국민 건강상태 증진을 위한 개인별 운동 추천시스템

정시원, 배수근, 임선우, 주원우

소속: 부산대학교 통계학과

『2024 문화 디지털 혁신 및 문화데이터 활용 공모전 기획서』 - 데이터 분석 -		
공모분야	데이터분석	
참가자 구분	☑일반부(19세 이상)	☐청소년부(13~18세)
분석 주제	국민 건강상태 증진을 위한 개인별 운동 추천시스템	
분석 툴(Tool)	☑Python ☑R ☐Tableau ☐기타( )	

Contents

1) 분석주제: 국민 건강상태 증진을 위한 개인별 운동 추천시스템

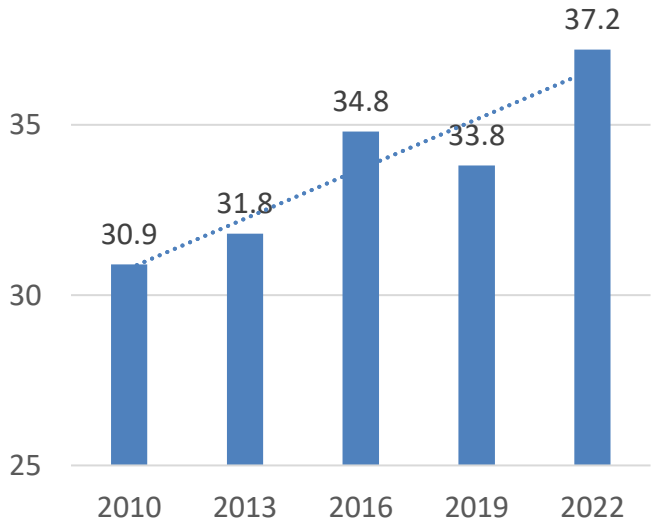
최근 들어 운동에 대한 관심도가 점차 높아지고 있는 반면, 국민의 건강상태는 이를 반영하지 않고 있다. 이에 본 분석에서는 기존의 운동 추천시스템 모델과 달리 개인의 신체 정보에 혈압, 체지방률 등을 반영하여 운동이 오히려 독이 될까 우려했던 사람들에게 보다 개인화된 추천을 제공한다. 본 분석이 제안하는 운동 추천 시스템은 K-mode를 통한 군집화와 협업 필터링 중 특이값 분해(SVD) 알고리즘의 잠재 요인 모델을 기반으로 설계하고, 정확도(Accuracy)를 통해 모델 구현의 합리성을 검정하였다. 이후 타 문화데이터와의 융합을 통해 추천 시스템의 확장을 기대하며 이로 인해 사용자들은 보다 효과적으로 운동을 선택할 수 있을 것이다.

2) 분석 배경 및 목적

"[표 1] 국내 전체비만율 추이'의 그래프를 확인해보면 2010년 이후로 **약 7% 정도의 비만율**이 **상승**한 것을 확인할 수 있다. 이는 국민의 건강상태의 향방이 다소 부정적임을 시사한다. 이에, 국민 건강 증진에 대한 적극적인 노력과 관심이 필요한 때라고 볼 수 있다.

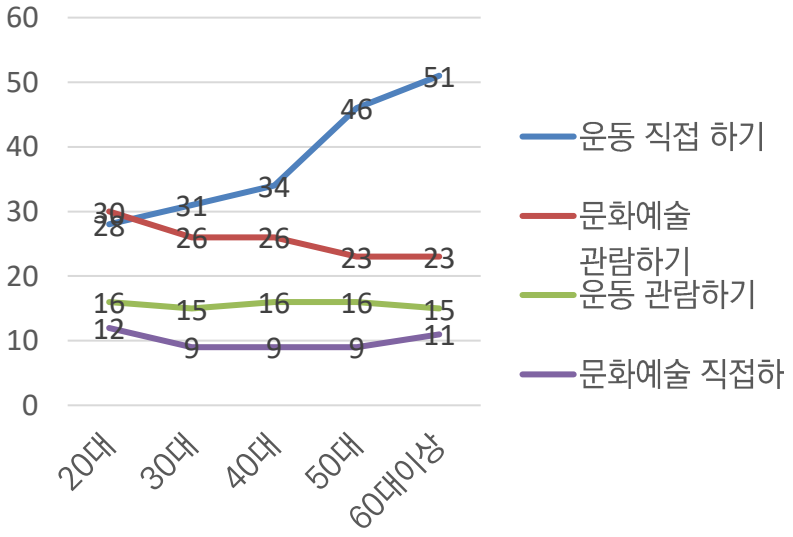
반면에, '[표 2] 문화 예술 활동 측면 여가활동 관심도 연령별 추이'를 보면 각 연령대별로 가장 **관심도가 높은 것은 '운동 직접하기'**라고 할 수 있고, 이는 국민들이 운동에 큰 관심을 가지고 있음을 의미한다.

[표 1] 국내 전체비만율 추이



출처: 국가발전지표, 지표누리

[표 2] 문화 예술 활동 측면 여가활동 관심도 연령별 추이



* 6개월간(2021년 10월~2022년 3월) 1만1281명을 대상

출처: 컨슈머인사이트

이러한 상황에서 개인별 운동 추천시스템을 제공한다면 **현재의 관심도를 그대로 실행으로 연결시켜주는 일종의 징검다리 역할**을 할 수 있을 것으로 생각한다. 본 연구는 체력측정 및 운동처방 종합 데이터의 연령, 신장, 몸무게, 혈압, BMI, 운동처방결과 등을 활용하여 진행하였고, 최종적으로는 **사용자가 본인의 신체정보를 입력하면 그에 적합한 운동을 추천하는 것을 목표**로 한다.

3) 분석 내용 및 결과

(1)데이터 전처리

체력측정 및 운동처방 종합 데이터의 2020년 1월 csv 파일부터 2023년 4월까지의 모든 파일을 R에서 dplyr package를 활용하여 병합하였다. 병합에서 R을 사용한 이유는 쉬운 데이터 조작과 데이터 프레임을 조작하고 병합하는 작업이 직관이어서 활용하였다. 이후 분석의 모든 과정에서 python 툴을 활용하였고, 본 분석에서 사용한 라이브러리는 sklearn, pandas, numpy, seaborn, kmodes, matplotlib, collection를 사용하였다.

변수선택

변수의 경우 51개의 변수 중 대다수의 값이 결측치인 변수는 사용할 수 없다고 판단하여 변수선택 과정에서 제외하였다. 또한, 상대악력에 대해 좌수악력과 우수악력을 나타내는 변수는 상관계수가 높게 나오기에, 마찬가지로 제외하였다. **최종적으로 '[표 4] 변수설명표'과 같이 12개의 변수를 선택하였다.**

[표 3] 악력 상관계수

	좌수악력	우수악력	상대악력
좌수악력	1	0.94	0.75
우수악력	0.94	1	0.79
상대악력	0.75	0.78	1

[표 4] 변수설명표

Column 명	변경한 변수 명	설명	데이터타입	Column 명	변경한 변수 명	설명	데이터타입
AGRDE_FLAG_NM	AGE_FLAG	연령대구분명	VARCHAR	MESURE_IEM_003_VALUE	Body_Fat	측정항목_3값 : 체지방율(%)	VARCHAR
MESURE_AGE_CO	AGE	측정연령수	DECIMAL	MESURE_IEM_004_VALUE	Waist	측정항목_4값 : 허리둘레(cm)	VARCHAR
SEXDSTN_FLAG_CD	SEX	성별구분코드	VARCHAR	MESURE_IEM_005_VALUE	Low_BP	측정항목_5값 : 이완기최저혈압(mmHg)	VARCHAR
MESURE_IEM_001_VALUE	Height	측정항목_1값 : 신장(cm)	VARCHAR	MESURE_IEM_006_VALUE	High_BP	측정항목_6값 : 수축기최고혈압(mmHg)	VARCHAR
MESURE_IEM_002_VALUE	Weight	측정항목_2값 : 체중(kg)	VARCHAR	MESURE_IEM_028_VALUE	Relative_Grap_strength	측정항목_28값 : 상대악력(%)	VARCHAR
MESURE_IEM_018_VALUE	BMI	측정항목_18값 : BMI(kg/㎡)	VARCHAR	MVM_PRSCRPTN_CN	Exercise_PRSCRPTN	운동처방내용	CLOB

유소년의 분리와 NA 처리

AGE_FLAG 열이 유소년인 경우, 체지방률이 무의미하여 측정하지 않아 1개의 행을 제외하고 모두 NA값으로 표시된다. 추후 체지방 칼럼의 경우, 범주형 데이터로 변환 예정이기에 유소년의 NA를 모두 0으로 바꾸었다. 또한, 원본 데이터셋의 NA값들에 대해서는 신체정보이기에 보간을 하기보다는 제거하는 게 타당하므로 **결측치를 보유하고 있는 행들에 대해서는 제거**하였다.

이상치 처리

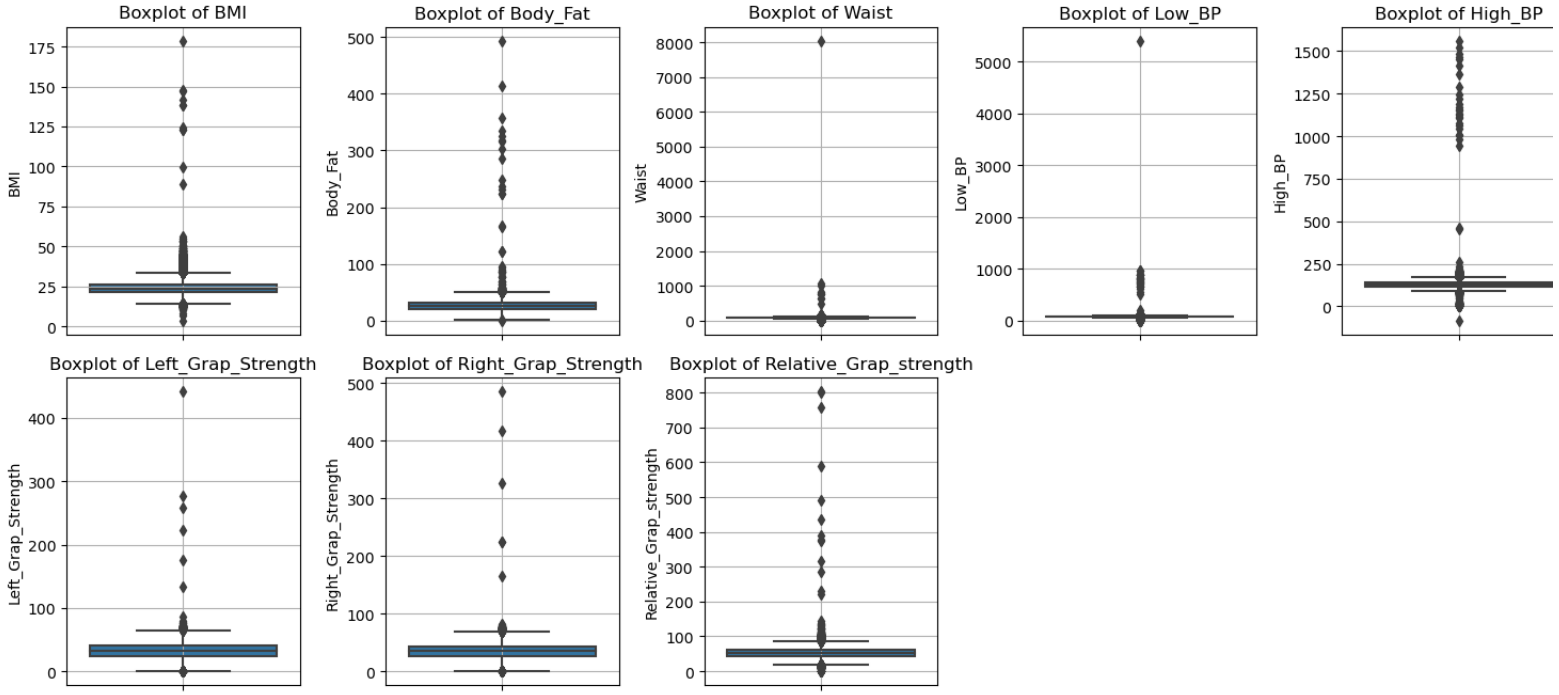
수치형 칼럼에 대해 Box-plot을 그려보면 이상치가 다수 있는 것을 확인할 수 있다. 이에, **[하한선(Q1-1.5 X IQR), 상한선(Q3+1.5 X IQR)]** 구간 내의 값들만 포함하도록 이상치를 처리하였다.

범주화

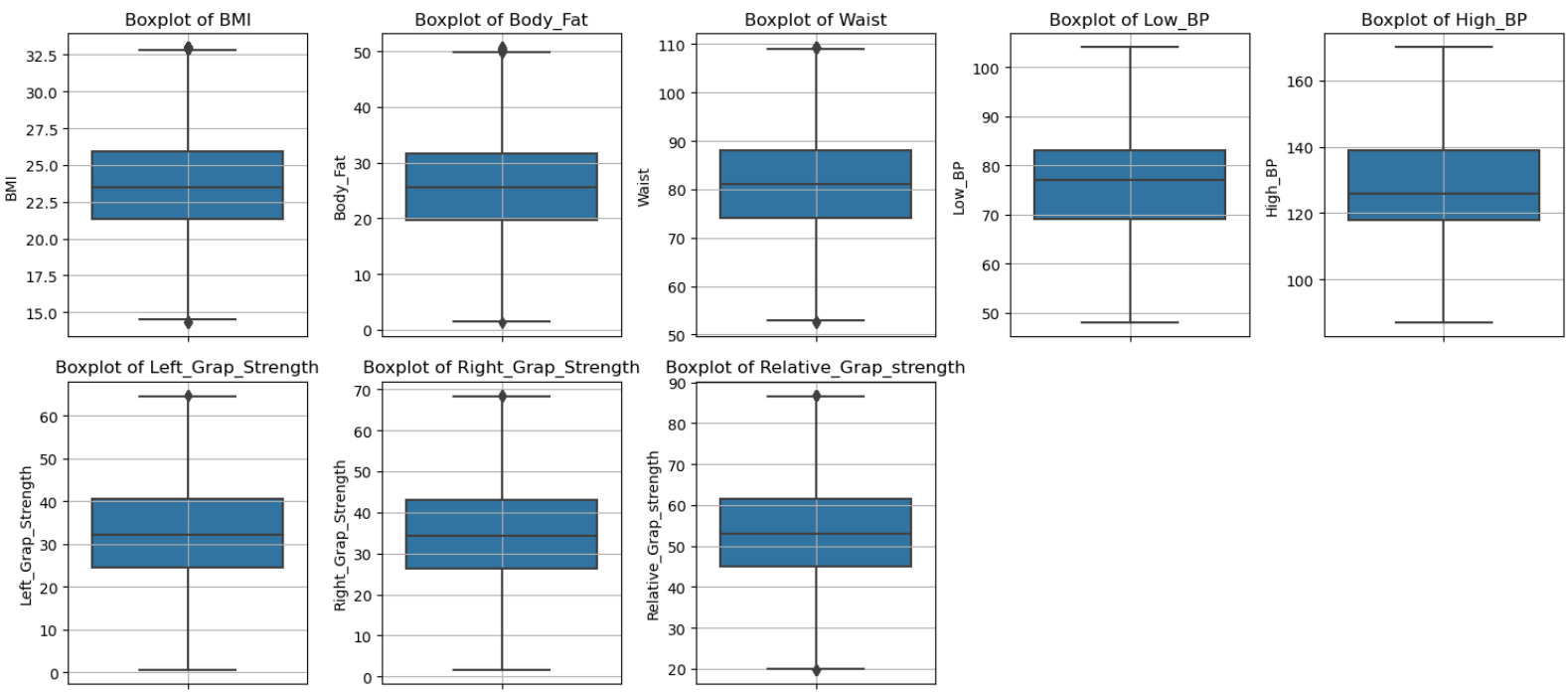
BMI, Blood_Pressure(혈압), Body_Fat(체지방률)과 같이 **사회에서 통념적으로 범주화**가 되어있는 열들은 해당 기준을 사용하였고, 나머지는 **등구간 범주화**를 진행하였다. 등구간 범주화를 진행함에 있어 describe 함수를 사용하여 평균, 표준편차, 최대, 최소, 1,3분위수를 참고하여 **비율적으로도 일부 조정**하였다.

전처리가 끝난 후 최종 데이터의 사이즈는 270245 X 9이다.

[그림 1] 이상치 처리 전 Boxplot



[그림 2] 이상치 처리 후 Boxplot



(2) 군집화 (K-modes)

군집화의 목적

왜 하는가? 현재 전처리 이후 약 27만 개의 행이 생성되어 있다. 각 행에 대해 유사도를 비교할 시 막대한 계산량으로 인해 추천을 받고자 하는 데에 시간이 많이 소모된다. 이에, 계산량과 추천 운동 결과를 도출해내는 데까지의 **시간을 줄이기 위해 군집화를 선택**하였다.

군집화에 사용한 method

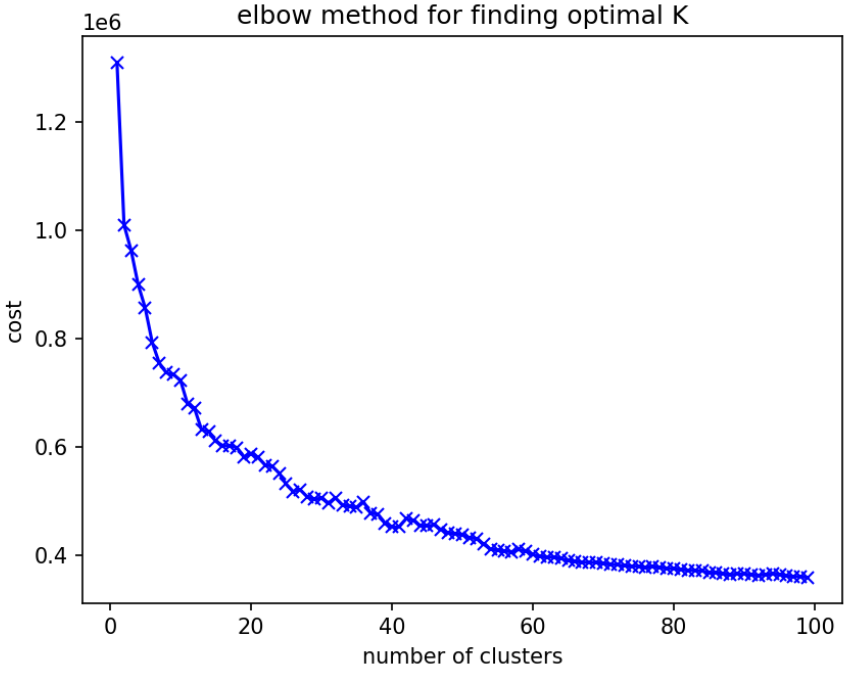
현재 변수들의 경우, 전부 범주형 column으로 되어있다. 이에, **범주형 변수들에 대한 군집화 기법**인 K-modes Clustering을 사용하였다. 여기서 K-modes Clustering이란, k개의 중심을 정하여 각 범주형 데이터 포인트와 그 데이터가 속한 군집의 중심 간의 해밍(Hamming) 거리를 최소화하는 알고리즘이다. 아래는 K-modes Clustering의 진행 단계이다.

- Step 1) k개의 중심점 설정: 범주형 데이터의 초기 중심점을 무작위로 선택한다.
- Step 2) 데이터 할당: 각 데이터 포인트를 가장 가까운 중심점과 같은 그룹에 할당합니다. 이때 해밍 거리를 사용하여 유사성을 측정한다.
- Step 3) 중심점 업데이트: 각 그룹의 중심점을 해당 그룹 내 가장 빈번한 값(모드)으로 업데이트
- Step 4) 반복: 중심점이 더 이상 업데이트되지 않을 때까지 2번과 3번 과정을 반복한다.

최적의 k값 찾기

최적의 k값을 찾기 위해 군집의 수를 1~100 까지 K-modes Clustering을 적용하였다. k-modes의 cost는 클러스터 내의 데이터 포인트와 해당 클러스터의 중심 간의 불일치 정도를 나타내는 척도이다. '[그림 3] Elbow method' 을 보면, x축의 경우 군집의 수(K)이고 Y축은 앞서 언급한 Cost를 나타낸다. 여기서 상당히 완만해지는 구간이 60에서 발생한다고 판단하여 **최종적으로 최적의 K값을 60**개로 설정하였다.

[그림 3] Elbow method



최적의 군집 수로 비지도 학습 분류

군집화의 마지막 단계로 실제 KModes 함수를 사용하여 분류를 시행하였고, 각 군집에 대해 출력해본 결과 각 행들의 차이가 크게 나타나지 않았다. 하여, 분류된 cluster들에 대해 **새로운 칼럼인 'clusters'를 생성**하였다.

(3) SVD Algorithm

협업 필터링 방법 중 차원 축소를 통해 대규모 item-user martix를 저차원으로 변환시키는 **특이값 분해(SVD) 알고리즘**의 잠재 요인 모델을 적용하여 데이터의 확장성, 희소성 문제를 완화한 추천 시스템을 생성하였다. 본 분석이 대규모 데이터셋을 사용한다는 점을 고려하여 이전 절의 클러스터링 결과를 반영한 알고리즘 즉, 전체 item-user martix를 사용하는 대신, 클러스터와 운동 처방 간의 관계로 대체하여 분석을 진행했다. 이를 통해 SVD의 계산 복잡도를 줄이고 효율성을 높였다.

운동 처방 column 텍스트 처리 & item-user martix

우선 원본 데이터셋에서 아이템인 Exercise_PRSCRPTN(운동처방) 하나의 열 안에서 준비/본/마무리운동으로 나열되어 있던 운동들을 텍스트 처리를 통해 3개의 열로 분할하였다.

각 클러스터에서 각 운동 처방이 몇 번 추천되었는지 그 빈도를 집계하여 '클러스터-운동 처방 행렬'을 만들고 이를 item-user martix의 역할로 사용한다. 약 27만 개의 **사용자 대신 군집을 사용**함으로써 데이터의 noise를 줄여 SVD의 신뢰성을 향상시켰다.

최적의 n_component 찾기

SVD 알고리즘을 적용하기 위한 최적의 n_component를 찾은 결과, 누적 설명 분산 95% 기준으로 3개의 component를 선택했다. 준비/본/마무리운동 각각의 item-user martix에 sklearn 라이브러리의 TruncatedSVD 함수를 사용하여 SVD를 계산함으로써 **클러스터 잠재 요인 행렬**과 **아이템 잠재 요인 행렬**을 생성한다.

[표 5] item-user martix(본운동)

	cluster	exercise	Rating
0	0	V자 사이클	10
1	0	W 스텝	2
2	0	가슴 스트레칭	39
…	…	…	…
37677	59	턱걸이	72
37678	59	트레드밀에서 걷기	56
37679	59	파워클린	2

similarity 계산하여 사용자와 가장 유사한 군집 출력

데이터의 희소성 문제를 고려하여 사용자 데이터셋인 신체 특성 정보를 **정규화** 후 진행한다.

[식 1] 코사인 유사도

$$Cosine\ Silmilarity = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}$$

수치 정보를 포함하고 있지 않은 SEX 변수의 처리를 위해 원-핫인코딩을 통해 데이터셋에 변경된 값으로 입력하였다. sklearn 라이브러리의 StandardScaler, OneHotEncoder 함수를 사용하여 처리 후의 데이터를 통해 각 클러스터의 중심(평균)을 구하고, 새로운 사용자 입력값에 동일한 범주화 및 전처리 과정을 수행한 후 **유사한 클러스터를 찾아 할당**한다.

코사인 유사도를 활용하여 유사한 운동 처방 찾기

할당된 클러스터의 결과와 클러스터/아이템 잠재 요인 행렬을 기반으로 새로운 사용자의 잠재요인벡터를 구하고 위와 마찬가지로 item 잠재 요인 행렬과의 코사인 유사도를 계산하여 **상위 5개의 유사한 운동 처방을 추천**해준다.

(4) 모델의 평가

평가 방법 설명

본 분석의 추천시스템의 구현성을 확인해보기 위해 분류 모델의 평가 척도 중 가장 직관적이고 널리 사용되는 정확도의 방법을 수정/적합하여 이를 평가한다. 정확도(Accuracy)란 전체 데이터에서 모델이 얼마나 정확하게 예측했는지를 나타내는 값으로 모델의 전체 예측 수 중 올바른 예측의 비율을 의미한다.

$$\text{정확도}(Accuracy) = \frac{\text{올바른 예측의 수}}{\text{전체 예측의 수}}$$

이를 본 분석에 적용하기 위해 test set의 실제 처방에 추천 운동 처방이 포함되는 경우를 올바른 예측으로 간주하고 정확도 척도를 수정한다.

Sampling 설명

실험을 위해 각 군집당 10개씩 총 600명의 사용자를 랜덤으로 추출하여 이들이 실제로 처방받은 운동 처방과 알고리즘을 통해 추천받은 운동을 평가 방법을 사용하여 추천의 정확성을 평가한다.

최종 Accuracy

이를 통해 추출한 accuracy값은 77.333%로 동일한 신체 정보 입력을 가짐에도 다른 운동을 처방받은 원본 데이터의 한계에도 충분한 성능을 가졌음을 시사하며 아래에서 서술할 추후 변수의 추가를 통해 보다 개인화된 추천시스템으로 확장 가능성이 있음 또한 알 수 있다.

(5) Experiment(실제 입력하여 추천받기)

아래 '[표 6] 지인 신체 정보'와 같이 주변 지인들로부터 신체정보를 제공받아 추천시스템에 입력하였다. '[표 7] 추천 운동 결과물'은 입력한 정보에 의해 추천시스템이 출력한 최종결과물이다.

[표 6] 지인 신체 정보

사용자	나이	성별	키	몸무게	BMI	체지방률	허리둘레	상대악력	이완기최저 혈압	수축기최고 혈압
A	13	여성	163cm	70	26.3	28	75	45	80	125
B	18	남성	173.9	53.5	17.7	14.8	67	75.7	78	127
C	63	남성	169.1	79.6	27.8	37.3	103.9	34.2	91	139
D	30	여성	167	60	21.5	24.0	70	58	73	121

[표 7] 추천 운동 결과물

	추천 준비운동	추천 본운동	추천 마무리운동
A	'윗몸올리기', '전진 점프하며 발 뒤꿈치 찍기', '앉았다 일어서기', '상지 루틴 스트레칭', '줄넘기'	'다리 뻗어 올리기', '연속 하향 점프하기', '줄넘기', '깍지 끼고 상체 숙이기', '누워서 전신 뻗기'	'하지 루틴 스트레칭2', '전신 루틴 스트레칭', '서서 발목 뒤로 당기기', '상지 루틴 스 트레칭', '넙다리 뒤쪽 스트레칭'
B	'전진 점프하며 발 뒤꿈치 찍기', '상지 루틴 스트레칭', '하지 루틴 스트레칭1', '줄넘기', '윗몸올리기'	'한 발 뒤로 들어올리기', '다리 뻗어 올리기', '달리기', '앉아서 가슴 모으기', '바로서서 상체 숙이기'	'걷기', '허리 스트레칭', '스텝박스', '전신 루틴 스트레칭', '상지 루틴 스트레칭'
C	'걷기', '작은 공을 이용한 동적 루틴 스트레칭', '전신 루틴 스트레칭', '목 스트레칭', '몸통 들어올리기'	'달리기', '한쪽 다리펴고 상체 숙이기', '한 발 뒤로 들어올리기', '앉아서 가슴 모으기', '등/어깨 뒤쪽 스트레칭'	'걷기', '허리 스트레칭', '밴드 걸고 앉아서 발등 굽힘', '스텝박스', '엉덩이 스트레칭'
D	'걷기', '하지 루틴 스트레칭1', '작은 공을 이용한 동적 루틴 스트레칭', '전신 루틴 스트레칭', '목 스트레칭'	'달리기', '한쪽 다리펴고 상체 숙이기', '한 발 뒤로 들어올리기', '앉아서 가슴 모으기', '등/어깨 뒤쪽 스트레칭'	'걷기', '허리 스트레칭', '밴드 걸고 앉아서 발등 굽힘', '스텝박스', '하지 루틴 스트레칭1'

4) 시사점 및 기대효과

[국민 모두를 위한 추천 시스템]

개인의 연령, 성별, 체질량 지수(BMI), 체지방률, 허리둘레, 근력 상태, 혈압 상태 등 **다양한 건강 지표를 고려**하여 각 개인에게 가장 적합한 운동을 추천하였다. 개인의 특성을 고려한 맞춤형 서비스이므로 단순하고 일률적인 처방보다 효과적이다. 이를 통해 운동의 효과를 극대화하고, 질환 예방 및 관리가 수월해지며, **국민의 전반적인 건강 수준을 향상**시킬 수 있다. 특히, 연령대와 성별, 체지방률 등 개인화 요소들이 모두 추천 결과에 영향을 주는 요인들로 작용하여 개개인이 각각 다른 신체를 가지고 있어도 추천이 가능하다. 즉, 유소년부터 노년층, 저체중부터 과체중까지 다양한 상황들을 다룰 수 있는 **국민 모두에게 적용가능한 범용성이 높은** 추천 시스템이다.

[부담의 감소]

운동을 통한 질환 예방 및 관리가 효율적으로 이루어짐에 따라, 장기적으로 의료 비용을 절감할 수 있다. 이는 **사회적 부담을 줄이고**, 국가 경제에도 긍정적인 영향을 미칠 것이다.

접근성이 용이한 공공데이터를 활용하여 사회적 문제에 도움이 되는 모델을 만들었다는 점에서 사회 문제에 **누구나 쉽게 접근할 수 있다는 점**을 보여 지속적인 사회적 변화를 이끌어 낼 수 있다. 또한 본 연구를 계기로 운동 처방과 관련한 다양한 데이터 수집의 포문을 열어 추후에 보다 개선된 분석 모델을 이끌어낼 수 있을 것으로 기대한다.

[간편해지는 운동 처방]

본래, 51개의 칼럼으로 구성되어 있던 데이터를 사용하였다. 이는 운동처방을 받기 위해서 수십가지의 테스트를 받아야 하는 번거로움이 있었다. 허나, 이제는 9개의 칼럼으로 기존과 비교하여 다소 **간소화된 정보만을 요구**하여 보다 효율적으로 추천 운동을 처방 받을 수 있다.

기존의 운동처방은 직접 현장으로 가서 수십가지의 체력 측정을 한 뒤에 운동 처방을 하였다. 하지만 이제는 본 연구의 추천시스템을 활용하면 **집에서도 간편하게 운동을 처방** 받을 수 있게 된다. 또한, 현재 제공된 운동처방 데이터의 운동처방결과 칼럼을 확인하면 비슷한 체력과 체형의 사람에게도 상이한 운동을 처방하는 경우가 이따금씩 발견된다. 이에, 앞선 분석에서 활용한 군집화 결과를 이용한다면 위와 같은 경우를 방지할 수 있는 어느 정도의 **척도의 역할을 수행**할 수 있다.

5) 문화데이터 활용성

o 문화 데이터 명

- 체력측정 및 운동처방 종합 데이터

o 데이터 소개

- 서울올림픽기념 국민체육진흥공단에서 관리하고 있는 국민체력측정데이터의 항목별 측정 정보와 운동처방결과를 종합적으로 제공하는 데이터이다. 체력측정 센터명, 연령대, 신장, 체중, 윗몸일으키기, BMI, 제자리 멀리뛰기 등의 체력측정 값 및 그에 따른 운동처방결과를 조회할 수 있다.

o 데이터출처

- 서울올림픽기념국민체육진흥공단

o URL

https://www.bigdata-culture.kr/bigdata/user/data_market/detail.do?id=b3924850-aa65-11ec-8ee4-95f65f846b27

o 활용 내용 및 획득방법

- 2020년 1월부터 2023년 4월까지의 월별 데이터를 전부 활용하였다. 해당 데이터들은 위 URL에서 회원가입 후 다운 받을 수 있고 총 40개의 csv파일을 병합하여 하나의 csv 프레임을 만들었고 총 573,027개의 행과 51개의 열로 이루어져 있다. 각 열에 대한 설명은 위 URL의 컬럼 정의서를 참고하면 된다.

o 타분야 공공·민간 데이터 융복합

- 체력측정 및 운동처방 종합 데이터와 시간적, 공간적, 경제적 여건 등의 다양한 정보를 종합적으로 고려한 개인 맞춤형 추천 시스템을 구현할 수 있습니다. 이를 통해 사용자의 개별 상황과 필요에 최적화된 운동 프로그램을 제공할 수 있습니다.
- 시간적, 공간적, 경제적 여건 등 다양한 정보를 종합적으로 고려하여, 체력측정 및 운동처방 종합 데이터를 활용한 보다 개인화된 추천 시스템을 제공할 수 있습니다.

6) Reference

- 이한나, 백수빈, 박두순 (2019) R에서 협업필터링과 개인화 요소를 이용한 개인 맞춤형 운동 추천 시스템, 2019년 추계학술발표대회 논문집, Vol. 26, No. 2
- 정우용, 경찬욱, 이승우, 김수현, 선영규, 김진영 (2022) 신경망 협업 필터링을 이용한 운동 추천시스템, 한국인터넷방송통신학회 논문지, 22:6, 173-17
- 이하영, 정옥란. (2021). 차원축소 알고리즘을 이용한 개인화된 운동 추천 시스템. 한국컴퓨터정보학회논문지 , 26(6), 19-28.