# 사용자 맞춤형 관광지 추천 시스템

# Customized Tourist Destination Recommendation System

정원홍
2020320089
Korea University
thetwo0525@korea.ac.kr

박지원
2021160043
Korea University
studypjw0104@naver.com

우혜민
2021340035
Korea University
woohm3@naver.com

최윤재
2022320303
Korea University
jyunchoi0710@naver.com

## ABSTRACT

본 연구는 여행 로그 데이터를 기반으로 한 개인 맞춤형 관광지 추천 시스템을 제안한다. 제안된 시스템은 다단계 분석 파이프라인을 중심으로 구되며, 각 단계는 연속적이고 통합된 방식으로 연결된다.

우선, 관련성이 낮은 장소 유형을 필터링하고, 분산된 다중 테이블을 통합하여 데이터 전처리를 수행한다. 사용자별 평균 만족도, 재방문 의도, 추천 의도 점수에 대해 사분위수 범위(IQR)를 기반으로 이상치를 탐지하고, 해당 범위를 벗어나는 극단값을 포함하는 기록은 분석의 정확도를 저해할 수 있으므로 제거하여 보다 정제된 데이터셋을 확보한다.

이후, 성별, 연령 등 인구통계학적 특성과 더불어 다양한 여행 스타일을 기준으로 비지도 학습 기반의 사용자 클러스터링을 수행한다. 사용자 클러스터링 이후에는 각 그룹의 대표적인 여행 성향을 분석하고, 이를 기반으로 해석 가능한 사용자 프로파일을 도출한다. 특히 성별 및 연령대에 따른 여행 스타일의 차이를 클러스터링 결과를 통해 확인하며, 이는 추천의 정밀도를 향상시키는 기초 자료로 활용된다.

다음 단계에서는 각 클러스터 내에서 빈발 패턴 마이닝을 적용하여 자주 함께 방문되는 장소 트랜잭션을 도출하고, 이를 바탕으로 추천 후보지를 선정한다.

마지막으로, 분류 모델을 활용하여 각 후보지에 대한 사용자의 만족도를 예측하고, 예측된 만족도에 따라 추천 대상을 정렬 및 정제함으로써 최종 추천 결과를 도출한다.

본 시스템은 각 분석 단계를 통합적으로 구성함으로써, 사용자 맞춤형이며 맥락을 반영한 관광지 추천의 가능성을 탐색하고자 하였다. 시스템 설계 과정을 통해, 사용자 특성과 행동 양식을 반영한 맞춤형 추천의 가능성과 그 한계를 함께 조명하였다. 서비스화 실질적인 성능 향상보다는, 추천 시스템의 구조와 데이터 활용 방식에 대한 탐색적 기여에 의의를 둔다. 향후 추가적인 데이터 보완과 방법론 개선을 통해 실질적인 추천 정확도를 높일 수 있을 것으로 기대된다.

## 1 INTRODUCTION

### 1.1 Motivation and Problem Setting

오늘날 관광 산업은 개인화되고 데이터 기반의 경험을 중심으로 빠르게 진화하고 있다. 그러나 기존의 대부분의 추천 시스템은 여전히 인기순, 리뷰 점수와 같은 일반적인 지표에 의존하고 있으며(Figure 1), 이러한 시스템은 사용자 개인의 선호나 여행 맥락을 고려하지 않고 모든 이용자에게 동일한 목적지 목록을 제시한다. 이와 같은 획일화된 접근 방식은 개별 여행자의 다양한 요구를 충족시키지 못하는 한계를 지닌다.

이는 지역 관광 분야에서도 문제로 인식될 수 있다. 지역 관광에서는 방문객의 행동을 정확하게 이해하는 것이 효과적인 추천과 참여를 이끌어내는 핵심이기 때문이다. 여행 목적지를 보다 적절하게 추천하기 위해서는 사용자가 누구인지, 어디를 다녀왔는지, 무엇을 즐겼는지에 대한 정보가 반드시 고려되어야 하지만, 이러한 정보는 대부분의 기존 시스템에서는 반영되지 않는다.

출처: 여기어때

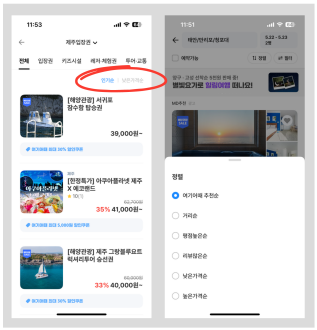


**Figure 1: '여기어때' 앱 UI**

### 1.2 Goal

이러한 문제의식에 기반하여, 본 프로젝트는 실제 여행 로그 데이터를 바탕으로 개인 맞춤형 관광지 추천 시스템을 설계하는 것을 목표로 한다. 본 연구에서 활용한 데이터셋은 성별과 연령을 포함한 인구통계 정보, 자연/도시, 새로운 장소/익숙한 장소, 인기 지역/한적한 지역, 선호 지역, 여행 목적(휴식 또는 체험), 여행 성향(계획형 또는 즉흥형), 사진 촬영의 중요도와 같은 여행 스타일 설문 응답, 방문 장소, 만족도 및 재방문 의사 점수 등을 포함한 12,000명 이상의 국내 여행자에 대한 정보를 담고 있다.

### 1.3 Significance

본 연구 설계 목표의 의의는 개별 여행자에게 보다 의미 있는 추천을 제공함과 동시에 지역 관광 활성화에 기여할 수 있는 가능성에 있다. 본 시스템은 행동 기반 데이터를 활용하고, 클

러스터링, 패턴 마이닝, 만족도 예측 기법을 통합함으로써, 사용자 개별성과 추천 정확도 간의 간극을 해소하고자 한다.

## 2 DATASET

본 프로젝트에서는 AI Hub Korea에서 제공한 실제 국내 여행 로그 데이터를 활용하였다. 원본 데이터는 여행객 정보, 여행 기록, 방문지 정보, 동행자 정보로 구성된 총 네 개의 테이블로 분산되어 있었으며, 이를 분석 목적에 맞게 하나의 통합 테이블로 구성하였다.

이 과정에서 다음과 같은 주요 컬럼을 선별하였다.

- **여행객 정보**: 여행자 고유 ID, 성별 (GENDER), 연령대 (AGE_GRP), 여행 스타일 관련 항목:
  - 자연/도시: TRAVEL_STYL_1
  - 익숙한 지역/낯선 지역: TRAVEL_STYL_3
  - 인기 관광지/한적한 관광지: TRAVEL_STYL_5
  - 휴양 여행/체험활동 위주 여행: TRAVEL_STYL_6
  - 계획적 여행/즉흥적 여행: TRAVEL_STYL_7
  - 사진 촬영 중요도: TRAVEL_STYL_8
- **여행 기록**: 여행 고유 ID, 여행 시작일 (TRAVEL_START_YMD), 여행 종료일 (TRAVEL_END_YMD), 이동 수단 (MVMN_NM), 총 여행일 수 (TRAVEL_DAYS), 여행 시기 (MAIN_TRAVEL_MONTH)
- **방문지 정보**: 방문지명 (VISIT_AREA_NM), 방문지 유형 (VISIT_AREA_TYPE_CD), 방문지 만족도 (DGSTFN), 재방문 의사 (REVISIT_INTENTION), 추천 의사 (RCMDTN_INTENTION)

데이터 품질 확보를 위해 다단계 전처리를 수행하였다. 먼저, 방문지 유형 코드가 관광과 직접적으로 관련된 값인 1, 2, 3, 4, 5, 6, 7, 8, 12, 13에 해당하는 레코드만을 유지하고, 주거지, 교통시설, 식당 등 일상적 장소는 제외하였다. 또한 여행 기간이 비정상적으로 짧거나 긴 경우, 만족도나 재방문/추천 의향과 같은 핵심 변수들이 누락된 경우 역시 제거하였다.

이와 함께, 사용자별 평균 만족도 (DGSTFN), 재방문 의사 (REVISIT_INTENTION), 추천 의사 (RCMDTN_INTENTION) 점수를 기준으로 사분위수 범위(IQR)를 활용한 이상치 탐지를 적용하였다. 각 변수별 IQR을 계산하여 $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$ 범위를 벗어나는 사용자에 대해서는 극단값으로 판단하고 전체 레코드를 제거함으로써, 분석의 신뢰도를 높였다.

최종적으로 정제된 데이터셋은 분석의 효율성을 높이기 위해 권역 단위(수도권, 서부권, 동부권, 제주권)로 분할 저장하였으며, 이후 단계에서 클러스터링, 패턴 마이닝, 분류 모델 입력값으로 사용되었다.

## 3 METHODOLOGY

본 추천 시스템은 데이터 전처리, 사용자 클러스터링, 패턴 마이닝, 만족도 예측을 통합한 다단계 분석 파이프라인으로 구성된다.

### 3.1 User Clustering

사용자의 여행 성향에 따라 유사한 그룹으로 분류하기 위해, 전처리된 데이터셋에 대해 클러스터링을 수행하였다. 먼저 성별(남성/여성)과 연령대(20대, 30대, 40대, 50대, 60대)를 기준으로 총 10개의 하위 그룹으로 나눈 뒤, 각 하위 그룹 내에서 독립적으로 클러스터링을 적용하였다.

클러스터링은 사용자의 여행 성향을 기반으로 수행되었으며, 해당 성향은 6개의 설문 항목에 대한 응답을 0 1 사이의 연속값으로 정량화한 6차원 벡터로 구성된다. 문항은 다음과 같다:

- 자연을 좋아하시나요? (값이 낮을수록 도시 선호)
- 새로운 지역을 좋아하시나요? (값이 낮을수록 익숙한 지역 선호)
- 휴양이나 휴식을 좋아하시나요? (값이 낮을수록 체험 활동 선호)
- 나만의 방문지를 좋아하시나요? (값이 낮을수록 유명 관광지 선호)
- 계획적인 여행을 좋아하시나요? (값이 낮을수록 즉흥적 여행 선호)
- 사진 촬영을 원하지 않나요? (값이 낮을수록 사진 촬영을 중요하게 생각)

클러스터링 기법은 **K-means**, **hierarchical clustering**, **DBSCAN**을 실험적으로 비교한 후 결정되었다. DBSCAN은 실루엣 계수가 대부분 음수에서 0.1 수준으로 낮게 나타났으며, hierarchical clustering은 실루엣 계수는 가장 우수했지만 일부 클러스터에 데이터가 과도하게 집중되는 불균형 현상이 발생하여 실용성이 떨어졌다. 이에 따라 실루엣 계수와 해석의 편의성을 고려하여 **K-means 알고리즘**을 최종적으로 채택하였다.

군집 간 유사도 측정에는 **유클리드 거리**를 사용하였으며, 클러스터 수 $k$는 2에서 20까지 변경하면서 실루엣 계수를 관찰하였다. 클러스터 수 변화에 따른 실루엣 계수의 차이가 크지 않았기 때문에, 해석의 일관성과 적용의 간결성을 고려하여 각 하위 그룹 내 클러스터 수를 **3개로 고정**하였다.

최종적으로 전체 사용자 기반에서 총 30개의 클러스터가 생성되었으며, 각 사용자에게 부여된 클러스터 라벨은 이후 패턴 마이닝 및 만족도 예측 단계에서 개인 맞춤형 추천의 기반으로 활용되었다.

### 3.2 Frequent Pattern Mining

사용자 그룹별로 맞춤형 목적지 추천 규칙을 도출하기 위해, 각 클러스터 내에서 **FP-Growth 알고리즘**을 적용하였다. Apriori 알고리즘에 비해 연산 효율성이 뛰어나며, 다양한 지지도 및 신뢰도 조건을 빠르게 적용할 수 있다는 점에서 FP-Growth를 선택하였다.

패턴 마이닝에 앞서, 전체 데이터에서 30회 미만으로 등장한 방문지는 희소 지역 또는 이상값으로 간주하여 제외하였다. 이후 각 거래(transaction)는 특정 클러스터와 방문지 이름으로 구성된 2-항목(item) 집합으로 구성하였다. 이는 특정 클러스터 그룹과 관광지 간의 동시 발생(co-occurrence)을 나타낸다.

마이닝 조건으로는 **최소 지지도 0.8%**, **최소 신뢰도 0.35**를 설정하였다. 이는 통계적으로 유의미한 규칙만을 추출하면서도, 지나치게 약하거나 드문 연관 규칙은 걸러내기 위함이다. 도출된 규칙의 형태는 다음과 같다:

> "Cluster X에 속한 사용자는 Y 관광지를 Z%의 확률로 함께 방문하였다."

예시 규칙은 다음과 같다:

- Cluster_0에 속한 사용자가 ***에버랜드***를 방문한 경우, ***국립중앙박물관***도 함께 방문했을 확률은 38.6%였다.
- Cluster_2에 속한 사용자가 ***경복궁***을 방문한 경우, ***캐리비안 베이***도 함께 방문했을 확률은 46.9%였다.

당초 기대와는 달리, 신뢰도(confidence) 값이 60% 이상에 달하는 강한 패턴은 거의 발견되지 않았으며, 대부분의 신뢰도 값은 0.5 이하에 머물렀다. 이는 특정 경향은 존재하더라도 클

러스터 간 여행 패턴이 명확히 분화되어 있지는 않다는 점을 의미한다.

도출된 연관 규칙을 보다 직관적으로 해석하기 위해, **NetworkX**를 활용한 그래프 시각화를 수행하였다. 텍스트 테이블에 비해 클러스터-관광지 간의 관계를 대규모로 시각화하는 데 효과적이며, 연관 구조를 한눈에 파악할 수 있는 장점이 있었다.

향후에는 클러스터 분할 방식을 조정하거나 지지도/신뢰도 조건을 완화함으로써 더욱 실행력 있는 연관 규칙을 발견할 수 있을 것으로 기대된다. 또한, 여행 목적이나 체류 기간과 같은 추가 속성을 반영하여 다차원적 규칙 탐색을 시도한다면 예측 성능이 향상될 가능성이 있다.

## 3.3 Satisfaction Prediction

사용자가 추천된 관광지를 얼마나 만족스러워할지를 예측하기 위해, 본 단계에서는 이진 분류(binary classification) 문제로 정식화하였다. 구체적으로, 만족도 점수(DGSTFN)가 4.0 이상인 경우를 긍정 클래스(1), 그 미만인 경우를 부정 클래스(0)로 정의하여 새로운 이진 라벨을 생성하였다.

입력 특성으로는 방문지 이름(VISIT_AREA_NM), 여행 시작일로부터 추출한 방문 월(Month), 재방문 의사(REVISIT_INTENTION), 추천 의사(RCMDTN_INTENTION) 등을 활용하였다. 범주형 변수인 방문지 이름은 원-핫 인코딩을 통해 벡터화하였으며, 이를 특징 변환과 로지스틱 회귀 모델을 결합한 파이프라인으로 구성하였다.

모델 학습은 계층적 비율(stratified split)을 유지한 70:30의 학습-테스트 데이터 분할 방식으로 진행하였으며, 클래스 불균형 문제를 완화하기 위해 클래스 가중치를 balanced로 설정하였다.

- **모델:** logistic regression (max_iter = 1000, class_weight = 'balanced')
- **입력 특성:** 방문지 이름, 방문 월, 재방문 의사, 추천 의사
- **목표 변수:** 이진 만족도 라벨 (만족도 DGSTFN ≥ 4.0이면 1)

모델 성능 평가는 정확도(Accuracy), 혼동 행렬(Confusion Matrix), 정밀도(Precision), 재현율(Recall), F1-score를 포함한 전체 분류 리포트를 기반으로 수행하였다. 결과적으로 본 모델은 만족/불만족 가능성을 구분하는 데에 일정 수준 이상의 예측력을 보였으며, 이는 최종 추천 결과를 필터링하거나 랭킹화하는 데 유용하게 활용될 수 있다.

# 4 EXPERIMENTS AND EVALUATION

## 4.1 Clustering

### 4.1.1 클러스터별 데이터 분포

각 성별-연령대 그룹 내에서 3개의 클러스터를 구성한 결과, 전체적으로 다음과 같은 데이터 개수 분포를 보였다.

**Table 1: 클러스터별 데이터 개수 분포**

| Age Group | Gender | Cluster 0 | Cluster 1 | Cluster 2 | Total |
|---|---|---|---|---|---|
| 20s | Male | 390 | 381 | 445 | 1216 |
| 20s | Female | 972 | 625 | 588 | 2185 |
| 30s | Male | 370 | 604 | 381 | 1355 |
| 30s | Female | 601 | 492 | 751 | 1844 |
| 40s | Male | 284 | 264 | 318 | 866 |
| 40s | Female | 343 | 339 | 390 | 1072 |
| 50s | Male | 174 | 158 | 140 | 472 |
| 50s | Female | 164 | 158 | 201 | 523 |
| 60s | Male | 64 | 84 | 47 | 195 |
| 60s | Female | 66 | 102 | 60 | 228 |

### 4.1.2 차원 축소 시각화 (PCA)

각 그룹별로 클러스터링 결과를 **PCA (Principal Component Analysis)**를 이용해 2차원 평면에 시각화하여, 군집 간 분포와 분리를 직관적으로 확인하였다.

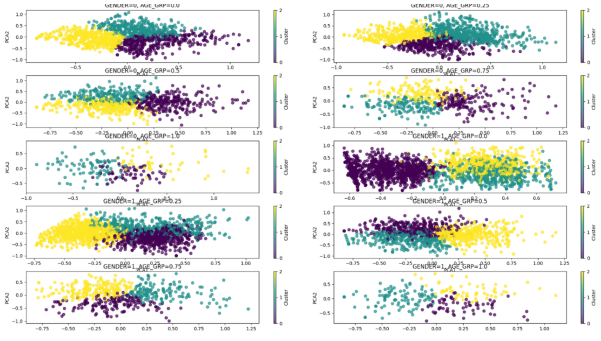


**Figure 2: 성별 및 연령대별 PCA 시각화 결과**

### 4.1.3 클러스터링 평가 지표 (Silhouette Coefficient)

클러스터 품질 평가는 실루엣 계수(Silhouette Coefficient)를 사용하였다. 이 지표는 응집도와 분리도를 동시에 고려하는 대표적인 평가 기준으로, 값이 1에 가까울수록 클러스터 간 분리가 잘 되었음을 의미한다.

**Table 2: Silhouette Coefficient by Group (Full Grid Format)**

| Age Group | Gender | Silhouette Coefficient |
|---|---|---|
| 20s | Male | 0.1580 |
| 20s | Female | 0.1609 |
| 30s | Male | 0.1492 |
| 30s | Female | 0.1540 |
| 40s | Male | 0.1459 |
| 40s | Female | 0.1476 |
| 50s | Male | 0.1618 |
| 50s | Female | 0.1562 |
| 60s | Male | 0.1568 |
| 60s | Female | 0.1813 |
| **Average** | | **0.1572** |

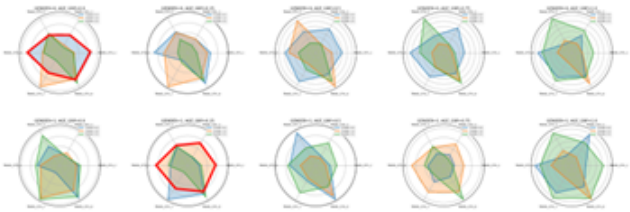
전반적으로 0.15 수준의 낮은 실루엣 계수를 보였으며, 이는 클러스터 간 분리가 명확하지 않다는 것을 의미한다. 이러한

결과는 사용자 성향이 이산적인 집단보다는 연속적인 스펙트럼에 가까움을 시사한다.

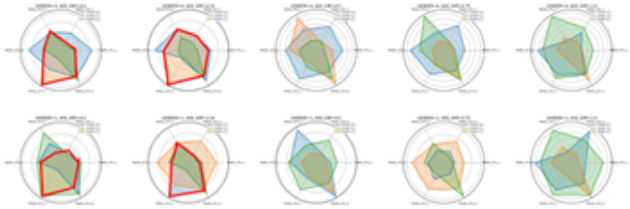
#### 4.1.4 Centroid 시각화 및 유형 분류

각 클러스터의 중심 좌표(centroid)를 **레이더 차트**로 시각화하여, 특징적인 성향을 가진 대표 유형을 분류할 수 있었다. 주요 centroid 유형은 다음과 같다:

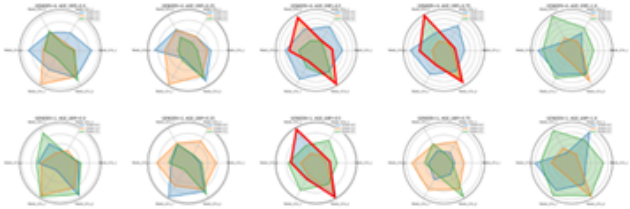
- **유형 1:** 전반적으로 값이 0.5에 가까운 중립형.



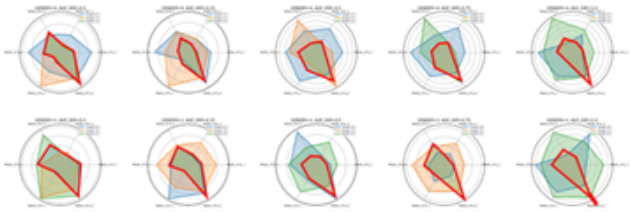
- **유형 2:** 도시, 익숙한 지역 선호, 계획적 성향 (주로 20–30대).



- **유형 3:** 도시, 익숙한 지역, 휴양 선호, 즉흥적 여행, 사진 촬영 선호 (40–50대).



- **유형 4:** 도시, 익숙한 지역, 체험 활동, 유명 관광지, 즉흥적 여행, 사진 촬영 선호 (전 성별/연령대 관찰됨).



#### 4.1.5 한계점 및 개선 방안

실루엣 계수는 전반적으로 약 0.15 0.18 수준으로, 클러스터 간 구분이 매우 뚜렷하지는 않음을 시사한다. 이는 클러스터 간 경계가 애매하거나, 사용된 변수들이 클러스터를 명확히 구분 짓기에는 다소 제한적이었을 가능성을 시사한다. 또한, 일부 클러스터의 centroid들이 유사한 양상을 보여 이들을 재통합하고 이를 기반으로 classification을 시도해보았지만 기존 모델 대비 성능 향상은 확인되지 않았다. 향후에는 더 다양한 feature를 활용하거나, 텍스트 응답, 위치 데이터 등 비정형 데이터를 포함하거나, feature engineering을 통해 더 고차원적 특성을 추출함으로써 클러스터 품질을 향상시킬 수 있을 것으로 기대된다.

### 4.2 Pattern Mining

#### 4.2.1 데이터셋 및 전처리

클러스터링을 통해 얻은 CLUSTER 분류 값을 TRAVELER_ID와 JOIN해준 데이터를 사용했다.
(/Pattern_Mining/Final/travel_data_final.csv)

사용 항목: GENDER, AGE_GRP, TRAVEL_STATUS_ACCOMPANY, VISIT_AREA_NM, CLUSTER_NEW

두 번째로, 클러스터링 결과에서 유사성을 띄는 항목으로 추가로 수행한 패턴 마이닝의 경우 같은 방식으로 데이터를 처리해 사용했다. 해당 파일들은 /Pattern_Mining/Final/ver2/에 존재한다.

#### 4.2.2 패턴 마이닝 과정 및 결과

고객의 성별, 연령대, 여행 스타일을 기준으로 분류한 클러스터링 결과를 바탕으로 '어떤 유형의 클러스터에 속한 사람은 어떤 장소를 여행할 가능성이 높은가?'를 계량적으로 분석해 1차적인 여행지 추천 목록을 선별하고자 했다.

Apriori 대비 속도의 이점이 있는 **FP-Growth** 방식을 사용했고, 신뢰도 임계치 필터링을 통해 어느 정도 패턴으로 볼 수 있는 트랜잭션만 남겼다. 또한 **outlier**로 인해 신뢰도가 너무 내려가는 것을 방지하기 위해 특정 횟수 이상 중복해서 등장하는 여행 지역만 패턴 마이닝을 실시했다.

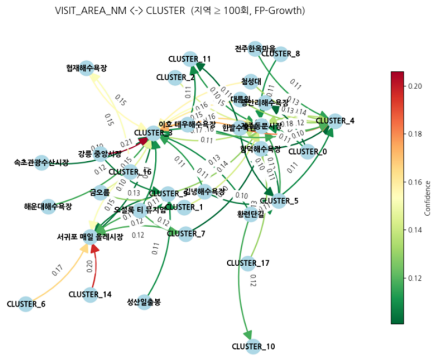



**Figure 3: 패턴 마이닝 수정 전**

위는 100회 이상 등장하는 지역을 바탕으로 패턴 마이닝을 처음 수행한 결과인데, 30가지의 클러스터 각각의 비율이 서로 일정하지 않아서 동일한 임계치를 기준으로 패턴 마이닝을 수행했더니 위와 같이 몇몇의 클러스터에만 패턴이 매칭되는 문제가 발생했다.

해당 문제를 해결하고자 반복문을 삽입해 매칭된 패턴이 존재하지 않는 클러스터의 경우 수정된 임계치 값으로 다시 패턴 마이닝을 실시해서 각 클러스터에 최소 5개의 여행지 패턴이 보장되도록 수정했다.

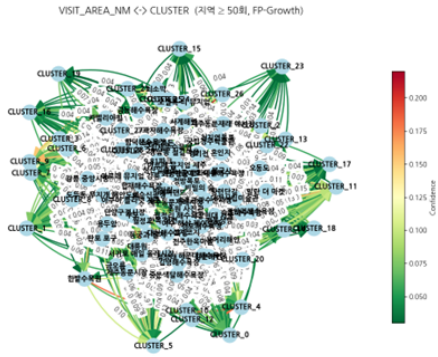



**Figure 4: 패턴 마이닝 수정 후**

이로써 모든 클러스터에 최소 5개의 여행지 패턴이 매칭되는 1차적인 여행지 추천 목록을 수집할 수 있었다.

패턴 마이닝의 결과로 적절한 양의 추천 목록을 얻을 수 있었지만, 각 클러스터에 매칭되는 패턴의 수가 클러스터에 따라 많게는 20가지 이상 차이나는 문제점이 존재했다. 하지만 해당 문제점은 데이터 자체가 일정한 비율로 고객-여행지 트랜잭션을 가지고 있지 않기 때문에 각 클러스터별로 패턴 수 편향이 발생하는 것은 어쩔 수 없는 문제점으로 판단된다.

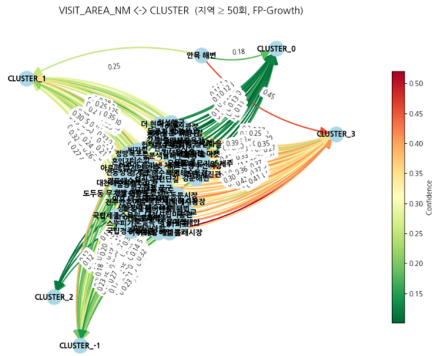



**Figure 5: 유사 클러스터 통합 후**

추가로, 클러스터링 결과의 유사성을 바탕으로 새로이 구성한 클러스터를 바탕으로도 앞선 패턴 마이닝 과정을 동일하게 수행해서 결과를 살펴보았는데, 위와 같이 기존의 문제점인 패턴 수 편향의 정도가 더욱 심해져서 기존보다 성능이 더 좋지 않은 것으로 판단된다.

#### 4.2.3 한계점 및 개선 방안

패턴 마이닝을 통해 클러스터별로 매칭되는 1차적인 추천 여행지 목록을 얻을 수 있었지만, 각 클러스터에 매칭되는 패턴의 수가 클러스터별로 생각보다 큰 값으로 차이나는 **편향 문제**가 발생하는 한계점이 존재했다.

이는 대상 데이터의 클러스터별 트랜잭션 수 자체가 차이가 나서 발생하는 문제로 보이기 때문에, **근본적인 원인을 해결해 정확도 높은 추천 시스템**을 구축하기 위해서는 이후 **추가적인 데이터를 수집**하여 각 클러스터별로 동일한 수의 고객-여행지 트랜잭션을 맞춰줄 필요가 있을 것으로 보인다.

## 4.3 Classification

#### 4.3.1 데이터셋 및 전처리

기준 데이터는 `preprocessed_with_cluster_ALL.csv`로, 여행지 정보, 사용자 정보, 클러스터링 결과가 포함되어 있다. 타겟 변수는 만족도(DGSTFN)로, 1에서 5 사이의 정수형 점수이다. 예측 변수는 범주형과 수치형으로 구성된다.

- **범주형:** `VISIT_AREA_NM`, `MAIN_TRAVEL_MONTH`, `TRAVEL_STATUS_ACCOMPANY`, `RELATION_TYPE`, `MVMN_NM`, `GENDER`, `AGE_GRP`, `VISIT_AREA_TYPE_CD`
- **수치형:** `TRAVEL_STYL_1`, `TRAVEL_STYL_3`, `TRAVEL_STYL_5`, `TRAVEL_STYL_6`, `TRAVEL_STYL_7`, `TRAVEL_STYL_8`

#### 4.3.2 분류 모델링 과정

*(1) Baseline* ***모델****: Logistic Regression (****이진 분류****).*
처음에는 DGSTFN 만족도 점수를 이진 분류 방식으로 변환하여 분석을 시작하였다.

- **기준:** 4점 이상은 만족(1), 그 미만은 불만족(0)으로 설정하였다.
- **모델:** Logistic Regression
- **특징:** `REVISIT_INTENTION`, `RCMDTN_INTENTION` 변수에 예측력이 집중되었다.
- **결과:** 해당 변수 제거 시 정확도가 급감하였고, 이는 특정 변수에 지나치게 의존한다는 점을 의미하였다.
- **한계:** 만족도를 이진으로 나누는 것은 중간 점수대(3 4점)를 적절히 반영하지 못하였다. 따라서 다중 클래스 분류로의 전환 필요성이 확인되었다.

*(2) Regressor Model.*
회귀 모델을 통해 문제를 해결하려는 시도를 진행하였다.

- **기준:** DGSTFN 만족도를 연속형 수치로 예측한 후, 반올림하여 정수형 데이터로 설정하였다.
- **모델:**
  - Random Forest
  - XGBoost ($R^2 \approx 0.0389$, MAE $\approx 0.64$)
- **결과:**
  - 예측력이 나아지지 않았으며, 여전히 낮거나 중간 점수대를 제대로 반영하지 못하였다.
  - SMOTE, `class_weight` 등의 다양한 시도를 해보았으나 눈에 띄는 변화는 없었다.
  - `VISIT_AREA_NM` 변수에 대해 frequency encoding을 시도하였고, 그 결과 $R^2 \approx 0.0499$로 소폭 개선되었다.
- **한계:** 데이터의 불균형으로 인해 낮은 점수의 데이터가 부족한 문제가 있었다.

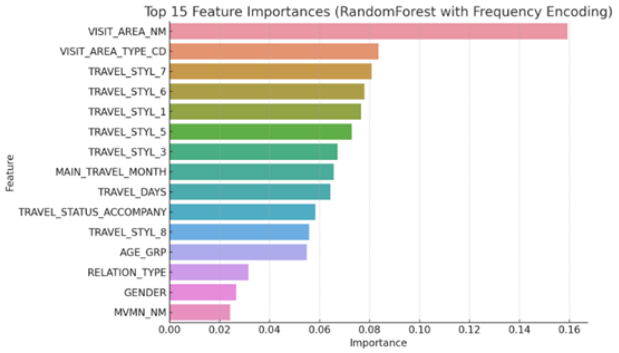



**Figure 6: Random Forest model**

*(3)* ***다중 클래스 분류*** *(Random Forest 등).*
DGSTFN 점수를 세 개의 클래스로 변환하여 다중 클래스 분류 모델을 적용하였다.

- **초기 기준:** 1~3점은 불만족(0), 4점은 보통(1), 5점은 만족(2)으로 설정하였다. 그러나 불만족과 보통 클래스에 대한 예측력이 매우 낮게 나타났다.
- **최종 기준:**
  - 1~2점 → 불만족 (0)
  - 3~4점 → 보통 (1)
  - 5점 → 만족 (2)
- **모델:**
  - Random Forest
  - XGBoost
  - 다양한 하이퍼파라미터 실험을 함께 수행하였다.
- **결과:**
  - Random Forest가 가장 안정적인 성능을 보였다.
  - 그러나 클래스 불균형으로 인해 소수 클래스에 대한 예측력이 떨어졌다.
  - 따라서 데이터 불균형을 해결할 필요성이 확인되었다.

*(4) SMOTE + class_weight* ***적용 모델 (최종 통합모델 채택).***
데이터 불균형 문제를 해결하기 위해 **SMOTE**와 **class_weight**를 함께 적용한 모델을 구성하였다.

- **SMOTE 적용:**
  - 모든 클래스에 대해 oversampling을 적용한 경우와 소수 클래스에만 oversampling을 적용한 경우를 비교하였다.
  - 실험 결과, **모든 클래스에 SMOTE 적용**한 경우가 더 나은 성능을 보여 해당 방식을 선택하였다.
- **class_weight 조정:**
  - 불만족(0) 클래스에 **가중치 3**, 보통(1), 만족(2) 클래스에는 각각 **가중치 1**을 부여하였다.
- **이상치 제거 여부:**
  - 모든 클래스 oversampling과 다수 클래스(5점) oversampling을 비교 실험하였다.
  - 데이터 수가 줄어들면서 전반적인 성능이 저하되어, **이상치 제거는 제외**하였다.
- **모델:** Random Forest
- **결과:**
  - Accuracy: 0.638
  - Macro F1: 0.460
  - 따라서 본 모델을 **최종 통합 예측 모델**로 채택하였다.
- **한계:** 모델 성능이 크게 향상되지 못한 점이 아쉬운 한계로 남는다.

Random Forest Classifier (SMOTE + class_weight)

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.28 | 0.06 | 0.09 | 308 |
| 1 | 0.63 | 0.63 | 0.63 | 5513 |
| 2 | 0.65 | 0.68 | 0.66 | 5961 |
| accuracy | | | 0.64 | 11782 |
| macro avg | 0.52 | 0.45 | 0.46 | 11782 |
| weighted avg | 0.63 | 0.64 | 0.63 | 11782 |

Accuracy: 0.6382617552198269
Confusion Matrix:
[[ 17 157 134]
[ 32 3467 2014]
[ 12 1913 4036]]

**Figure 7:** *SMOTE + class_weight* **적용 모델**

**a. 클러스터별 개별 학습 실험**
클러스터링을 기반으로 한 개별 모델 학습 실험을 수행하였다.

- 기준: 성별(GENDER), 연령대(AGE_GRP)로 그룹을 나눈 뒤, 각 그룹에서 클러스터 변수 유무에 따라 예측 모델 성능을 비교 (1,000개 미만 그룹은 제외)
- 각 그룹마다 Random Forest 모델(최종 채택 모델 동일) 학습
- 일부 클러스터만 개별 모델로 대체하는 hybrid 구조도 검토
- 결과: 일부 나은 그룹도 있었으나 대체할 수준은 아니라고 판단했다. 모든 클러스터에서 통합 모델의 결과가 더 우수했다.

| | GENDER | AGE_GRP | Samples | Accuracy | F1_불만족 | F1_보통 | F1_만족 | Macro_F1 |
|---|---|---|---|---|---|---|---|---|
| 6 | 1 | 0.50 | 4217 | 0.611374 | 0.261194 | 0.570194 | 0.710015 | 0.513801 |
| 3 | 0 | 0.75 | 1954 | 0.575809 | 0.291667 | 0.517815 | 0.683087 | 0.497523 |
| 7 | 1 | 0.75 | 2317 | 0.574713 | 0.200000 | 0.536204 | 0.678523 | 0.471575 |
| 5 | 1 | 0.25 | 7302 | 0.570059 | 0.192469 | 0.502376 | 0.685315 | 0.460053 |
| 1 | 0 | 0.25 | 5103 | 0.583279 | 0.147368 | 0.496032 | 0.703222 | 0.448874 |
| 0 | 0 | 0.00 | 5019 | 0.583665 | 0.150538 | 0.484211 | 0.704431 | 0.446393 |
| 2 | 0 | 0.50 | 3041 | 0.555312 | 0.108696 | 0.500000 | 0.674044 | 0.427580 |
| 4 | 1 | 0.00 | 8886 | 0.542011 | 0.185859 | 0.432616 | 0.662325 | 0.426933 |

**Figure 8: 개별 모델 학습 결과**

**b. 유사 클러스터 통합 학습 실험**
클러스터링 과정에서 유사한 양상을 보인 클러스터 그룹끼리 묶어 학습하였다.

- 기준 데이터: `preprocessed_with_cluster_numbering.csv`
- 통합 대상 클러스터: (0, 10), (1, 4, 7, 9), (13, 15, 20), (2, 3, 8, 11, 14, 16, 19, 23, 25, 28)

– **(0,10)**

Accuracy: 0.6320564516129032
Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.0000 | 0.0000 | 0.0000 | 25 |
| 1 | 0.6253 | 0.6439 | 0.6345 | 469 |
| 2 | 0.6552 | 0.6526 | 0.6539 | 498 |
| accuracy | | | 0.6321 | 992 |
| macro avg | 0.4268 | 0.4322 | 0.4295 | 992 |
| weighted avg | 0.6246 | 0.6321 | 0.6282 | 992 |

– **(1, 4, 7, 9)**

Accuracy: 0.6229936543486375
Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.1333 | 0.0270 | 0.0449 | 74 |
| 1 | 0.6019 | 0.5724 | 0.5868 | 1202 |
| 2 | 0.6437 | 0.6978 | 0.6696 | 1403 |
| accuracy | | | 0.6230 | 2679 |
| macro avg | 0.4596 | 0.4324 | 0.4338 | 2679 |
| weighted avg | 0.6108 | 0.6230 | 0.6152 | 2679 |

– **(13, 15, 20)**

Accuracy: 0.6475138121546962
Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.2500 | 0.0741 | 0.1143 | 27 |
| 1 | 0.6332 | 0.5868 | 0.6091 | 409 |
| 2 | 0.6641 | 0.7335 | 0.6971 | 469 |
| accuracy | | | 0.6475 | 905 |
| macro avg | 0.5158 | 0.4648 | 0.4735 | 905 |
| weighted avg | 0.6378 | 0.6475 | 0.6399 | 905 |

– **(2, 3, 8, 11, 14, 16, 19, 23, 25, 28)**

Accuracy: 0.640252138676272
Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.1538 | 0.0333 | 0.0548 | 120 |
| 1 | 0.6314 | 0.6087 | 0.6198 | 2057 |
| 2 | 0.6527 | 0.7011 | 0.6760 | 2265 |
| accuracy | | | 0.6403 | 4442 |
| macro avg | 0.4793 | 0.4477 | 0.4502 | 4442 |
| weighted avg | 0.6293 | 0.6403 | 0.6332 | 4442 |

- 방법: 해당 그룹에 대해 별도 모델 학습
- 결과: Accuracy, Macro F1 모두 통합 모델보다 낮거나 유사
- 시사점: 클러스터 별 학습과 유사 클러스터 통합 학습 모두 성능이 개선되지 않은 것을 보아, 클러스터와 예측 모델 간의 연관성이 의심됐다.

#### 4.3.3 실제 예측 적용 예시 실험

본 실험은 모델의 실용 가능성을 검토하기 위해 수행하였다. 입력 예시는 다음과 같다:

- MAIN_TRAVEL_MONTH: 8
- TRAVEL_STATUS_ACCOMPANY: 2인 여행(가족 외)
- RELATION_TYPE: 친구
- MVMN_NM: 대중교통 등
- GENDER: 1 (여성)
- AGE_GRP: 0.75 (50대)
- VISIT_AREA_TYPE_CD: 1
- TRAVEL_STYL_1: 0.7
- TRAVEL_STYL_3: 0.4
- TRAVEL_STYL_5: 0.8
- TRAVEL_STYL_6: 0.2
- TRAVEL_STYL_7: 0.9
- TRAVEL_STYL_8: 0.6

해당 조건을 기반으로 산출된 소속 클러스터는 CLUSTER_NUM = 22이다. 22번 클러스터에 대해 수행한 패턴 마이닝 결과, 다음과 같은 추천 후보 장소가 도출되었다:

- 경기전, 금능해수욕장, 비자림, 서귀포 매일 올레시장, 성산일출봉, 속초관광수산시장, 오설록 티 뮤지엄, 전주한옥마을, 정방폭포, 제주동문시장, 제주동문재래 야시장, 제주동문재래시장, 천지연폭포, 초원 사진관, 혼인지

해당 리스트를 대상으로 본 예측 모델을 적용한 결과, 예측된 만족 상태 리스트는 아래 표와 같고, “천지연폭포”와 “혼인지”가 최종 추천 장소로 선정되었다. 이를 통해 예측 결과가 실제 추천 시스템에 연동 가능한 형태로 도출 가능함을 확인하였다.

| VISIT_AREA | Predict | Satisfied_Prob | Neutral_Prob | Dissatisfied_Prob |
|---|---|---|---|---|
| 천지연폭포 | Satisfied | 0.5 | 0.4 | 0.1 |
| 혼인지 | Satisfied | 0.44 | 0.42 | 0.14 |
| 금능해수욕장 | Neutral | 0.43 | 0.44 | 0.13 |
| 속초관광수산시장 | Neutral | 0.42 | 0.46 | 0.12 |
| 오설록 티 뮤지엄 | Neutral | 0.42 | 0.44 | 0.14 |
| 서귀포 매일 올레시장 | Neutral | 0.41 | 0.49 | 0.1 |
| 제주동문재래시장 | Neutral | 0.41 | 0.44 | 0.14 |
| 성산일출봉 | Neutral | 0.4 | 0.5 | 0.1 |
| 정방폭포 | Neutral | 0.4 | 0.48 | 0.12 |
| 전주한옥마을 | Neutral | 0.38 | 0.48 | 0.14 |
| 비자림 | Neutral | 0.38 | 0.48 | 0.14 |
| 제주동문시장 | Neutral | 0.38 | 0.45 | 0.17 |
| 경기전 | Neutral | 0.37 | 0.49 | 0.14 |
| 제주동문재래 야시장 | Neutral | 0.36 | 0.52 | 0.12 |
| 초원 사진관 | Neutral | 0.35 | 0.51 | 0.14 |

**Figure 9: 예측 만족도 확률 분포 및 예측 결과**

#### 4.3.4 한계점 및 개선 방안

본 과정에서는 만족도 예측을 위한 다양한 분류 및 회귀 모델을 실험하였으나, 예측 성능 향상에는 구조적인 한계가 존재하였다. 가장 근본적인 문제는 데이터 분포 자체에 있었으며, 특히 만족도 5점 데이터가 전체의 절반을 차지하는 심각한 클래스 불균형(class imbalance) 문제가 있었다. 이는 학습된 모델이 대부분의 데이터를 5점으로 예측하더라도 일정 수준 이상의 정확도(약 50%)를 달성할 수 있게 만들었고, 결과적으로 정확도 단독으로는 모델의 성능을 평가하기 어려운 구조가 되었다.

또한, 클래스 불균형으로 인해 상대적으로 소수 클래스인 1~2점(불만족)의 예측 정확도와 F1-score가 매우 낮게 나타났으며, 이는 만족도 분류 모델의 실질적 활용 가능성을 저해하는 요인이 되었다. 이에 대한 대응으로 SMOTE, class_weight 조

정, 이상치 제거 등의 기법을 적용하였으나, 완전한 해결책은 되지 못하였다.

Feature importance 분석을 통해 타겟 변수(DGSTFN)와 관련성 있는 feature를 도출하려는 시도도 하였지만, 특정 변수 하나가 만족도에 결정적인 영향을 미친다고 보기 어려운 결과가 나타났으며, 전체적으로 feature와 만족도 간의 명확한 연관성은 약하게 나타났다.

결과적으로, 성능 개선을 위한 모델 선택 및 최적화 과정에서 명확한 기준을 도출하기 어려웠고, 시도한 여러 모델 간의 수치상 차이도 미세하여 최종 모델 채택에 있어 정량적인 평가 외에도 정성적인 판단이 요구되었다.

개선 방안은 다음과 같다.

- 클래스 불균형 문제를 완화하기 위한 데이터 보강 및 추가 수집이 필요하다.
- 사용자의 여행 이력 등 추가적인 feature 도입을 고려할 수 있다.
- 만족도 점수를 단일 수치가 아닌 경향성 기반 척도 또는 범주형 만족도로 전환하여 예측 타겟을 재설계할 수 있다.
- 협업 필터링 등의 추천 알고리즘 기반 모델 적용도 시도할 수 있다.

# 5 DISCUSSION AND CONCLUSION

## 5.1 Limitations

이번 프로젝트에서 실험한 다양한 모델들은 기대에 비해 뚜렷한 성능 향상을 보여주지 못하였다. 그러나 이는 특정 알고리즘이나 분석 방식의 실패라기보다는, 현재 보유한 데이터의 한계가 명확히 드러난 결과라고 해석할 수 있다.

가장 큰 구조적 문제는 클래스 불균형이었다. 만족도 5점 데이터가 전체의 절반 이상을 차지하면서, 모델이 모든 데이터를 5점으로 예측해도 50

또한, 피처와 타겟 간의 연관성이 뚜렷하게 나타나지 않은 것도 제한 요인 중 하나였다. 특정 변수 하나가 만족도에 결정적으로 작용하지 않았고, 전체적으로 피처 중요도가 분산되어 있었다. 이로 인해 성능 개선을 위한 방향성이 뚜렷하게 보이지 않았고, 실험 간 성능 차이도 미세하여 정량적인 기준만으로 최적 모델을 결정하는 데 어려움이 있었다.

## 5.2 Insight and Feedback

그럼에도 불구하고 본 프로젝트는 여러 흥미로운 관찰을 통해 시사점을 남겼다. 여행이라는 활동은 선택지가 구조적으로 제한된 도메인이라는 점을 다시금 확인할 수 있었다. 예컨대 "서울로 여행 간다"고 했을 때 방문지는 남산, 북촌, 경복궁 등으로 집중되며, 성별이나 나이, 여행 목적이 달라도 겹치는 관광지가 발생할 수밖에 없는 구조라는 것이다.

이러한 특성 때문에 클러스터 간 방문지 차이가 통계적으로 명확히 나타나지 않았고, 패턴 마이닝 결과 또한 강한 연관 규칙을 만들기 어려웠다. 그러나 이는 데이터가 무의미하다는 뜻이 아니라, 여행이라는 영역이 원래부터 단선적인 선택지를 중심으로 구성되어 있다는 점을 반영한 자연스러운 결과로 해석해야 한다.

이러한 현실적 제약 속에서도 프로젝트가 가진 가장 큰 의의는, 바로 그 '작은 차이'를 포착하려는 시도에 있다. 예를 들어, 트렌드에 민감한 20대 커플과 조부모님과 아들 딸이 함께 떠나는 대가족 여행은 같은 지역을 방문하더라도 그 안에서 선택하는 식당, 숙박, 사진 촬영 중요도 등이 분명히 다를 수 있다는 점을 고려한 분석이 이루어졌다.

또한 프로젝트는 단지 높은 정량적 성능을 목표로 하기보다는, 실제 서비스에 가까운 구조에서 현실적으로 해석 가능한 결과를 도출하는 데 방점을 두었으며, 이는 실험적 접근이 갖는 가치를 잘 보여주는 예시가 되었다.

## 5.3 Meaning

이번 프로젝트는 단순히 정확도가 높은 모델을 찾는 데 집중한 실험이 아니었다. 우리는 여행자의 성향과 맥락을 반영하여 단순한 인기 기반 추천을 넘어선 개인화된 추천 시스템이 구현 가능한지 그 가능성을 검토하고자 하였다.

비록 만족스러운 결과가 도출되지 않았더라도, 그 과정을 통해 어떤 데이터를 어떻게 활용해야 개인화된 추천이 실현 가능한가에 대한 중요한 단서를 얻을 수 있었다. 특히 소비 내역, 텍스트 리뷰, 동선 등 추가적인 맥락 데이터와의 결합이 이루어진다면, 지금보다 훨씬 정교한 개인화가 가능할 수 있다는 방향성을 확인하였다.

또한, 이번 프로젝트는 단순한 알고리즘 구현을 넘어서, 데이터를 해석하고 실제 서비스 설계나 정책 수립으로의 확장 가능성까지 고려한 실질적 실험 기반을 제시했다는 데에서 의미가 있다.

실험 결과만 놓고 보면 전체적인 예측 성능은 제한적이었지만, 그 과정에서 우리는 '데이터의 현실적 제약'을 반영한 결과가 오히려 자연스럽고 의미 있는 해석이 될 수 있다는 점을 발견했다.

예컨대 만족도 예측이 잘 되지 않았던 이유는 모델의 미흡함이 아니라, 애초에 만족도라는 지표 자체가 사용자마다 상대적이고 맥락에 따라 다르게 형성된다는 점 때문일 수 있다. 즉, 우리가 사용한 만족도 점수는 절대적인 '정답'이라기보다, 그 자체로 노이즈와 해석 여지가 많은 변수였던 것이다.

그럼에도 불구하고, 우리는 '작은 차이' 속에서 사용자 그룹 간의 경험 차이를 찾으려 했고, 그 시도는 단지 결과 수치보다도 더 중요한 발견이었다. 결과적으로 이번 프로젝트는 데이터 기반 개인 맞춤형 관광지 추천 시스템이 어떤 요소들을 고려해야 실현 가능한가에 대한 실험적 통찰을 제공하였다.

# 6 ROLES AND RESPONSIBILITIES

**정원홍**
: 프로젝트 구조 기획 및 발표 자료 제작, Pattern Mining, 실험 및 결과 분석, 최종 발표

**박지원**
: 프로젝트 구조 기획 및 발표 자료 제작, Data Preprocessing, Clustering, 실험 및 결과 분석, 최종 발표

**최윤재**
: 프로젝트 구조 기획 및 발표 자료 제작, Classification, 실험 및 결과 분석, 최종 발표

**우혜민**
: 프로젝트 구조 기획 및 발표 자료 제작, 프로젝트 워크로드 조정, Introduction, Discussion and Conclusion 작성, LaTeX 문서 작업, 중간/최종 발표