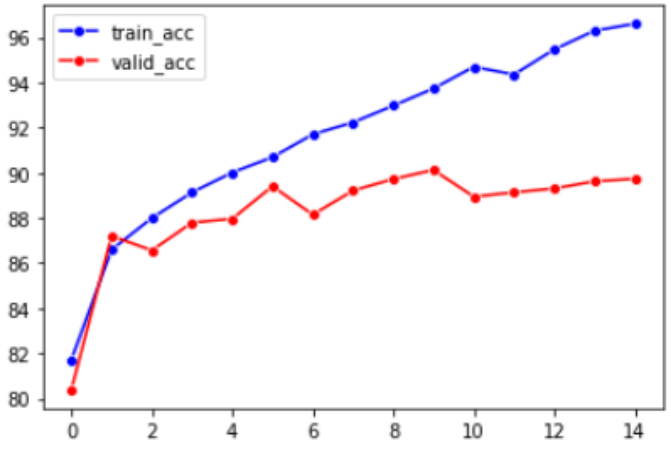
**문제 정의 2.**

masterCategory, subCategory의 계층적 특징을 고려한 Hierarchical Classification(HCM) 모델 개발

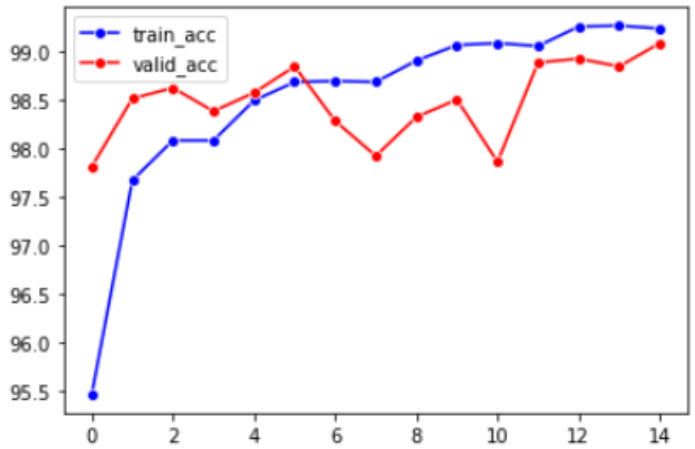
## gender, masterCategory, subCategory 분류 모델 개발 (학습)

- 사용 모형: Resnet50
- 모형 선택 사유: 이미지 분석에 흔히 사용하는 모형으로 baseline 성능을 확인하기 위해 사용하였으나
  3가지 target에 대한 모델 모두 우수한 성능을 보여 최종 선택
- 학습 방법
  1) Train/Validation/Test 는 6:2:2로 분할.
  2) pretrained weight를 이용하여 전체 네트워크에 대해 transfer learning (15 epoch)
- 학습 결과
  3개 모델 모두 validation 정확도 90% 이상 기록하였으며, masterCategory의 경우 과적합 없이 굉장히 우수한 성능 보임

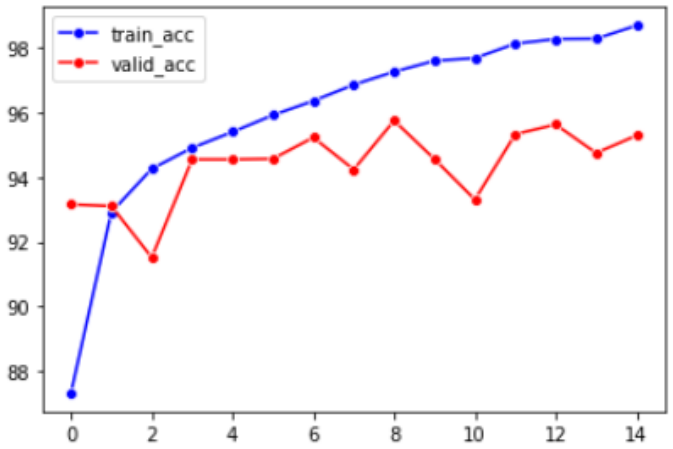
Gender 모델 학습 History




masterCategory 모델 학습 History

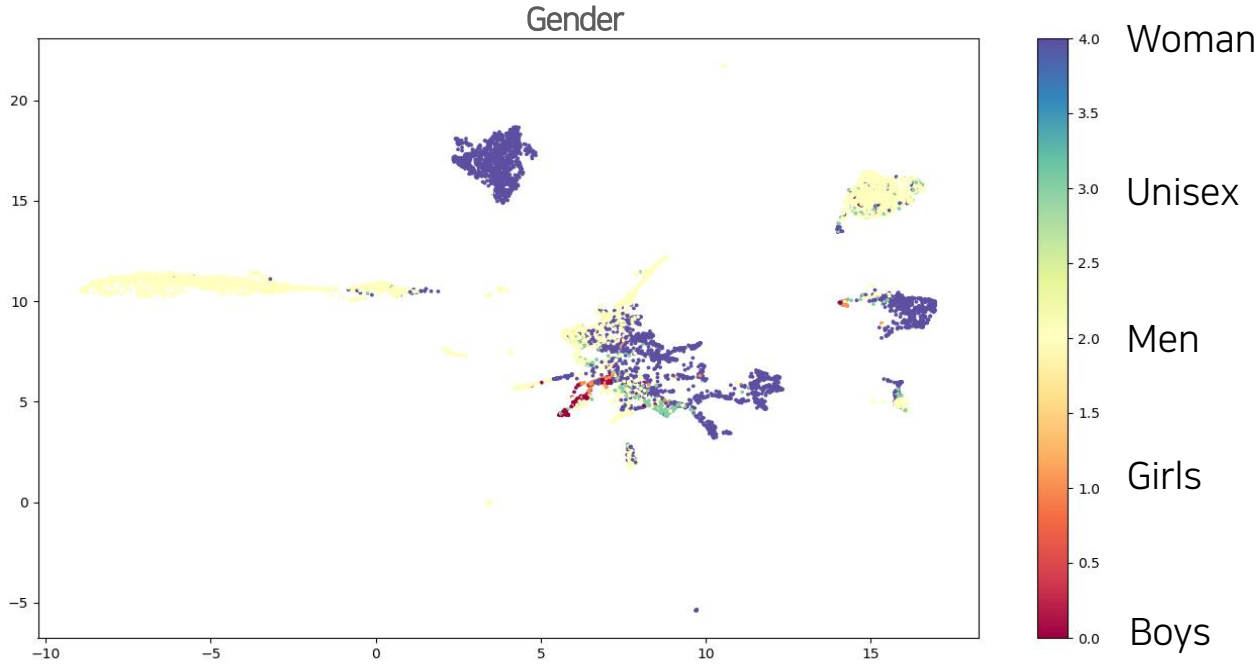



subCategory 모델 학습 History



## gender, masterCategory, subCategory 분류 모델 개발 (추론)

- (test data) 추론 결과
  1) gender → Accuracy: 89.57%, F1-score: 0.7048
  2) masterCategory → Accuracy: 99.01%, F1-score: 0.9722
  2) subCategory → Accuracy: 94.68%, F1-score: 0.8667
- UMAP[2]을 이용하여 test data의 100,352 차원 embedding vector를 2차원으로 표현 후 시각화
  → Men과 Women의 경우 외곽 지역에서 cluster를 잘 형성하고 있지만, 중심부에는 서로 다른 class의 데이터가 혼재되어 있음



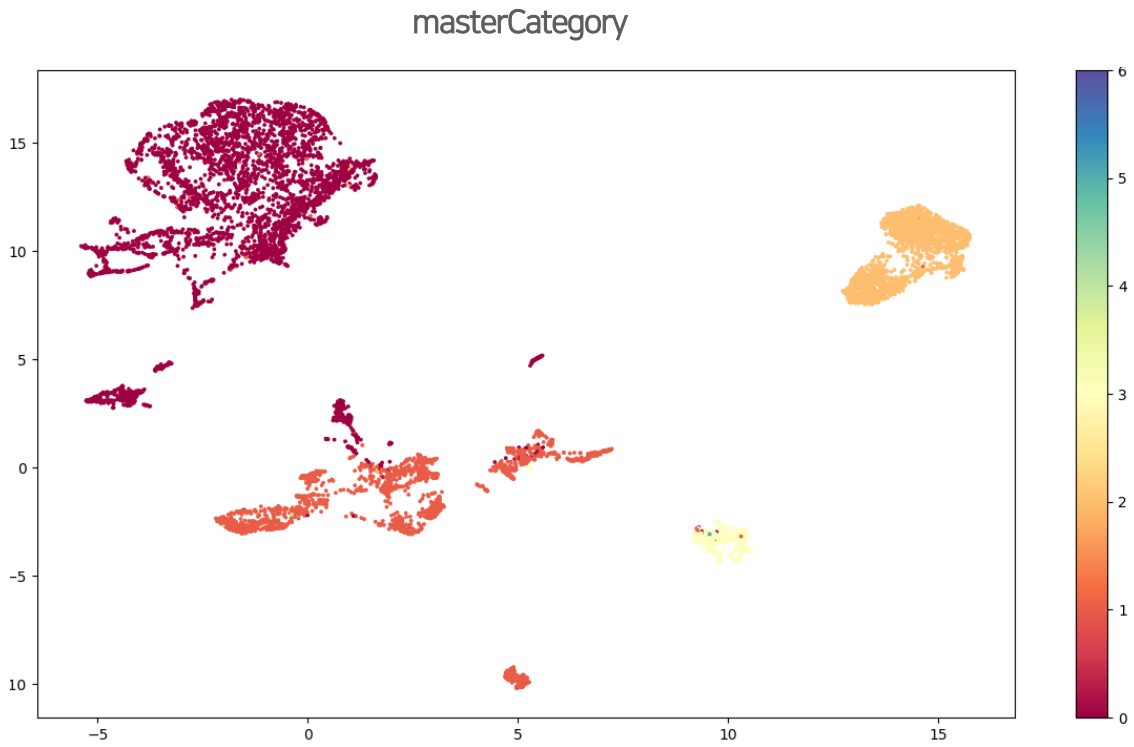

[2] McInnes, Leland, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction." *arXiv preprint arXiv:1802.03426* (2018)..

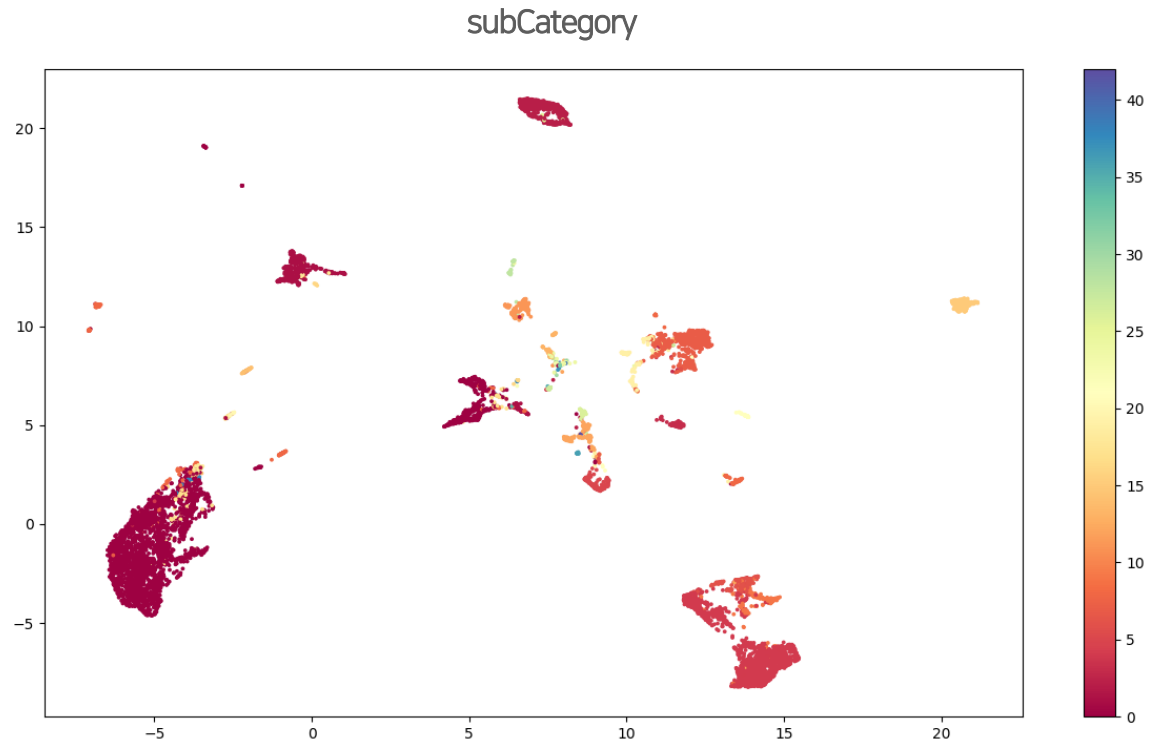
# AI 모델링

2) 분석 과정

## gender, masterCategory, subCategory 분류 모델 개발 (추론)

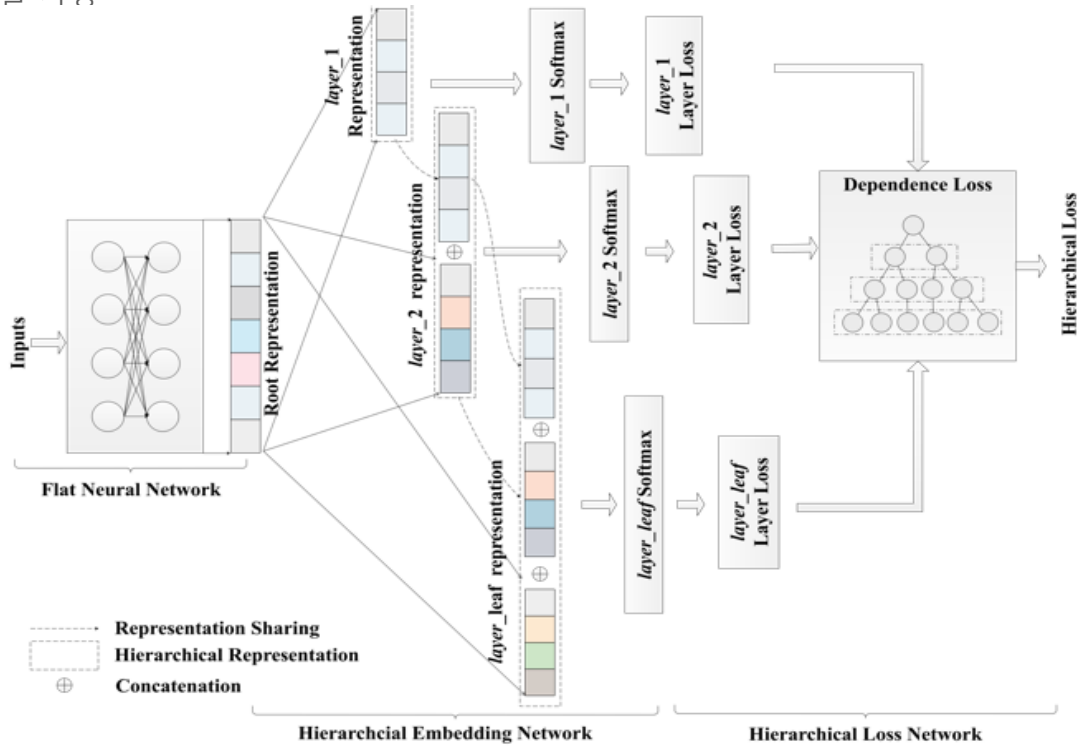
- UMAP을 이용하여 test data의 100,352 차원 embedding vector를 2차원으로 표현 후 시각화
  → masterCategory의 경우 7가지 class 모두 각각 cluster를 잘 형성하고 있음을 확인
  → subCategory의 경우 전반적으로 class별로 cluster를 잘 형성하고 있으나, 일부 minor class의 경우 다른 class와 혼재 되어 있는 경우가 있음



## masterCategory, subCategory 계층 분류 모델(HCM) 개발

- 사용 모형: Deep Hierarchical classification model[3)]
- 모형 선택 사유: masterCategory, subCategory의 계층적 특징을 고려하여 2가지 class에 대해 한번에 학습 가능
  일반적인 multi-label classification task로 접근 시 두 class의 계층 구조가 어긋나는 결과를 만들 수 있음.
- 아키텍처 구조: Flat Neural Network(FNN), Hierarchical Embedding Network(HEN), Hierarchical Loss Network(HLN)으로 구성
  → FNN으로 Resnet50을 사용




[3)] Gao, Dehong, et al. "Deep hierarchical classification for category prediction in E-commerce system." *arXiv preprint arXiv:2005.06692* (2020).

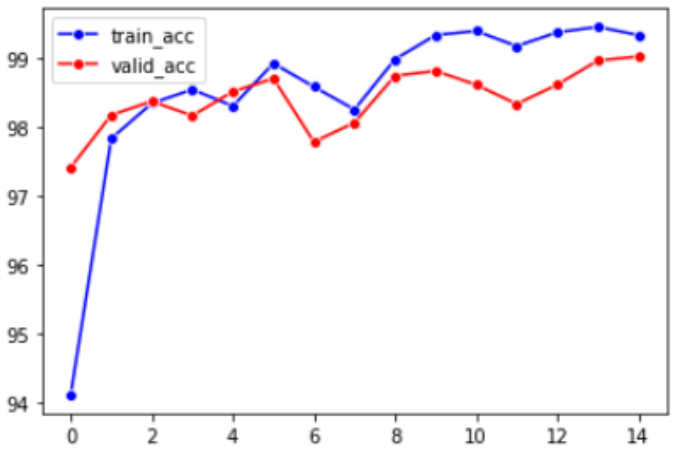
# AI 모델링



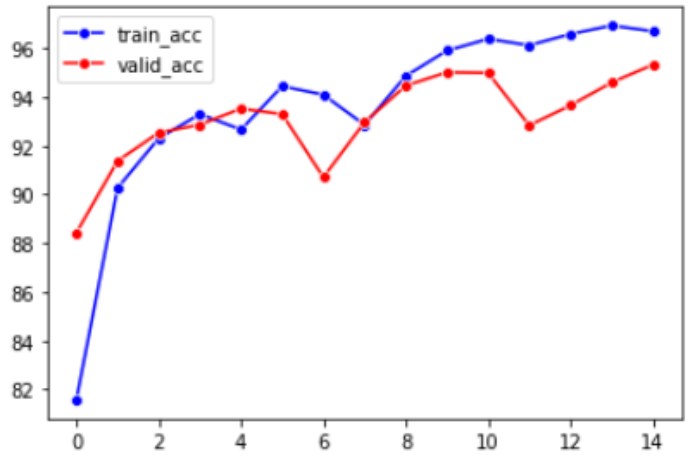
## masterCategory, subCategory 계층 분류 모델(HCM) 개발

- 학습 결과
  masterCategory, subCategory 모두 과적합 없이 학습 진행, subCategory의 경우 개별 모델 대비 과적합 정도가 개선된 것으로 보임
- (test data) 추론 결과
  1) masterCategory → Accuracy: 99.1%, F1-score: 0.9659
  2) subCategory → Accuracy: 94.88%, F1-score: 0.8601
  → 성능적으로 개별 모델과 유의한 차이가 없음

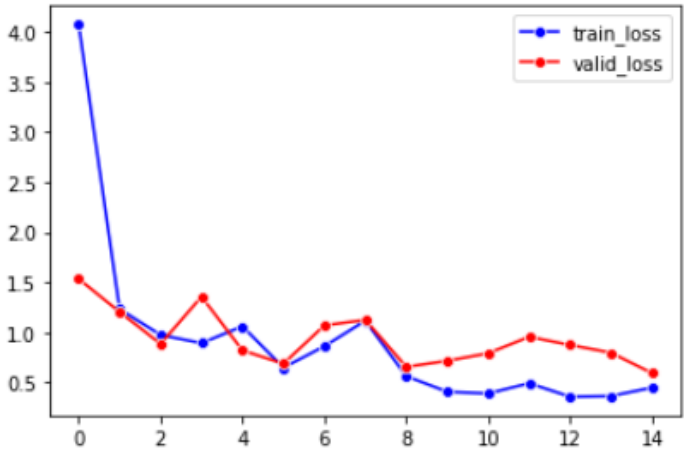
masterCategory 모델 학습 History




subCategory 모델 학습 History




HLN loss History



**결과 정리**

1) Resnet50 기반 3가지 개별 모델 모두 정확도 및 F1-score에서 우수한 성능을 보임
2) UMAP을 통해 계산한 저차원 Embedding을 보았을 때, class별로 cluster를 잘 형성하는 것을 통해 학습이 대체로 잘 이루어졌음을 확인
3) HCM의 경우 개별 모델과 성능이 비슷하며, 일부 과적합이 완화되어 보이는 결과가 나타남

**향후 개선점**

1) baseColour 정보를 추가하여 gender 분류모형의 성능이 개선되는지 확인 필요(Multi-modal learning)
2) class imbalance 문제에 대한 추가적인 분석 필요
3) GradCAM 등의 방법을 이용하여 data instance level의 결과 해석
4) HCM 모델에 gender class를 학습할 수 있는 layer를 추가하여 3가지 class 모두 학습할 수 있는 네트워크 구성