COSE362 Final-term Report

17조
2022320057 강우혁
2022320120 차승민
2022320157 노규주

1. 문제 정의

흡연이나 음주는 우리 몸에 어떠한 영향을 미치는가? 우리는 주변에서 흡연 또는 음주를 하는 사람들을 어렵지 않게 찾아볼 수 있다. 대부분의 사람들은 흡연 또는 음주를 하면서 이들이 우리 몸에 해롭다고 알고 있으나 술과 담배가 실제로 해로운지, 해롭다면 정확히 어떤 부분에서 안 좋은 영향을 미치는지는 자세히 알지 못한다.

이번 Term Project에서는 흡연이나 음주가 우리 몸에 어떤 영향을 미치는지 다양한 body signal들을 지표로 하여 분석하는 것을 목표로 한다. 성별, 나이 그리고 혈압, 단백뇨, 간 수치 등의 body signal들과 흡연 상태, 음주 여부를 포함하는 데이터셋을 이용해 사람들을 그룹화하여 각 집단들을 관찰 및 분석하고 이를 토대로 흡연 여부와 음주 여부가 그룹을 나누는 기준에 크게 영향을 미치는지, 만약 그렇다면 흡연이나 음주 상태에 따라서 사람들의 건강 상태에 어떤 차이가 존재하는지 분석할 것이다.

아래는 학습 모델에 사용된 kaggle 데이터셋이다.
kaggle datasets download -d sooyoungher/smoking-drinking-dataset

2. 방법론

1) 데이터 설명

kaggle 에서 받아온 데이터셋은 아래와 같은 column들을 포함한다.

sex(성별)
age(나이)
height(키)
weight(몸무게)
waistline(허리둘레)
sight_left/right(좌우 시력)
hear_left/right(좌우 청력)
SBP(수축기 혈압) / DBP(이완기 혈압)
BLDS(공복혈당)
hemoglobin(헤모글로빈)
urine_protein(요단백 또는 단백뇨) / serum_creatinine(혈청 크레아티닌)
triglyceride(중성지방) / tot_chole(총 콜레스테롤) / HDL_chole / LDL_chole
SGOT_AST / SGOT_ALT / gamma_GTP
SMK_stat_type_cd(흡연 상태)
DRK_YN(음주 여부)

모델 학습을 위한 데이터 전처리 과정은 방법론과 함께 설명하겠다.

2) 방법론 설명

흡연이나 음주 상태와 사람의 건강 상태 사이의 관계를 분석하기 위해 clustering, 그 중에서 k-means algorithm을 선택하였다. (midterm에서 언급한대로 agglomerative clustering, spectral clustering, 마지막으로 dbscan까지 추가적인 3가지를 구현하려고 시도했으나, colab 자체의 ram이 모든 방법에서 다운되어 구현에 실패했다.) 만약 흡연 혹은 음주 여부가 영향을 미친다면, 데이터셋을 이용해 사람들을 여러 그룹으로 나눴을 때 그에 따른 그룹화 현상이 나타날 것이라는 가설에서 출발하였다. clustering을 이용하여 데이터를 군집화하면 body signal에 따른 여러 집단을 분류할 수 있고 이를 바탕으로 사람들의 건강 상태를 분석하기 용이할 것이라 생각하였다.

즉, ‘흡연 혹은 음주가 실제로 우리 몸에 크게 영향을 준다면, 흡연이나 음주를 하는 사람들의 비율이 높은 그룹이 존재할 것’이라고 예측하였으며, 이 그룹(들)을 관찰하여 다른 군집과 유의미한 차이가 나는 feature들을 관찰할 수 있을 것이다. 또한 새로운 건강 상태 데이터들을 training한 모델에서 어떠한 군집에 속하는지 확인한 후, 군집의 특징을 관찰할 수 있다. 또한, 우리가 사용하는 데이터셋을 통해 얻을 수 있는 한 사람에 대한 column의 수는 수십 개가 존재한다. 모든 column들을 사용한다면, dimension의 수가 커져 연산이 복잡해질 우려가 있다. 따라서 데이터 전처리를 통해 일차적으로 데이터를 다루기 편한 방향으로 처리하고, PCA로 차원을 축소시킨 후 구현이 비교적 쉽고 계산 속도가 빠른 k-means algorithm을 적용하였다.

clustering을 진행하기 앞서, 우선 데이터 전처리 과정을 거쳤다. 시력과 청력은 흡연과 음주와 무관하다고 판단하여 feature에서 제거했다. 그리고 혈압은 평균값을 산출하여 pressure로 변수 이름을 바꾸었다. 가독성을 높이기 위해 BLDS(공복혈당)는 blood_sugar로 변수명을 재설정하였다. 콜레스테롤과 관련된 HDL_chole, LDL_chole, triglyceride는 tot_chole(총 콜레스테롤)만 남기고 제거하였다. 마찬가지로, 간 수치 지표인 SGOT_AST와 SGOT_ALT는 gamma_GTP만 남기고 제거해주었다.

카테고리컬한 데이터인 성별과 음주 여부/흡연 여부에 대하여 성별은 male은 1, female은 0으로 mapping 해주었다. DRK_YN(음주 여부)도 역시 yes면 1, no면 0으로 바꿔주었다. SMK_stat_type_cd(흡연 상태)는 총 세 가지로 나뉜다. 기존 데이터셋에서는 비흡연자의 경우 1, 흡연했었으나 끊은 경우 2, 흡연자의 경우 3으로 표현하였으나 일단은 각각의 값에서 2를 빼주어 -1, 0, 1로 처리하였다.

데이터를 둘러보던 중 waistline 항목에서 999라는 수치가 발견되었다. 다른 데이터에서도 이상치는 존재하였으나, 정상적인 수치와 완전히 동떨어지지 않았으며 인간의 건강 상태라는 데이터 특성상 ‘충분히 존재할 수 있는’ 이상치로 판단하여 제거하지 않기로 하였다. 그러나, waistline에서 999라는 수치는 999를 제외하고 가장 높은 값이 149.1인 상황에서 약

1백만 개의 데이터 중 999라는 값이 57개의 row에서 검측된 것으로 보아 명백한 outlier이며, 허리 둘레가 999일 경우 지름이 약 3미터로 계산되는 것도 고려하여 57개의 row를 제거하기로 했다. 999라는 값을 다른 값으로 imputation하는 것은 어떻겠냐는 의견이 있었으나,약 1백만 개 중 57개 밖에 되지 않기에 제거하는 것으로 결론지었다. 999 이상치를 제거하고 다시 correlation을 계산한 결과 weight와 waistline 두 feature의 correlation이 0.788103으로 나와서 두 개의 feature 중 하나만 남기는 방향으로 고려했으나, clustering을 진행한 결과를 분석하는 중 상대적으로 weight는 낮지만 waistline이 높은 group이 발견되었고, 이는 내장비만과 관련이 있다고 판단하여 둘 다 사용하기로 결정했다.

```
[4]  #이상치 제거
     ex['waistline'] = ex['waistline'].replace(999, float('nan'))
     ex = ex.dropna()
```

〈Figure 1. 이상치 처리 과정〉

	sex	age	height	weight	waistline	sight_left	sight_right	hear_left	\
0	Male	35	170	75	90.0	1.0	1.0	1.0	
1	Male	30	180	80	89.0	0.9	1.2	1.0	
2	Male	40	165	75	91.0	1.2	1.5	1.0	
3	Male	50	175	80	91.0	1.5	1.2	1.0	
4	Male	50	165	60	80.0	1.0	1.2	1.0	

	hear_right	SBP	...	LDL_chole	triglyceride	hemoglobin	urine_protein	\
0	1.0	120.0	...	126.0	92.0	17.1	1.0	
1	1.0	130.0	...	148.0	121.0	15.8	1.0	
2	1.0	120.0	...	74.0	104.0	15.8	1.0	
3	1.0	145.0	...	104.0	106.0	17.6	1.0	
4	1.0	138.0	...	117.0	104.0	13.8	1.0	

	serum_creatinine	SGOT_AST	SGOT_ALT	gamma_GTP	SMK_stat_type_cd	DRK_YN
0	1.0	21.0	35.0	40.0	1.0	Y
1	0.9	20.0	36.0	27.0	3.0	N
2	0.9	47.0	32.0	68.0	1.0	N
3	1.1	29.0	34.0	18.0	1.0	N
4	0.8	19.0	12.0	25.0	1.0	N

〈Figure 2. 데이터 전처리 전 columns〉

```
   sex  age  height  weight  waistline  pressure  blood_sugar  tot_chole  \
0    1   35     170      75       90.0     100.0         99.0      193.0
1    1   30     180      80       89.0     106.0        106.0      228.0
2    1   40     165      75       91.0      95.0         98.0      136.0
3    1   50     175      80       91.0     116.0         95.0      201.0
4    1   50     165      60       80.0     110.0        101.0      199.0

   hemoglobin  urine_protein  serum_creatinine  gamma_GTP  SMK_stat_type_cd  \
0        17.1            1.0               1.0       40.0              -1.0
1        15.8            1.0               0.9       27.0               1.0
2        15.8            1.0               0.9       68.0              -1.0
3        17.6            1.0               1.1       18.0              -1.0
4        13.8            1.0               0.8       25.0              -1.0

   DRK_YN
0       1
1       0
2       0
3       0
4       0
```

〈Figure 3. 데이터 전처리 후 columns〉

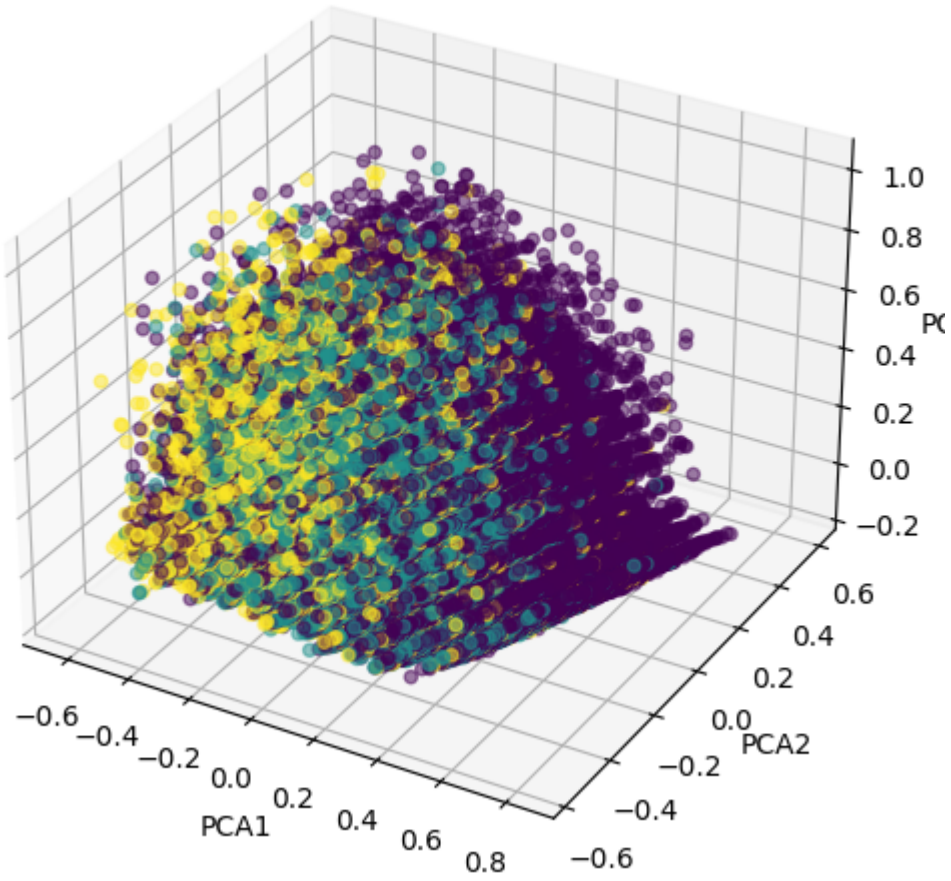
이후, k-means algorithm을 적용하기에 앞서 차원을 축소하기 위해 주성분 분석(이하 PCA)을 진행하였다. 우선, 각 column에 대한 수치를 정규화하였다. 각 차원마다 표준편차 값의 차이가 크기 때문에 PCA 적용 이전에 정규화는 반드시 거쳐야 하는 중요한 과정이다. 또한, 카테고리얼한 feature인 성별, 음주 여부, 흡연 여부는 제외하고 PCA를 수행하였다.

PCA에서 component의 개수 결정은 Proportion of Variance(이하 PoV)를 확인 후 결정했다. 수업에서는 보편적으로 PoV 값이 0.9 이상이 되는 component의 개수(이하 n)에서 멈춘다고 배웠으나, 두 가지 이유로 인해 n=3 으로 결정하였다. 첫째, 확인해본 결과 n=4 부터, 즉 네 번째 component부터는 PoV를 증가시키는데 있어서 큰 영향을 미치지 않았다. 둘째, PCA 결과와 후에 진행할 클러스터링 결과를 시각적으로 보여주는 데 있어 4차원 이상으로 진행하는 것은 어려울 것이라 판단하였다.

```
array([0.51758587, 0.24876695, 0.07125479])
```

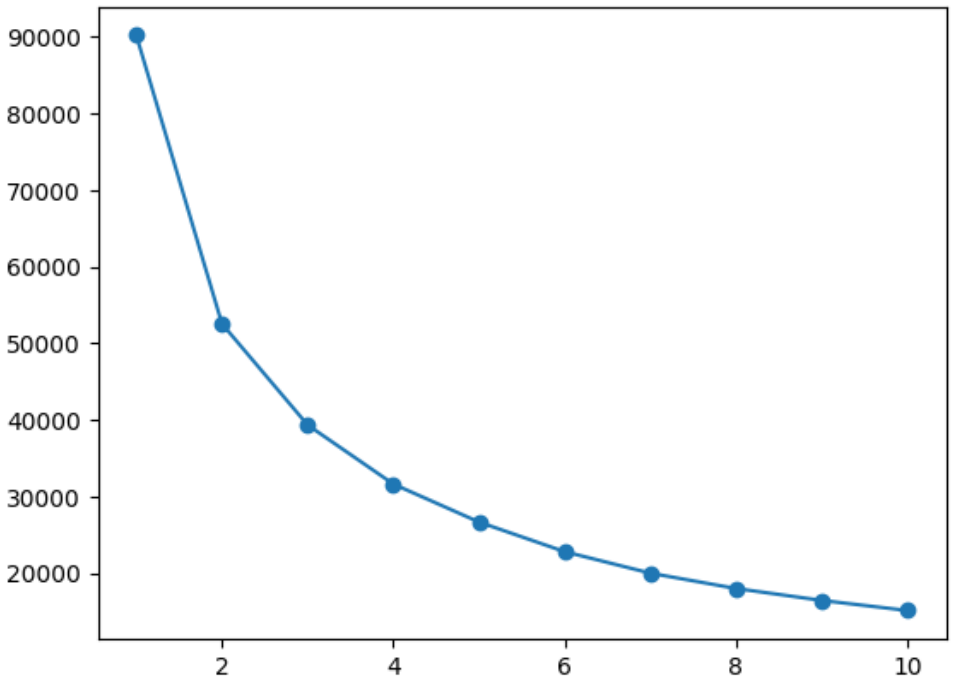
〈Feature 4. n번째 주성분의 전체 데이터를 설명하는 비율〉

위 과정으로 결정한 n값으로 전처리한 데이터에 PCA를 진행한 결과를 시각화한 결과가 아래와 같다. 데이터들의 색깔은 흡연 여부를 나타낸다.



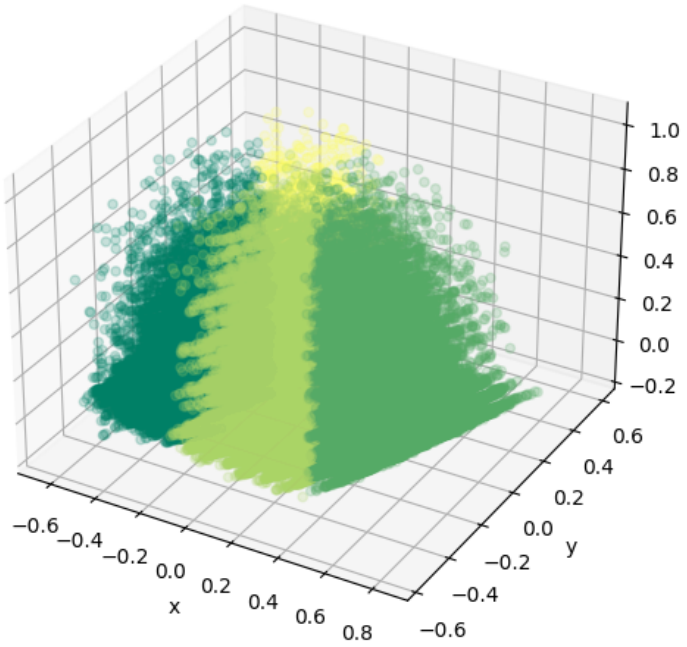
〈Feature 5. n=3으로 PCA 진행 후 시각화한 결과〉

이후 clustering 과정에서, 최적의 클러스터의 개수(이하 k)를 구하기 위해 1부터 10까지 cluster의 개수를 바꿔가며 inertia를 구했고, 이를 그래프로 나타낸 것이 다음과 같다.



〈Feature 6. 각 k값의 inertia〉

elbow로 보이는 k=4를 cluster의 개수로 결정하고 k-means Clustering을 진행하고 시각화한 결과는 아래와 같다. k-means Clustering 과정에서 sklearn 모듈을 이용했는데, 여기서 지원하는 k-means++를 적용해 centroid를 기존의 k-means보다 신중하게 선정하여 진행했다.



〈Feature 7. k=4로 clustering 후 시각화한 결과〉

3. 결과와 결과에 대한 해석

앞서 설명한 방법론을 토대로 진행한 결과에서 cluster 각각의 feature들의 평균을 확인해보았다. 그 전에, 흡연했었으나 끊은 경우와 흡연자를 묶어 1로, 비흡연자를 0으로 다시 mapping한 후 확인했다.

cluster	sex	age	height	weight	waistline	pressure	blood_sugar	tot_chole	hemoglobin
0	0.935175	33.902055	172.403919	75.151167	85.709674	100.340401	98.145406	197.518600	15.416832
1	0.249031	64.470252	153.117692	55.484630	80.594518	101.267820	103.604832	195.509192	13.442186
2	0.855506	53.759625	166.549335	69.400620	86.175230	102.817242	107.276910	196.914271	14.937173
3	0.118204	38.724900	157.353652	53.987347	72.952455	93.053624	93.044062	192.571303	13.215818

〈Figure 8. cluster별 feature들의 mean〉

urine_protein	serum_creatinine	gamma_GTP	SMK_stat_type_cd	DRK_YN	x	y	z
1.080833	0.948062	46.474920	0.643105	0.737287	-0.285200	-0.077687	0.001029
1.109703	0.820630	30.125493	0.170846	0.207890	0.309940	0.011702	0.002296
1.115553	0.940295	50.603455	0.632384	0.603912	0.034770	-0.135667	-0.004570
1.071757	0.742916	22.345359	0.140774	0.452957	-0.059765	0.187391	0.001402

〈Figure 9. cluster별 feature들의 mean〉

결과 해석에 앞서, 흡연 여부와 음주 여부 그리고 성별을 0과 1로 나누었기 때문에 각 cluster 마다 mean 값은 흡연자/술을 마시는 사람/남성의 비율과 같다.

흡연자들이 치우친 그룹은 cluster 1과 cluster 2(이하 흡연자 그룹)로 둘 다 그룹의 약 63%를 흡연자가 차지하고 있었다. 반대로 cluster 0과 cluster 3(이하 비흡연자 그룹)에서는 각각 약 17%, 약 13%가 흡연자로 분류되었다.

흡연율을 기준으로 흡연자 그룹과 비흡연자 그룹으로 이분했을 때 차이가 나는 feature들은 성별, serum_creatinine, gamma_GTP(이하 간 수치) 그리고 음주 여부가 눈에 띈다.

1) 성별과의 관계

흡연자 그룹은 남성이 대부분을 차지하고 있다. 이는 우리나라 흡연자 그룹 내에서의 성별 비율은 남성이 여성보다 더 높다는 것을 고려하면 타당하게 보인다.

2) serum_creatinine과의 관계

serum_creatinine의 mean 값도 흡연자 그룹에서 더 높았다. 그러나 이는 크레아틴이 근육에서 합성되기 때문에 근육량이 더 높은 남성 그룹(1에서 알 수 있듯이 흡연자 그룹)에서 높은 수치가 나오는 것이 타당하다고 보인다. 키 역시 마찬가지 이유로 남성 그룹에서 mean 값이 더 크다.

3) 간 수치, 음주 여부와의 관계

흡연자 그룹은 비흡연자 그룹에 비해 간 수치와 음주 비율 역시 높게 나타났다. 음주는 간 수치에 직접적인 영향을 미치는 원인으로 알려져 있기에 당연한 결과라고 판단하였다.

그러나, 흥미로운 점은 흡연자 그룹 내의 cluster 1보다 cluster 2에서 간 수치가 더 높은데 반해, 두 그룹의 몸무게와 허리둘레의 mean 값을 살펴봤을 때 허리둘레는 거의 유사했으나 몸무게는 cluster 2에서 더 낮은 것을 확인할 수 있었다. 이는 간 기능이 저하되어 내장 비만이 심화되는 현상과 관련이 있을 수 있다고 해석하였다.

4. 향후 가능한 추가 연구 개발 방향

혹은 기존에 사용한 clustering이 아니라, supervised learning을 이용한 완전히 새로운 방법론을 시도해볼 수 있다. 우리가 분석하고자 하는 데이터셋의 column들 중에는 평가하기에 적합한 output 값이 뚜렷하게 보이지 않아 clustering으로 진행하긴 하였으나, 흡연 상태나 음주 여부를 classification 하여 새로운 건강 상태 데이터가 들어왔을 때 흡연을 하는 사람인지 하지 않는 사람인지, 음주를 하는 사람인지 하지 않는 사람인지 분류하는 모델을 만들어보는 것 또한 이 데이터셋을 활용하는 방법 중 하나가 될 것이다.

5. 마치며

사람의 건강이란, 사람의 시선에선 예측하기 힘든 데이터이다. 담배를 즐기는 사람이라 해서 폐가 꼭 나쁜 것도 아니며, 술을 즐기는 사람이라 해서 간이 꼭 나쁜 것도 아니다. 물론 사회의 관념이 말하기에, 술과 간, 담배와 폐는 선형이라고 하지만, 그 관념이 확증 편향인 게 아니라고 자부할 수 있냐고 하기에, 한 조원은 너무 염세적이고 세상을 믿지 못했지만, 다행히도 그 조원은 사람의 시선에서 예측하기 힘든 데이터를 사람이 아닌 존재의 시선에서 예측할 수 있는 시대에 태어났다. 그 조원이 말하길, 머신 러닝이란 물론 사람이 직관적으로 할 수 있는 것을 더 수월하고 편리하게 만들어 주기도 하지만, 비과학을 정량화 해 직관의 영역으로 끌어내려주는 것이라 제 스스로 생각한다 했다. 이 코드들은 물론 전문가들의 코드와 비교하기엔 그 구성이 부족할 지 모르지만, 상기한 목적을 수행했다는 점에서 훌륭한 머신러닝이 아니냐 했다.

그는 이 코드를 아낀다 했다. 코드를 짜기 전 구상부터 코드를 짜는 중 생긴 문제까지 모두 포함해 아낀다 했다. 더 나아가, 같이 코드를 짜며 서로가 보지 못한 관점을 끌어와 준 각자에게, 그리고 이런 프로젝트의 마중물이 되어 준 이 수업에 감사하다 했다. 그리고 이런 감사함을 final term report에 '마치며'라는 항목을 추가하면서까지 알리고 싶다 했다. 이 감사함이 전해졌길 바란다.