2024 문화 디지털혁신 및 문화데이터 활용 공모전

신체정보 및 건강상태 기반 맞춤형 운동 추천 시스템

SODA

Contents 목차

01 분석배경 및 목적

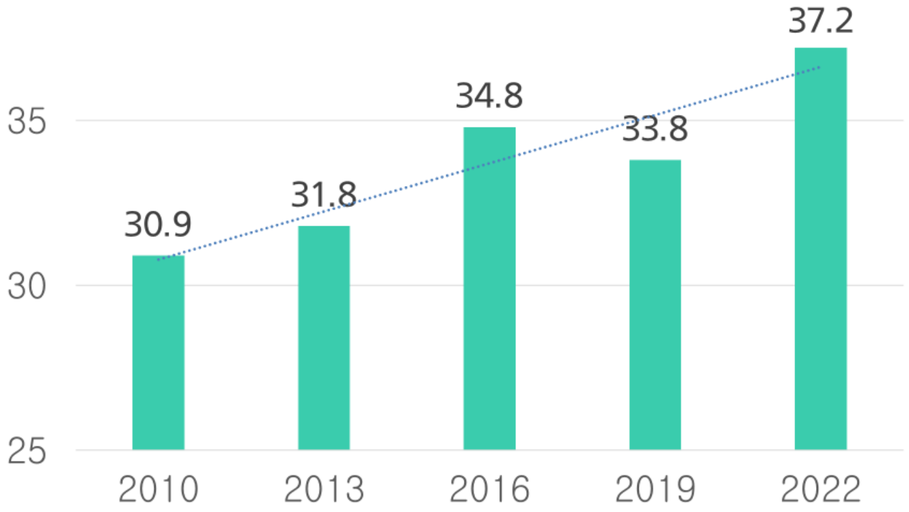
02 문화데이터 활용 정보

03 분석
- 데이터 전처리
- 군집화(K-modes)
- SVD 알고리즘
- 모델의 평가
- Experiment

04 시사점 및 기대효과

분석 배경 및 목적

[표 1] 국내 전체비만율 추이

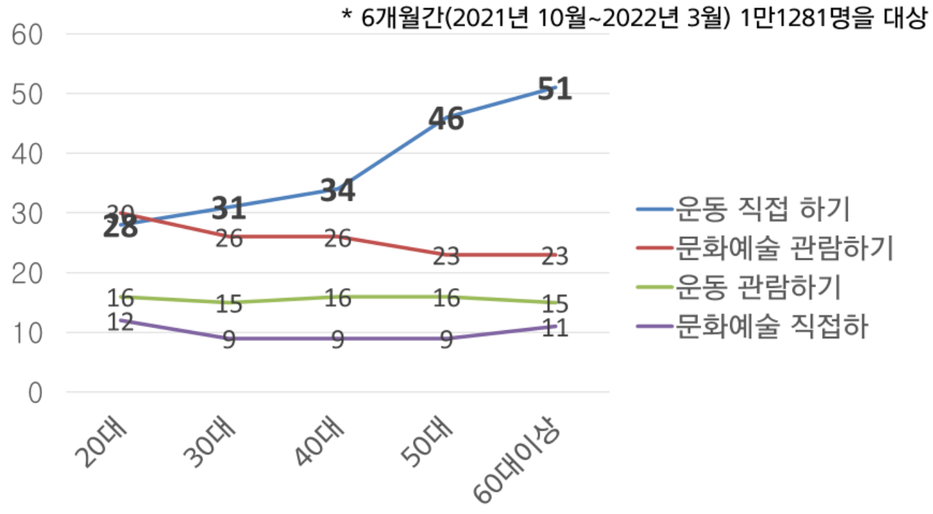


출처: 국가발전지표, 지표누리

1 증가하는 비만율

2010년 이후 약 7%p 정도의 비만율이 상승

[표 2] 문화 예술 활동 측면 여가활동 관심도 연령별 추이



출처: 컨슈머인사이트

2 높아지는 운동 관심도

개인별 운동 추천시스템을 제공 → 현재의 관심도를 연결시켜주는 일종의 <u>징검다리 역할</u>

분석 주제 문화데이터 활용정보(1)

1 문화 데이터 명

: 체력측정 및 운동처방 종합 데이터

2 데이터 소개

- 서울올림픽기념 국민체육진흥공단에서 관리하고 있는 국민체력측정데이터의 항목별 측정 정보와 운동처방결과를 종합적으로 제공하는 데이터
- 체력측정 센터명, 연령대, 신장, 체중, 윗몸일으키기, BMI, 제자리 멀리뛰기 등의 체력측정 값 및 그에 따른 운동처방결과 조회 가능

3 데이터 출처

: 서울올림픽기념 국민체육진흥공단

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

4 활용 내용 및 획득방법

- 2020년 1월부터 2023년 4월까지의 월별 데이터 융합
- 총 40개의 csv파일을 병합하여 하나의 csv 프레임을 만들었고 총 573,027개의 행과 51개의 열로 이루어짐

분석 주제 문화데이터 활용정보(2)

[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

분석 I. 데이터 전처리

1 Import Library

- 2020년 1월 csv 파일부터 2023년 4월까지의 모든 파일을 R에서 dplyr package를 활용하여 병합
- 이후 분석의 모든 과정에서 python 툴을 활용
- 라이브러리는 sklearn, pandas, numpy, seaborn, kmodes, matplotlib, collection를 사용

2 변수선택

- 대다수의 값이 결측치인 변수 제거
- 상대악력에 대해 좌수악력과 우수악력을 나타내는 변수는 상관계수가 높음으로 제거

3 결측치 처리

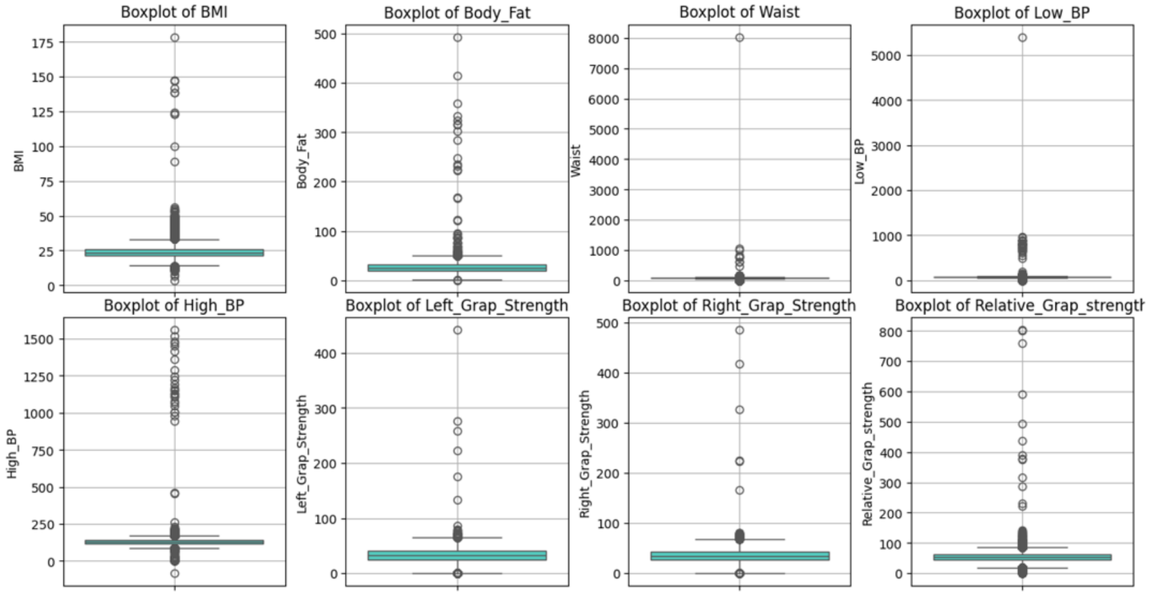
- 유소년의 경우, 체지방률이 무의미하여 측정하지 않아 1개의 행을 제외하고 모두 NA값으로 표시
- 체지방 칼럼의 경우, 범주형 데이터로 변환 예정이기에 유소년의 NA를 모두 0으로 보간
- 신체정보이기에 결측치를 보유하고 있는 행들에 대해서는 제거

[표 3] 악력 상관계수

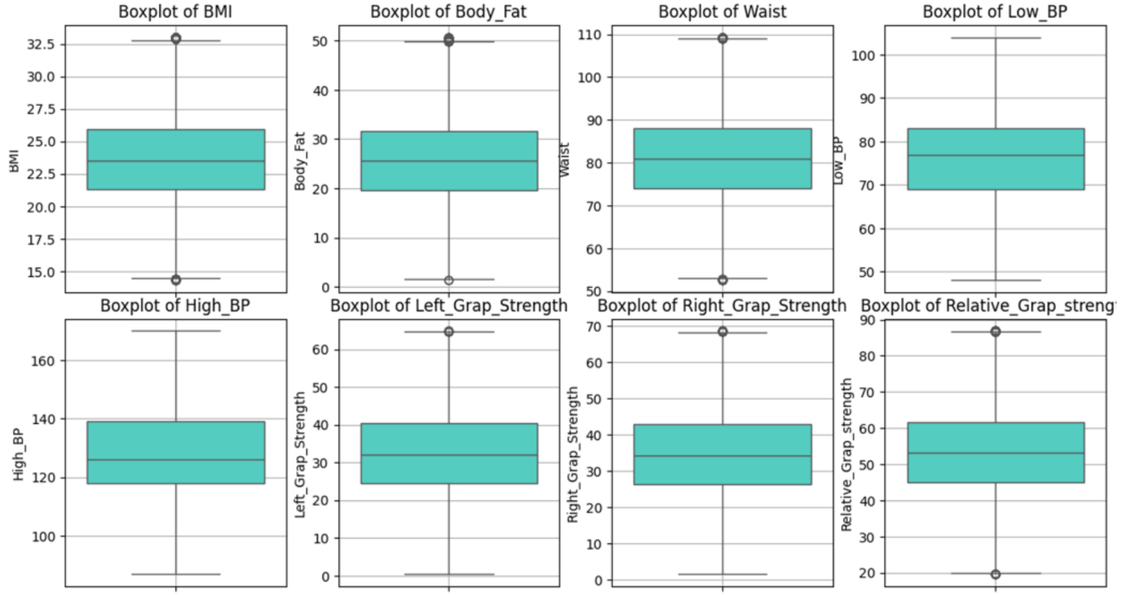
	좌수악력	우수악력	상대악력
좌수악력	1	0.94	0.75
우수악력	0.94	1	0.79
상대악력	0.75	0.78	1

분석 I. 데이터 전처리

[그림 1] 이상치 처리 전 Boxplot



[그림 2] 이상치 처리 후 Boxplot



4 이상치 처리

- [하한선(Q1-1.5 X IQR), 상한선(Q3+1.5 X IQR)] 구간 내의 값들만 포함하도록 이상치를 처리

5 범주화

- BMI, Blood_Pressure(혈압), Body_Fat(체지방률)과 같이 <u>사회에서 통념</u>적으로 범주화가 되어있는 열들은 해당 기준을 사용
- 나머지는 <u>등구간 범주화</u>를 진행하였다. 등구간 범주화를 진행함에 있어 describe 함수를 사용하여 평균, 표준편차, 최대, 최소, 1,3분위수를 참고하여 비율적으로도 일부 조정

전처리가 끝난 후 최종 데이터의 사이즈는 270245 X 9

분석 II. 군집화

1 군집화의 목적

- 각 행에 대해 유사도를 비교할 시 막대한 계산량으로 인해 추천을 받고자 하는 데에 시간이 많이 소모
- 클러스터와 운동 처방 간의 관계를 분석함으로써 **복잡성을 줄이며 효율성을 높이고 SVD의 신뢰성 향상**

2 군집화 사용 method

- 범주형 변수들의 군집화 기법인 K-modes Clustering을 사용
- K-modes Clustering이란, k개의 중심을 정하여 각 범주형 데이터 포인트와 그 데이터가 속한 군집의 중심 간의 해밍(Hamming) 거리를 최소화하는 알고리즘

3 K-modes Clustering의 진행 단계

Step 1) k개의 중심점 설정: 범주형 데이터의 초기 중심점을 무작위로 선택

Step 2) 데이터 할당: 각 데이터 포인트를 가장 가까운 중심점과 같은 그룹에 할당 (해밍 거리를 사용하여 유사성을 측정)

Step 3) 중심점 업데이트: 각 그룹의 중심점을 해당 그룹 내 최빈값(mode)으로 업데이트

Step 4) 반복: 중심점이 더 이상 업데이트되지 않을 때까지 2번과 3번 과정을 반복

분석 II. 군집화

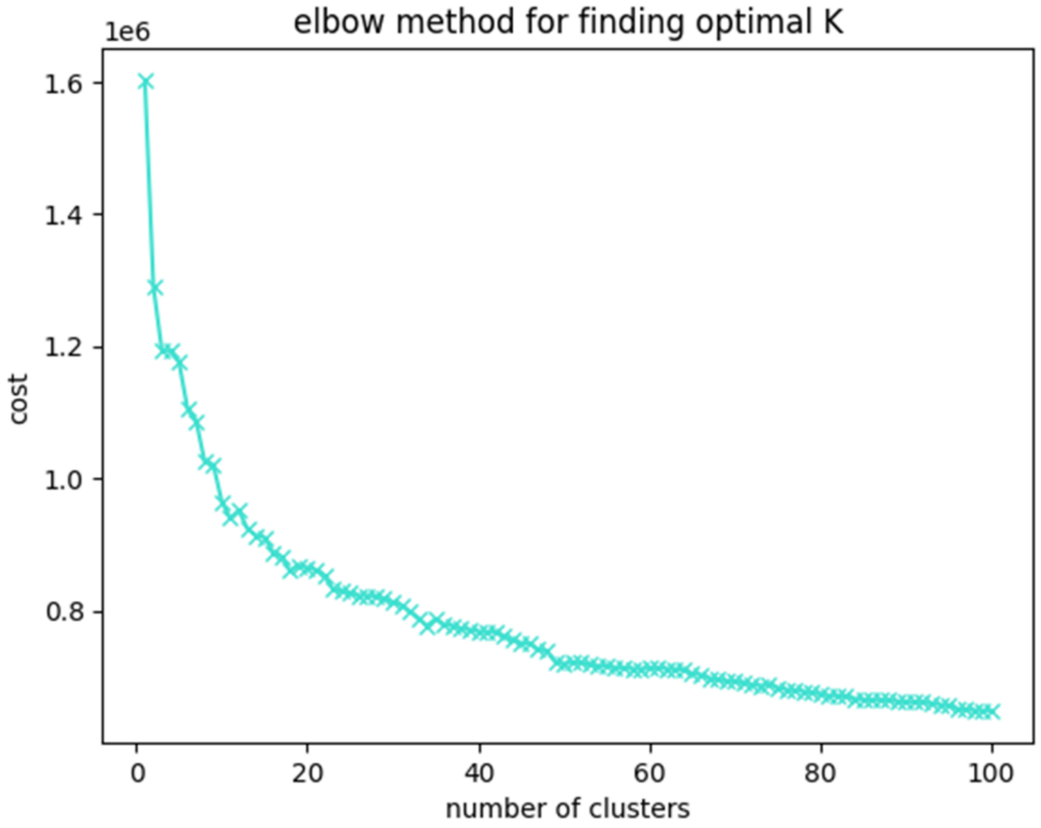
4 최적의 k값

- 최적의 k값을 찾기 위해 군집의 수를 1~100까지 K-modes Clustering을 적용
- K-modes의 cost는 클러스터 내의 데이터 포인트와 해당 클러스터의 중심 간의 불일치 정도를 나타내는 척도
- '[그림 3] Elbow method' 을 보면, 완만해지는 구간이 60에서 발생한다고 판단하여 <u>최종적으로 최적의 K값을 60개로 설정</u>

5 최적의 군집 수로 비지도 학습 분류

- K-modes 함수를 사용하여 분류를 시행하였고, 각 군집에 대해 출력해본 결과 각 행들의 차이가 크게 나타나지 않음
- 분류된 cluster들에 대해 <u>새로운 칼럼인 'clusters'를 생성</u>

[그림 3] Elbow method



분석 III. SVD Algorithm

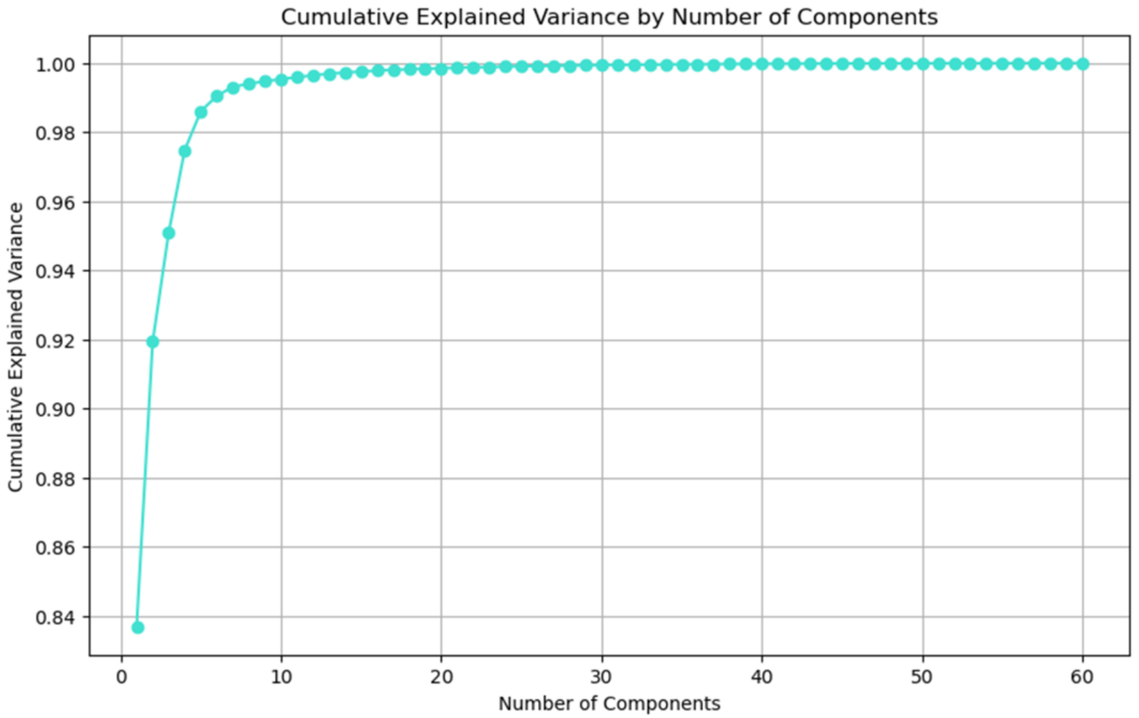
1 운동 처방 column 텍스트 처리 & item-user matrix

- Exercise_PRSCRPTN(운동처방) 하나의 열 안에서 준비/ 본/ 마무리 운동이 단순 나열되어 있던 운동들을 텍스트 처리를 통해 3개의 열로 분할
- 각 운동 처방이 군집별로 몇 번 추천되었는지 그 빈도를 집계하여 '클러스터-운동 처방 행렬' 생성

2 최적의 n_component 찾기

- 누적 설명 분산 95% 기준으로 3개의 component를 선택
- 기존의 Item-user matrix 추출 방식을 수정/보완
- 준비/본/마무리운동 각각의 item-user matrix에 SVD를 계산함으로써 클러스터 잠재 요인 행렬과 아이템 잠재 요인 행렬을 생성

[그림 4] Elbow Method for Number of Component



분석 III. SVD Algorithm

3 사용자와 가장 유사한 군집 출력

- 데이터의 희소성 문제를 고려하여 사용자 데이터셋인 신체 특성 정보를 정규화 후 진행
- 각 클러스터의 중심(평균)을 구하고, 새로운 사용자 입력값에 동일한 범주화 및 전처리 과정을 수행한 후 코사인 유사도를 계산하여 유사한 클러스터를 찾아 할당

4 유사한 운동 처방 찾기

- 할당된 클러스터의 결과와 클러스터/아이템 잠재 요인 행렬을 기반으로 새로운 사용자의 잠재요인벡터 생성
- item 잠재 요인 행렬과의 코사인 유사도를 계산하여 상위 5개의 유사한 운동 처방을 추천

[식 1] 코사인 유사도

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

분석 Ⅳ. 모델의 평가

1 평가 방법

- 분류 모델의 평가 척도 중 가장 직관적이고 널리 사용되는 정확도(Accuracy)의 방법을 사용
- 정확도(Accuracy)란, 전체 데이터에서 모델이 얼마나 정확하게 예측했는지를 나타내는 값으로 모델의 전체 예측 수 중 올바른 예측의 비율을 의미
- test set의 실제 처방에 추천 운동 처방이 포함되는 경우를 올바른 예측으로 간주하고 정확도 측정

2 Sampling & 최종 Accuracy

- 각 군집당 10명씩 총 600명의 사용자를 랜덤으로 추출하여 이들이 실제로 처방받은 운동과 알고리즘을 통해 추천받은 운동을 사용하여 정확성을 평가
- accuracy값은 74.83%로 동일한 신체 정보를 가진 사용자들이 다른 운동을 처방받은 원본 데이터의 한계에도 충분한 성능을 가졌음을 시사

[식 2] Accuracy

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

분석 Ⅴ Experiment(실제 입력하여 추천받기)

[표 6] 신규 사용자 신체 정보

사용자	나이	성별	키	몸무게	BMI	체지방률	허리둘레	상대악력	이완기최저혈압	수축기최고혈압
A	12	여성	145	40	19	-	65	39.0	70	110
B	35	남성	174	95	31.4	54.9	122	43.9	90	131
C	65	여성	155	50	20.8	30.1	80.4	36.0	75	120
D	25	남성	181	75	22.9	14.7	74	86.4	86	127

[표 7] 추천 운동 처방

	추천 준비운동	추천 본운동	추천 마무리운동
A	'팔굽혀펴기', '몸통 비틀기', '실외 자전거타기', '마주보고 종아리 펴기', '허리 굽혀 팔 뒤로 들기'	'동적 스트레칭 루틴프로그램', '팔굽혀 펴기', '레그 스윙', '조깅', '정적 스트레칭 루틴프로그램'	'가슴/어깨 앞쪽 스트레칭', '서서 어깨 들어올리기', '네발기기 자세로 팔 다리 들기', '조깅', '누워서 하늘 자전거'
B	'허리 굽혀 덤벨 들기', '하지 루틴 스트레칭2', '앉아서 몸통 움츠리기', '걷기', '전신 루틴 스트레칭'	'누워서 엉덩이 들어올리기', '뒤로 팔굽혀펴기', '스타 투 니 풀', '요통을 위한 스트레칭1', '앉아서 양팔 당기기'	'의자에 앉아 가슴 뒤로 젖히기', '하지 루틴 스트레칭2', '넙다리 뒤쪽 스트레칭', '상지 루틴 스트레칭', '허리 스트레칭'
C	'어깨 누르기', '아기자세', '짐볼 윗몸일으키기', '목 스트레칭', '바벨 당겨 올리기'	'누워서 팔 밀기', '한 발 뒤로 들어올리기', '앉아서 다리 굽히기', '상체 앞으로 숙이기', '팔꿈치 굽히기'	'서서 발목 뒤로 당기기', '허리 스트레칭', '걷기', '엉덩이 스트레칭', '등/어깨 뒤쪽 스트레칭'
D	'상지 루틴 스트레칭', '하지 루틴 스트레칭1', '버피 테스트', '앉아서 모으기', '맨몸운동 루틴프로그램'	'스타 투 니 풀', '누워서 엉덩이 들어올리기', '줄넘기', '나무자세', '달리기'	'하지 루틴 스트레칭2', '의자에 앉아 가슴 뒤로 젖히기', '넙다리 뒤쪽 스트레칭', '서서 어깨 들어올리기', '상지 루틴 스트레칭'

시사점 및 기대효과

1 국민 모두를 위한 추천 시스템

- 개인의 연령, 성별, 체질량 지수(BMI), 체지방률, 허리둘레, 근력 상태, 혈압 상태 등 다양한 건강 지표를 고려하여 <u>보다 개인화된 운동 추천시스템을 구현</u>
- 국민 모두에게 적용가능한 범용성이 높은 추천 시스템

2 간편해지는 운동 처방

- 9개의 변수로 기존과 비교하여 핵심적인 정보만을 요구

→ 기존의 '10미터 4회 왕복달리기' 등 측정이 까다로운 정보 불필요

- 비슷한 체력과 체형의 사람에게도 상이한 운동을 처방하는 경우 발생

→ 군집화 결과를 이용하여 척도의 역할을 수행

3 융복합, 활용 가능성

- 운동에 필요한 최소 자본 등 다양한 시간적, 공간적, 경제적 조건이 결부되어 추가적인 문화데이터와의 융복합을 통해 보다 정제되고 개인화된 운동 추천 시스템을 이끌어낼 수 있을 것으로 기대
- 향후 분석에서는 처방된 운동에 대한 선호도 조사를 통해 보다 정교한 추천시스템을 만들 수 있을 것으로 기대

File Edit Selection ··· code

recommand&accuracy_(5).py

recommand&accuracy_(5).py > ...

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.decomposition import TruncatedSVD
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.metrics.pairwise import cosine_similarity

df = pd.read_csv('seperate_exercise.csv', encoding='euc-kr', index_col=0)
prep_exercises_matrix = pd.read_csv('prep_exercises_matrix.csv', encoding='cp949', index_col=0
main_exercises_matrix = pd.read_csv('main_exercises_matrix.csv', encoding='cp949', index_col=0
cool_exercises_matrix = pd.read_csv('cool_exercises_matrix.csv', encoding='cp949', index_col=0

# 최적의 잠재요인수(n_components) 찾기
def find_n_components(user_item_matrix):
    svd = TruncatedSVD(n_components=60, random_state=42)
    svd.fit(user_item_matrix)

    explained_variance = svd.explained_variance_ratio_
    cumulative_explained_variance = np.cumsum(explained_variance)
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

Python
Python

```
qorud@LAPTOP-RVSAR8PS MINGW64 C:/Users/qorud/AppData/Local/Programs/Microsoft VS Code
$
 *  History restored


qorud@LAPTOP-RVSAR8PS MINGW64 C:/Users/qorud/AppData/Local/Programs/Microsoft VS Code
$
```

0 0 0 Ln 1, Col 19 Spaces: 4 UTF-8 CRLF Python 3.11.5 ('base': conda)

경청해 주셔서
감사합니다.