



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

석사학위논문

중국 개별관광객의 행태를 반영한
머신러닝기반 관광지추천 기법

제주대학교 대학원

컴퓨터 공학과

진 준 결

2021년 2월

중국 개별관광객의 행태를 반영한 머신러닝기반 관광지추천 기법

지도교수 변 영 철

진 준 걸

이 논문을 컴퓨터공학 석사 논문으로 제출함

2021年 2月

진준걸의 컴퓨터공학 석사학위 논문을 인준함

심사위원장 홍상훈 

위 원 이동철 

위 원 변영철 

제주대학교 대학원

2021年 2月

Machine learning-based tourist attraction
recommendation reflecting the behavior of individual
Chinese tourists

Jun-Jie Qin

(Supervised by professor Yung-Cheol Byun)

A thesis submitted in partial fulfillment of the requirement for
the degree of Master of Science in Computer Engineering

2021 . 2 .

This thesis has been examined and approved.

Sang-Joon Lee

Dong-Cheol Lee

Yung-cheol Byun

(Name and signature)

Date.

Department of Computer Engineering
GRADUATE SCHOOL
JEJU NATIONAL UNIVERSITY

중국 개별관광객의 행태를 반영한 머신러닝기반 관광지추천 기법
Machine learning-based tourist attraction recommendation reflecting the
behavior of individual Chinese tourists

컴퓨터공학과 진 준 걸
지도 교수 변 영 철

제주도는 중국 관광객에게 매력적인 방문지로 알려져 있다. 과거 수년 동안 제주도를 방문하는 모든 외국인 관광객 비중에서도 중국은 단연 1위를 기록해 왔다. 사드 배치로 인하여 단체관광이 금지됨에 따라 제주도를 방문한 중국 관광객은 주로 단체 관광객이 아닌 개별관광객으로 방문하고 있다. 대부분의 개별관광객은 여행하기 전에 인터넷 홈페이지나 애플리케이션을 이용해 관광지 정보, 음식 정보, 쇼핑 정보 등을 검색하여 마음에 드는 관광지를 선택하고 스케줄을 계획하는 특성이 있다. 따라서 적합한 추천 시스템은 관광객의 관광지 선택에 도움을 준다. 적절한 관광지 추천은 관광객에게 좋은 기억을 선사하고 재방문 의사도 높이는 반면에, 잘못된 추천은 손님의 금전적, 시간적 낭비를 초래할 뿐만 아니라 재미없고 힘든 여행이 초래할 수도 있다. 이로 인해 제주도 관광이 좋은 평가를 받지 못하고 관광 만족도도 떨어진다면 당연히 재방문 의사도 감소하여 제주관광에 부정적인 영향을 미치게 될 수 있다. 따라서 관광객의 선호에 따른 관광지 추천 시스템에 대한 연구가 필요하다고 할 수 있다.

기존의 추천 시스템은 크게 두 가지 방식이 있다. 아이템이나 사용자의 유사도 (similarity)를 모델링하고 측정하여 추천하는 협업 필터링 (collaborative filtering approach) 방식과 콘텐츠 자체를 분석하여 추천하는 콘텐츠 기반 필터링 (contents-based filtering approach) 방식이 있다.

관광지를 추천할 때 관광객의 성별, 연령대, 월수입, 제주도를 알게 된 경로, 방문 목적, 여행 동반자, 제주방문 횟수 등 여러 가지 속성과 형태에 따라 선호하는 관광지가 달라진다. 방문과 관련된 요소들이 모이고 데이터 분석을 통해 모델

을 만들어 사용자가 만족할만한 관광지를 추천할 수 있다. 따라서 본 연구에서는 이미 제주를 방문했던 관광객의 속성과 형태, 그리고 관광지에 대한 만족도 평가를 토대로 머신러닝 알고리즘으로 선호 예측 모델을 만들고, 이를 통해 머신러닝 기반 관광지 추천 시스템을 제안한다.

본 논문에서는 이전에 제주도를 방문했던 중국 개별관광객들로부터 얻은 데이터를 이용했으며, 이를 분석한 후 논리 회귀(logistic regression) 모델, 결정 트리(decision tree) 모델, K-근접 이웃(k-nearest neighbours) 모델, 서포트 벡터 머신(support vector machine, SVM) 모델, XGBoost 분류기 등을 이용하여 실험했다. 실험 결과 XGBoost 분류기를 이용한 관광지 추천이 평균 정확도가 83.59%에 나타나 다른 모델에 비해 나은 성능을 보여주었다.

목 차

국문 초록	i
목 차	iii
그림 목차	v
표 목차	vi
I. 서론	1
1.1 연구 배경 및 목적	1
1.1.1 연구 배경	1
1.1.2 연구 목적	3
1.2 연구 범위 및 방법	3
1.3 연구의 구성	4
II. 이론적 배경	5
2.1 제주도 방문한 중국 관광객 현황	5
2.2 개별관광객 개념	6
2.3 중국 개별관광객 특성	7
2.4 관광객의 속성과 형태	7
III. 관련 연구	8
3.1 추천 시스템	8
3.1.1 추천 시스템의 분류	8
3.1.2 콘텐츠 기반 필터링	9
3.1.3 협업 필터링	10
3.2 머신러닝 기법	12
3.2.1 근접 이웃(k-nearest neighbors) 기법	12
3.2.2 로지스틱 회귀(logistic regression) 기법	13
3.2.3 서포트 벡터 머신(support vector machines) 기법	14
3.2.4 결정 트리(decision trees) 기법	17
3.2.5 XGBoost (extreme gradient boosting) 기법	17
3.2.6 앙상블 Stacking 기법	19

3.3 머신러닝 모델 성능 평가 방법	19
3.4 관광지 추천 시스템 선행 연구	20
3.4.1 개인화 관광지 추천 시스템	20
3.4.2 빅데이터를 이용한 관광지 추천 시스템	21
3.4.3 성격유형에 따른 관광지 추천에 관한 연구	22
3.5 관광지 추천 시스템 문제점	23
IV. 제안하는 방법	24
4.1 데이터 수집	25
4.1.1 설문지 구성	25
4.1.2 설문조사 실시 내용	25
4.2 기본 데이터 분석	26
4.2.1 관광객 속성(인구통계학)	26
4.2.2 관광객 형태 분석	27
4.2.3 데이터 가시화 분석	29
4.3 머신러닝을 이용한 관광지 추천	33
4.3.1 구현 환경	33
4.3.2 머신러닝을 이용한 처리 과정	34
4.3.3 데이터 속성	34
4.3.4 학습 데이터 셋 구성	35
4.3.5 머신러닝 알고리즘에 따른 실험 결과	35
4.3.5.1 Logistic Regression	36
4.3.5.2 Decision Tree	36
4.3.5.3 K-Nearest Neighbours	37
4.3.5.4 Support Vector Machine (SVM)	38
4.3.5.5 XGBoost	38
4.3.5.6 Stacking	39
V 실험 결과	40
VI 결론	41
첨부자료	42

참고문헌	49
ABSTRACT	51

그림 목차

[그림 2-1] 제주 외래 관광객과 제주 중국인 관광객 현황	5
[그림 2-2] 개별관광객 속성 및 형태	7
[그림 3-1] 추천 시스템 흐름도	8
[그림 3-2] 추천 시스템의 분류	9
[그림 3-3] 사용자 기반 협업 필터링	10
[그림 3-4] 아이템 기반 협업 필터링	11
[그림 3-5] KNN로 데이터 분류	12
[그림 3-6] 시그모이드 함수의 그래프	13
[그림 3-7] 데이터1(빨강)과 데이터 2(파랑)를 구분하는 초평면 A1와 A2	15
[그림 3-8] 저차원 데이터	15
[그림 3-9] 고차원 전환된 데이터	16
[그림 3-10] 결정 트리 예식	17
[그림 3-11] 일반 트리 분석	18
[그림 3-12] XGBoost 분류기로 데이터 분석	18
[그림 3-13] Stacking 기법 흐름도	19
[그림 3-14] 개인화 추천 시스템	20
[그림 3-15] 빅데이터 기반 추천	21
[그림 3-16] Construction of a tourism recommendation system using a hybrid web app-based topic map	22
[그림 4-1] 관광지 추천 방법	24
[그림 4-2] 질문 답변 빈도도	30
[그림 4-3] 성별별, 나이별 자연 관광지 선택 빈도	31

[그림 4-4] 각 항목 간 상관관계	32
[그림 4-5] 머신러닝을 이용한 처리 과정	34
[그림 4-6] 학습 데이터 테스트 데이터 세팅 코드	35
[그림 4-7] Decision Tree 모델 분석 n_neighbors도	36
[그림 4-8] K-Nearest Neighbours 모델 분석 max_depth도	37
[그림 4-9] XGBoost 분류기 모델 분석 max_depth도	39

표 목차

<표 2-1> 개별관광객의 개념	6
<표 3-1> SVC 커널 (kernel trick) 트릭	16
<표 3-2> 머신러닝 평가 혼동행렬	19
<표 4-1> 설문지 구성표	25
<표 4-2> 설문조사 실시 내용	26
<표 4-3> 관광객 속성 통계분석(인구통계학)	27
<표 4-4> 관광객 형태 통계분석	29
<표 4-5> 구현 환경	33
<표 4-6> 관광객 속성, 행태 영어 약어로 표기	34
<표 4-7> 데이터 셋 구성	35
<표 4-8> Logistic Regression 모델 추천 정확도	36
<표 4-9> Decision Tree 모델 추천 정확도	37
<표 4-10> K-Nearest Neighbours 모델 추천 정확도	38
<표 4-11> Support Vector Machine (SVM) 모델 추천 정확도	38
<표 4-12> XGBoost 분류기 모델 추천 정확도	39
<표 4-13> Stacking 기법 모델 추천 정확도	39
<표 4-14> 알고리즘별 추천 정확도	40

I. 서론

1.1 연구 배경 및 목적

1.1.1 연구 배경

세계 경제 성장에 따라 각국 사람들이 소비 형태가 다양해지고 있다. 많은 사람이 다른 나라에 가서 휴식, 체험, 문화 탐방 관광 등 다양한 관광 활동이 활발히 일어나고 있다. 한국관광공사, 한국 관광통계(2019) 자료에 따르면 한국에 방문한 외래 관광객이 2017년 13,335,758명을 기록했고 2018년 15,346,879명으로 2017년보다 15.1%가 증가했으며 2019년 17,502,756명으로 2018년보다 14%가 증가했다¹⁾. 관광산업은 생산 유발 효과와 고용 창출 효과가 높은 고부가가치 산업으로 21세기 으뜸 산업으로 주목받고 있다. 대다수의 국가는 관광산업의 진흥을 위해 적극적으로 나서고 있으며 관광사업을 추진하기 위해 많은 유지 확대 방안을 마련하고 있다²⁾.

중국 경제의 꾸준한 발전과 주민 소득의 꾸준한 증가에 따라 도시 및 농촌 주민들의 소비 구조도 큰 변화가 발생하고 있다. 소비는 실물 소비에서 관광, 문화, 교육, 건강, 노후보장 등의 서비스 소비로 변화했으며, 현재 중국의 해외 관광객 인구수와 해외 관광 지출은 세계 1위를 기록했다³⁾. 2019년 중국인 해외 관광 인구수는 연간 1.55억 명으로 크게 증가했다. 2019년 중국인 해외여행 국가별 비중표를 보면 한국에 방문한 중국 관광객 수는 총 해외여행 관광객 수의 14%를 차지해 2위를 기록했다⁴⁾. 한국에 국가별 입국자 수 추이 표를 보면 2019년 한국을 방문한 중국 관광객 수는 총 6,023,021명으로 2018년 중국 관광객 입국자 수 총 4,789,512명에 비해 25.8%가 늘었다⁵⁾. 2008년 이후의 관광추세는 단체관광에서

1) 한국관광공사(2020). 2019 외래 관광객 조사 최종 보고서

2) 김종열(2014). 국내 관광산업의 관광수지 결정요인에 관한 연구, 부산대학교 경제 통상대학원 석사학위 논문, pp 1-2.

3) 중국 통신원 (2019).〈중국 정보 소비발전태세 및 전망보고서〉

4) 중국산업정보망(2020.3). <http://www.chyxx.com/industry/202003/845172.html>

5) 한국관광공사(2019). 2019년 한국 외래 관광객 실태조사 최종보고서

개별 관광으로 점차 변화하고 있다⁶⁾. 한국관광공사의 외래 관광객 실태조사에 따르면 2013년 한국 방문 관광객 중에 66.2%는 개별관광객이다⁷⁾.

제주도는 중국 관광객에게 매우 매력적인 방문지로 알려져 있다. 제주도 홈페이지 공개 정보에 따르면 2019년 기준 외국인 관광객은 총 1,726,132명이다. 그중 중국인 관광객이 1,079,133명으로 1위를 기록했다. 2018년도 중국인 제주도 방문 관광객 수 총 666,120명에 비하여 무려 62% 증가해 큰 폭의 성장세를 기록했으며 외국인 총 방문 관광객 숫자 비중도 54.3%에서 62.5%로 상승세를 보였다⁸⁾.

사드 배치로 인하여 중한 관계가 악화하여 한국을 방문한 단체 관광객이 급감하며 현재 거의 중단된 상태이다, 2019년 중국 문화관광부 자료에 따르면 한국에 방문한 중국단체 관광객은 거의 없었다고 발표했다⁹⁾. 이를 근거로 2019년에 한국을 방문하는 관광객은 개별관광객이라는 사실을 알 수 있다.

개별관광객은 여행사 도움 없이 스스로 여행 스케줄을 계획하고 타 문화의 고유성을 체험하며 항공권, 교통, 숙박 등 부분적으로 여행사의 도움을 받는 독립 여행자로 구분했다¹⁰⁾. 대부분 개별관광객은 해외 관광 전에 인터넷 홈페이지나 애플리케이션을 이용해 관광지 정보, 음식 정보, 쇼핑 정보 등을 검색하여 마음에 드는 관광지를 선택하고 스케줄을 계획한다.

이에 따라 관광지 추천 시스템은 관광객에게 여행지를 선택하는 데 도움을 주는 중요한 역할을 할 수 있다. 적절한 관광지 추천은 관광객에게 좋은 기억을 선사하고, 재방문 의사도 높이는 반면에, 잘못된 추천은 손님에 금전적, 시간적 낭비를 초래할 뿐만 아니라 재미없고 힘든 여행을 선사할 수 있다. 이로 인해 제주도 관광이 좋은 평가를 받지 못하고 관광 만족도도 떨어진다면 당연히 재방문 의사도 감소하여 제주 관광에 부정적인 영향을 미치게 될 수 있다. 따라서 관광객의 선호에 적합한 관광지 추천 시스템에 관한 연구가 필요하다고 할 수 있다.

관광지를 추천할 때 관광객의 성별, 연령대, 월수입, 제주도를 알게 된 경로, 방문 목적, 여행 동반자, 제주 방문 횟수 등 여러 가지 속성과 형태에 따라 선호

6) 최학현·오세형·이주현·김정희(2008). 실버세대의 사용 편의성을 위한 Digital(On-line) TV 인터페이스 디자인, p 384.

7) 한국관광공사(2014). 2013년 한국 외래 관광객 실태 조사 보고서.

8) 제주도청 홈페이지(2020) 관광객 입도 현황 표

9) 중국문화관광부(2020.8) <http://zwgk.mct.gov.cn/zfxxgkml/>

10) 김사현(2003). 비제도권 개별관광 자의 속성과 영향. 관광학 연구, 25(1), pp. 131-148.

하는 관광지가 달라진다. 방문과 관련된 요소들이 모이고 데이터 분석을 통해 모델을 만들어 사용자들이 만족할 만한 관광지 추천이 될 수 있다.

따라서 본 연구에서는 이미 제주를 방문했던 관광객의 속성과 형태, 그리고 관광지에 대한 만족도 평가를 토대로 머신러닝 알고리즘으로 선호 예측 모델을 만들고 이를 통해 머신러닝 기반 관광지 추천 시스템을 제안한다.

1.1.2 연구 목적

본 연구는 중국 개별관광객에게 알맞은 관광 정보를 추천하고 더 좋은 관광서비스를 제공하는 목적에서 다음 3가지 세부 목적을 중심으로 수행했다.

- 첫째 이전에 제주도에 방문했던 관광객한테 수집한 데이터를 인구통계학 기준으로 분석해서 관광객의 속성 및 형태를 알고자 한다.
- 둘째 머신러닝을 통해 데이터를 분석하여 선호 예측 모델을 만든 뒤, 이 모델을 기반으로 필터링을 하는 머신러닝 관광지 추천 시스템을 연구한다.
- 셋째 연구된 머신러닝 추천 시스템을 바탕으로 제주도 관광산업에 도움이 되 고자 한다.

1.2 연구의 범위와 방법

본 연구는 이전에 제주도에 방문했던 중국인 개별관광객의 속성, 형태 그리고 관광지에 대한 만족도 평가를 토대로 머신러닝 알고리즘으로 선호 예측 모델을 만들고 이를 통해 머신러닝 기반 관광지 추천 시스템을 개발하기 위해 다음과 같이 공간적 범위, 시간적 범위, 내용적 범위를 수행한다.

첫째 연구 공간적 범위로 한 여행사 관광서비스를 이용한 중국 개별관광객 대상으로 설문 조사를 수행했다. 둘째, 연구의 시간적 범위로 2020년 8월 27일부터 9월 10일까지 15일간 걸쳐 2019년 1월 - 2020년 2월 간 제주도 방문했던 관광객에게 온라인 설문 조사를 수행했다. 셋째, 연구의 내용적 범위로 관광객의 속성, 형태 및 관광지에 대한 선호도 평가 등을 선정했다.

구체적인 연구 방법은 다음과 같다. 첫째 관광시장에 대한 문헌 고찰을 통해

필요한 관광개념 및 현황에 대한 이론을 정립했다. 둘째, 인구통계학 질문, 제주도 관광지 테마별 선호도 질문 등을 준비해 설문지를 구성했다. 셋째 설문지를 배포하고 데이터를 수집했다. 넷째 머신러닝 알고리즘 및 기법을 이용하여 수집한 데이터를 분석하고 관광지 선호 예측 모델을 만들었다.

1.3 연구의 구성

본 연구는 총 5장으로 구성되었으며, 각 장의 내용을 요약하면 다음과 같다.

우선 제1장에서는 서론으로 본 연구의 연구 배경 및 목적, 연구 범위와 방법하고 연구 구성 등에 관해서 기술한다.

제2장은 이론적 배경에 대한 부분으로 제주도 방문한 중국 관광객 현황, 개별 관광객의 개념, 개별관광객의 특성에 대해 살펴본다.

제3장에서는 관련 연구로 추천 시스템, 머신러닝 기법 및 평가 방법, 관광지 추천 선행 연구하고 관광지 추천 시스템 문제점에 관해 설명한다.

제4장에서는 연구 설계로 데이터 수집, 머신러닝 분석, 모델 성능평가에 관해 기술한다.

제5장 결론으로 본 연구의 결론을 제시한다.

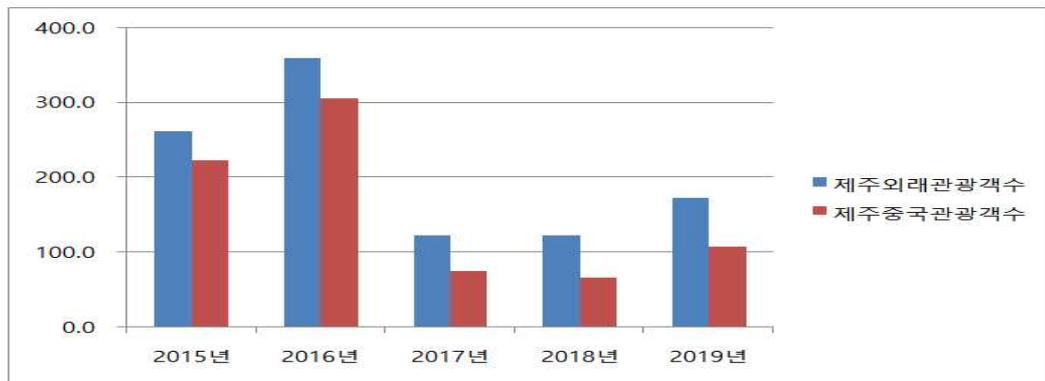
II. 이론적 배경

2.1 제주도 방문한 중국 관광객 현황

제주특별자치도 관광협회 제주 관광객 통계자료에 따르면 2015년에 제주를 방문한 외국인 관광객 수는 262만 명을 기록했으며 이 중에 중국 관광객은 223만 명을 기록했다. 2016년에 제주를 방문한 외국인 관광객 수는 360만 명을 돌파했고 이 중에 중국 관광객은 306만 명을 돌파했고 전체 외래 관광객 비중의 85%를 돌파했다.

사드(THAAD) 배치로 인해 단체 관광객의 한국 관광을 2017년 3월 15일부터 전면 금지함에 따라 2017년 제주도 방문한 중국 관광객 수가 74.4만 명으로 급감하면서 2016년 대비 24.4% 감소했다. 2018년에는 66.6만 명으로 감소했다¹¹⁾.

2019년에 한중 관계가 발전함으로 제주를 방문한 중국 관광객 수는 급증하여 107.9만 명을 기록했고 2018년 대비 61.7%로 상승했다¹²⁾. 제주 외래 관광객과 제주 중국인 관광객 현황은 [그림 2-1] 과 같다



[그림 2-1] 제주 외래 관광객과 제주 중국인 관광객 현황

출처: 제주특별자치도 관광협회(2019), 제주 관광 통계, 저자 작성

11) (오정 2018) 제주방문 중국인 개별관광객의 숙박시설 선택속성과 만족도에 관한 연구, p 56.

12) 한국관광공사(2020) 2019 외래 관광객조사 최종보고서

2.2 개별관광객 개념

개별관광객은 영어로 Foreign Independent Tour로 표시하고 약어로 FIT 말하고 있다. 한국관광공사(2006)는 개별관광객을 외래 개별관광객, 개별 또는 친구 가족 등 소그룹의 자발적인 여행 형태, 패키지 단체여행보다 상대적으로 자유로운 여행으로 정의하고 있다. 이는 패키지 단체여행에 비하면 상대적으로 자유롭다¹³⁾. 김사현(2003)은 개별 자유 여행자로 여행사 도움 없이 스스로 여행계획과 일정을 짜고 실행하여 타 문화의 고유성을 체험하는 자유 여행자로 항공권, 교통, 숙박 등 부분적으로 여행사의 도움을 받는 단독 여행자로 구분했다¹⁴⁾. 개별관광객의 개념 정리는 <표 2-1>과 같다.

기관 및 연구자	개념
한국관광공사(2006)	여행계획을 세우고 혼자, 친구, 가족 등 소규모로 자유롭게 여행하는 관광객으로 단체여행보다 시간, 교통, 숙박 등 더 독자적인 결정을 하는 개별관광객
김사현(2003)	항공권, 숙박, 여행 가이드, 교통 등의 도움 없이 자기 스스로 독자적인 여행계획을 세우고 관광이나 문화체험을 위해 여행하는 관광객
우려(2013)	단체여행보다 상대적으로 자유롭게 여행하는 단독 또는 소규모의 개별관광객
최경은(2011)	여행 일정을 관광객 스스로 독자적으로 결정하여 실행에 옮기는 관광객
추승우장준호(2010)	여행에 관한 전반적인 사항을 계획하는 개별관광객
조희진(2015)	여러 가지 옵션에 기반을 두어 에어텔(Air-tel)방식까지 포함하여 개별적으로 관광하는 관광객

<표 2-1> 개별관광객의 개념 (출처: 진문도, 2018, 저자 정리)

13) 한국관광공사(2008). www.visitkorea.or.kr.

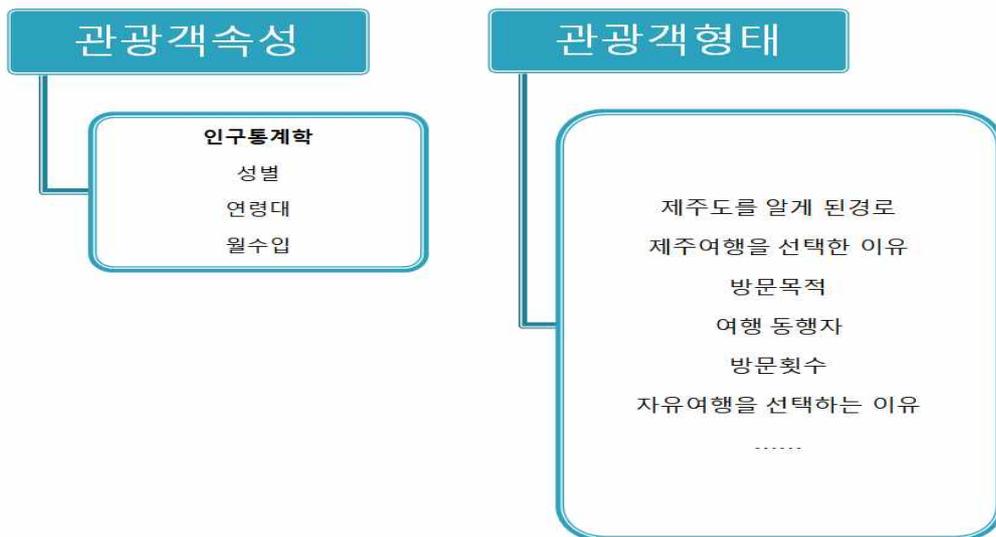
14) 김사현(2003). 비제도권 개별관광 자의 속성과 영향. 관광학 연구, pp. 131-148.

2.3 중국 개별관광객 특성

대부분 중국 개별관광객은 사전에 관광지 정보를 검색하여 관광지를 알아보는 특성이 있다. 중국 개별관광객이 선호하는 개별 자유여행 상품은 항공권, 숙박, 부가서비스(입장권, 일일 투어 등)이며, 중국인 개별관광객의 60%는 부가서비스를 구매하는 것으로 나타났다. 개별 관광 상품 구입경로는 온라인 주문(82%)이 대부분을 차지했다. 인터넷 검색을 통해 관광 일정을 직접 계획한 것으로 나타났다¹⁵⁾.

2.4 관광객의 속성과 형태

본 논문에서 관광객 속성은 성별, 연령대, 월수입을 토대로 조사했다. 또한 관광객 형태는 제주도를 알게 된 경로, 제주 여행을 선택한 이유, 방문 목적, 여행 파트너, 방문 횟수, 자유여행을 선택한 이유 등을 이용해서 조사했다.



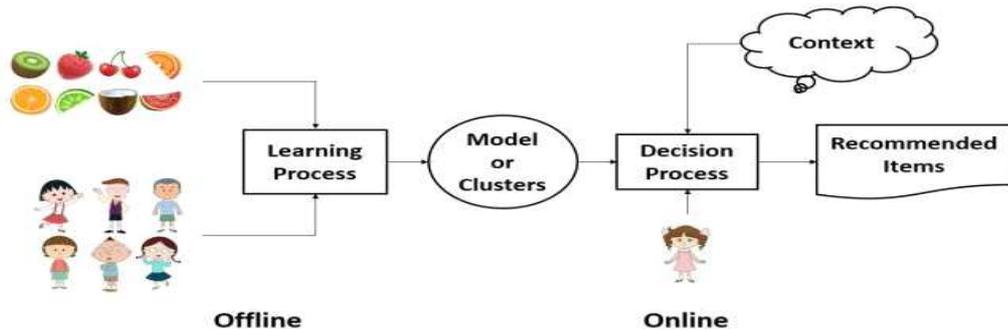
[그림 2-2] 개별관광객 속성 및 형태

15) 김도훈, 김영규(2017). 순수 관광목적을 가진 중국인 개별 자유 여행객의 한국 관광지 선택속성중요도-만족도

Ⅲ. 관련 연구

3.1 추천 시스템

추천 시스템은 시스템이 보유하고 있는 데이터를 활용하여 사용자의 특성에 맞게 사용자가 좋아할 만한 것을 각자에게 맞게 개인화하여 제공하는 것을 말한다¹⁶⁾. 정보 시스템은 사용자에게 개인 신상, 관심 분야, 선호도 등을 질의하여 사용자의 정보 프로파일을 획득하는 기법이다. 추천 시스템은 이러한 정보를 기반으로 고객의 심리 정보와 선호도 정보에 알맞은 정보 및 상품을 추천하거나 제공하는 방법이다. 영화, 음악, 뉴스, 책, 연구 주제, 탐색 질의, 상품 등 검색에 적용될 수 있다¹⁷⁾. 추천 시스템 흐름도는 아래 [그림 3-1] 과 같다.



[그림 3-1] 추천 시스템 흐름도

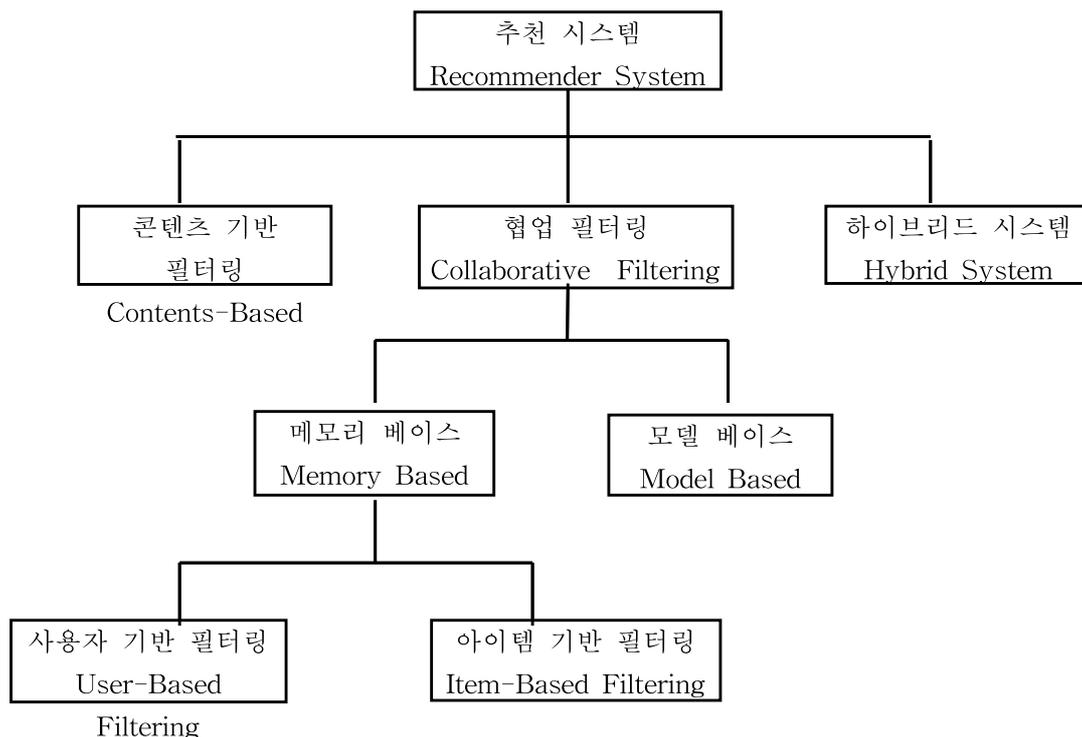
3.1.1 추천 시스템의 분류

추천 시스템은 크게 두 가지 방식이 있다. 아이템이나 사용자를 분석하여 비슷한 아이템을 추천한 콘텐츠 기반 필터링(contents-based filtering approach)과 아이템이나 사용자를 분석하지 않고 사용자의 평가 명세를 이용하는 협업 필터링

16) (조현제·이필규, 2014) 추천 시스템 기반 초등학교 인공지능교육프로그램 개발 및 적용, p8.

17) [네이버 지식백과] 추천 시스템

(collaborative filtering approach) 이 있다. 또한, 콘텐츠 기반 필터링과 협업 필터링을 섞어 사용하는 혼용 하이브리드 시스템(Hybrid System)이 있다. 아래 [그림 3-2] 와 같이 추천 시스템의 구체적인 구성 내용이다¹⁸⁾.



[그림 3-2] 추천 시스템 의 분류

3.1.2 콘텐츠 기반 필터링

콘텐츠 기반 접근(contents-based approach)은 item에 대한 설명(description) 과 사용자 선호에 대한 profile을 기반으로 한다¹⁹⁾. 아이템과 사용자 간의 액션을 분석하는 것이 아니라 콘텐츠 자체를 분석하기 때문에 많은 양의 사용자의 액션을 요구하지 않는다는 것이 장점이 있다. 영화추천을 많이 사용하고 있다. 예를 들면 사용자가 “신화” 영화를 봤다. 거기 주연 김희선이 출연되는 분석 결과를 추출했다. 만약 사용자는 김희선이 주연으로 출연된 천하 영웅을 안 봤으면 천하 영웅을 사용자한테 추천할 수 있다. 콘텐츠 기반 접근은 비슷한 아이템끼리만 추

18) 류슬기(2019) 음악 가사의 분석을 통한 사용자기반의 음악 추천 시스템 제안, pp 4-5.

19)[위키백과]. 추천 시스템 Content-based filtering

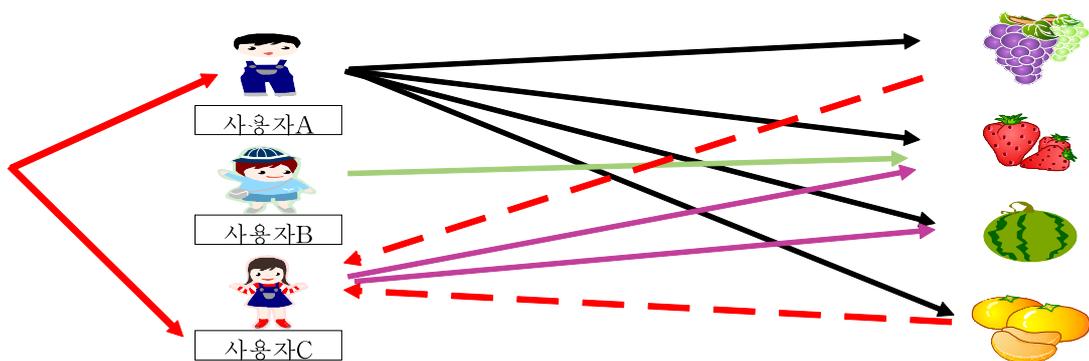
천을 가능하여 추천 범위가 제한된다. 또한 복잡한 콘텐츠를 분석하면 많은 시간이 필요하다는 단점은 있다.

3.1.3 협업 필터링

협업 필터링 (collaborative filtering approach) 방식은 사용자 행동, 활동 또는 선호도에 대한 많은 정보를 분석하고 모으고 다른 사용자와의 비슷함에 기초를 두고 사용자들이 무엇을 좋아할지를 예측하는 것에 기초를 두고 있다²⁰⁾. 아이템이나 사용자의 유사도(similarity)를 모델링하고 측정하여 추천해서 가장 큰 장점은 복잡한 아이템을 어렵게 분석하지 않아도 된다, 하지만 많은 사용자의 행동 패턴 정보들을 필요한 cold start 문제 있다.

협업 필터링(collaborative filtering approach)은 메모리 기반(Memory-based), 모델 기반(Model-based)과 총 2가지로 나뉜다. 메모리 기반 협업 필터링은 말 그대로 메모리, 즉 데이터 값 그 자체를 기반으로 필터링을 하는 것이다. 메모리 기반 협업 필터링은 다시 사용자 기반 협업 필터링(User-baser collaborative filtering)과 아이템 협업 필터링(Item-baser collaborative filtering) 두 가지 방법으로 나뉜다.

사용자 기반 협업 필터링(User-baser collaborative filtering)은 특정 사용자의 상품에 대한 선호도를 예측하기 위해 유사도 측정 지표를 이용해 유사한 선호도를 하는 이웃들을 정한다²¹⁾.



[그림 3-3] 사용자 기반 협업 필터링

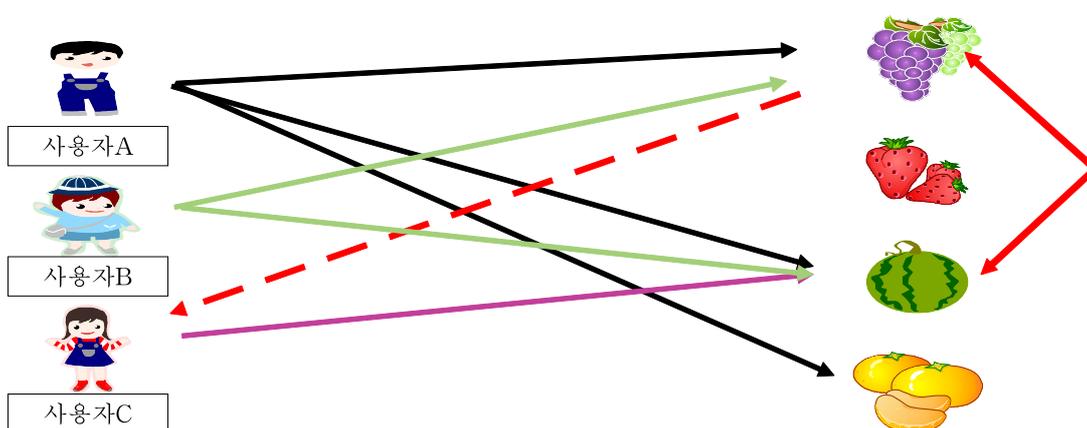
위에 [그림 3-3] 을 보면 사용자 A는 포도, 딸기, 수박, 감귤을 좋아한다. 사용

20) [위키백과]. 추천 시스템 <https://ko.wikipedia.org/wiki/>

21) 김하연(2018). 사용자 기반 협업 필터링 알고리즘의 이웃 확장 연구, pp 8-11.

자 B는 딸기를 좋아한다. 사용자 C는 딸기, 수박을 좋아한다. 그럼 사용자 A하고 사용자 B가 유사도 높다고 판단해서 사용자 A가 좋아하는 포도, 감귤을 사용자 C한테 추천할 수 있다.

아이템 기반 협업 필터링(Item-baser collaborative filtering)은 사용자가 과거에 평가한 각각의 상품들의 기록과 그 사용자가 선호도를 예측하고자 하는 상품의 유사도 정도를 계산하여 k개의 가장 유사도가 높은 상품을 선택해 추천하는 방법이다. 가장 비슷한 k개의 상품이 찾아지면 예측하고자 하는 상품을 추천한다²²⁾.



[그림 3-4] 아이템 기반 협업 필터링

위에 [그림 3-4] 를 보면 사용자 A는 포도, 수박, 감귤을 좋아한다. 사용자 B는 포도하고 수박을 좋아한다. 사용자 A하고 사용자 B는 동시에 포도하고 수박을 좋아해서 포도하고 수박의 유사도가 높다고 인식한다. 사용자 C는 수박을 좋아해서 사용자 C한테 포도를 추천할 수 있다.

모델 기반(Model-based) 협업 필터링은 데이터 값을 바로 사용하는 것 아니라 머신러닝을 통해 데이터를 분석하여 선호 예측 모델을 만든 뒤 이 모델을 기반으로 필터링을 하는 것이다²³⁾.

²²⁾ 김하연(2018). 사용자 기반 협업 필터링 알고리즘의 이웃 확장 연구, pp 5-8.

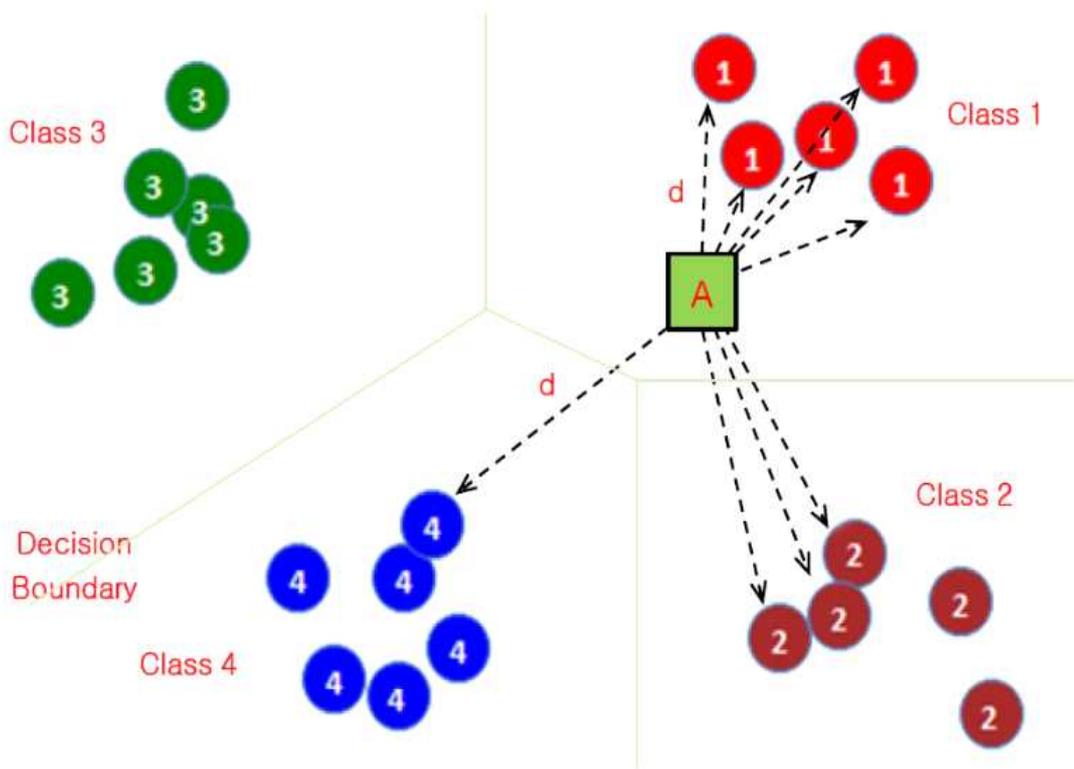
²³⁾ <https://blog.naver.com/rmby7484/221973700822>

3.2. 머신러닝 기법

3.2.1 근접 이웃(k-nearest neighbors) 기법

KNN (k-nearest neighbors) 알고리즘은 점 A로부터 가까운 거리에 있는 K개의 점을 선택한 후, K개의 점이 가장 많이 속한 클래스를 찾아 점 A가 그 클래스에 속한다고 판단하는 알고리즘이다.

[그림 3-5] 를 보면 A에서 가까운 9개의 점을 선택했으며 그중 5개는 클래스 1에 속하고, 3개는 클래스 2에 속하고, 1개는 클래스 4에 속한다면 점 A는 클래스 1에 속한다고 인식한다²⁴⁾.



[그림 3-5] KNN로 데이터 분류

²⁴⁾ [인사이트 캠퍼스] 지도학습의 예시 : KNN(K-Nearest Neighbors) 알고리즘 | 작성자 편인사이트

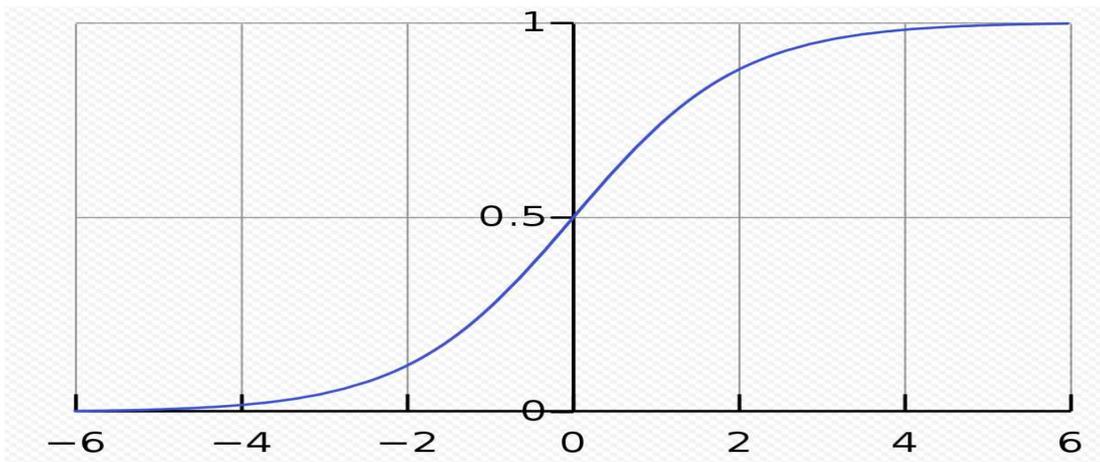
3.2.2 로지스틱 회귀(logistic regression)기법

로지스틱 회귀는 수많은 입력 변수들이 있을 때 이를 기반으로 주식을 살 것인가 팔 것인가 같은 의사결정을 할 때 주로 사용된다.

로지스틱 회귀의 목적은 일반적인 회귀 분석의 목표와 일치하게 종속 변수와 독립 변수 간의 관계를 함수로 나타내어 향후 예측 모델에 사용하는 것이다. 로지스틱 회귀의 Hypothesis는 시그모이드 함수를 이용한다. [그림 3-6] 을 살펴보면 로지스틱 함수는 x값 증가에 대해 처음에는 서서히 증가하다가 0.5 부근에 급격한 증가를 하고 1에 가까워질수록 서서히 증가하는 경향을 보인다. 이런 성질은 대부분의 특정 현상의 발생 확률을 예측하는 상황에서 매우 적합하다²⁵⁾.

시그모이드 함수
$$f(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{1 + e^x}$$
 이다

아래 이 함수의 그래프이다.



[그림 3-6] 시그모이드 함수의 그래프

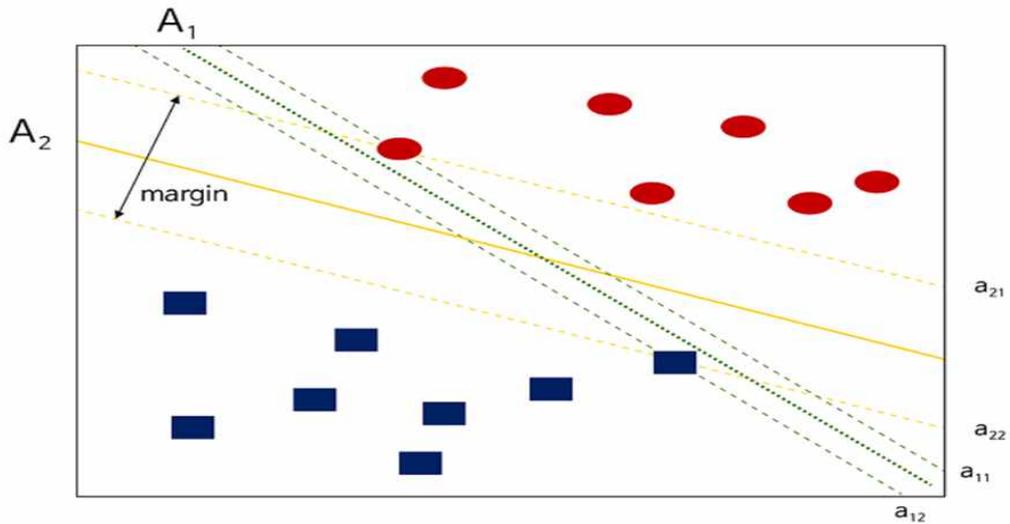
²⁵⁾ [위키백과] 로지스틱 회귀

3.2.3 서포트 벡터 머신(support vector machines)기법

서포트 벡터 머신은 주어진 샘플 그룹에 대해 그룹 분류(classification) 규칙을 찾아내는 기법의 하나다. 패턴인식(pattern recognition) 분야에서 제안된 알고리즘으로 1995년 러시아의 수학자 블라디미르바프닉(Valdimir N. Vapnik, 1936)이 처음 제시했다. 이후 글자 인식 분야에서 그 성능을 인정을 받아 기계학습 분야에서 분류를 위한 대표적인 알고리즘이 되었다. 각 샘플이 특성을 정의하는 N 개의 변수로 표현된다고 할 때, 샘플들은 N 차원 데이터 공간(data space)에 분포하게 된다. 서포트 벡터 머신의 목적은 N 차원의 데이터 공간에서 샘플 그룹들을 구분해 내는 최적의 분할 선(optimal decision boundary)을 찾아내는 것이다. 선형 판별 분석(linear discriminant analysis)은 선형 분할 선(linear decision boundary)을 찾아주지만, 서포트 벡터 머신은 커널 함수를 어떻게 정의하느냐에 따라 선형 또는 비선형 분할선(nonlinear decision boundary)을 가진다.

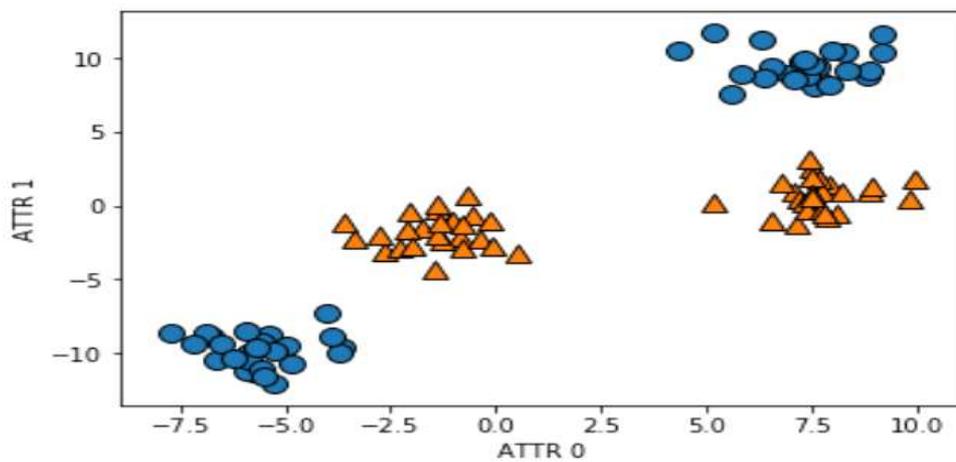
데이터 공간에서 샘플 그룹을 구분할 수 있는 초평면(hyperplane)의 개수는 무수히 많다. 이때 초평면에 의해 분리된 점 중 초평면과 가장 가까운 점과 초평면 사이의 거리를 마진(margin) 이라고 정의하며, 초평면마다 고유의 마진을 가지게 된다. 예를 들어 [그림 3-7] 과 같이 2개의 그룹으로 나누어져 있을 때, 초평면 A_1 과 A_2 모두 두 그룹을 구분하는 분할 선으로 사용이 가능하다. A_1 의 마진은 a_{11} 과 a_{12} 사이의 거리이며, A_2 의 마진은 a_{21} 과 a_{22} 사이의 거리가 된다. 서포트 벡터 머신이 지향하는 최적의 분할 선의 성질은 초평면들 중 최대의 마진을 가지는 초평면이며, 이러한 초평면을 최대-마진 초평면이라 한다²⁶⁾.

26) [네이버 지식백과] 서포트 벡터 머신 [support vector machine] (생화학 백과)

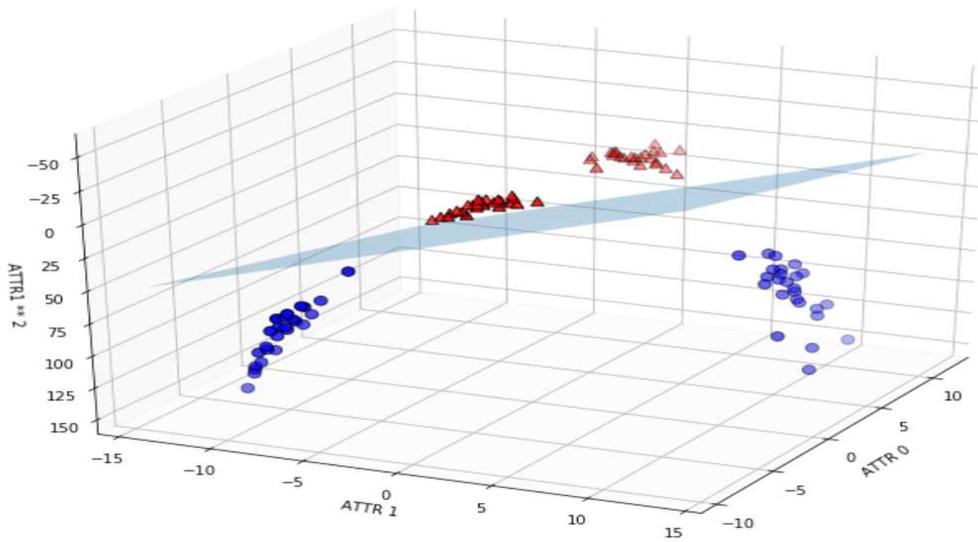


[그림 3-7] 데이터1(빨강)과 데이터 2(파랑)를 구분하는 초평면 A1와 A2
(출처: 생화학분자생물학회)

일반적으로 분류하고 싶은 데이터는 비선형일 경우가 많다. 따라서 SVM은 기본적으로 초평면을 사용하는 선형 분류 기법이지만 선형 분류가 되지 않는 경우, 이를 해결하기 위해 밑의 [그림 3-8] 저차원 데이터를 [그림 3-9] 고차원 데이터와 같이 커널 함수를 통해 저차원의 데이터를 고차원으로 대응하여 분류할 수 있게 한다. 이 작업이 바로 커널 트릭 (kernel trick) 기법이다.



[그림 3-8] 저차원 데이터



[그림 3-9] 고차원 전환된 데이터

서포트 벡터 머신(support vector machine, SVM)의 분류정확도가 SVC 커널에 따라 다르다. 아래 <표 3-1> SVC 커널 구분에 대한 내용이다.

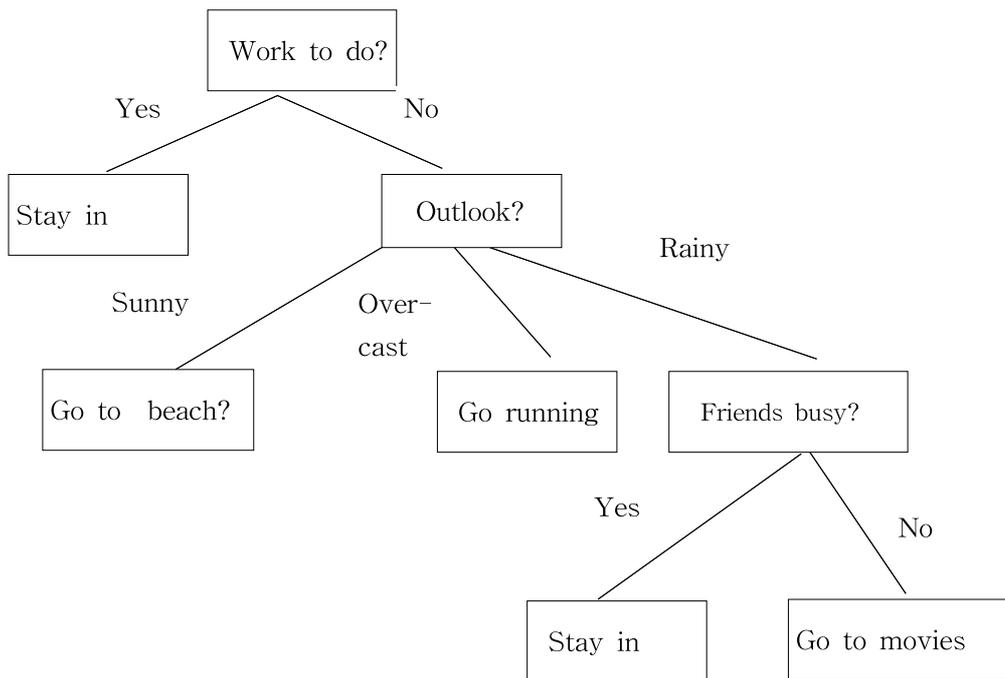
구분	적용 커널명	파라미터	설명
선형 커널	linear	C	소프트 마진 분류 (커널 트릭 X)
다항식 커널	poly	C degree coef0 gamma	지정한 차수까지 변환하여 확장 ex) 이미지 처리
RBF 커널	rbf	C gamma	지정한 차수까지 변환하여 확장 ex) 이미지 처리
시그모이드 커널	sigmoid	C coef0 (gamma)	지정한 차수까지 변환하여 확장 ex) 이미지 처리

<표 3-1> SVC 커널 (kernel trick) 트릭

(출처 <https://blog.naver.com/taewwon/221598401726> 작성자 공린이)

3.2.4 결정 트리(decision trees) 기법

결정 트리는 입력된 변수들의 값에 의하여 목표변수의 값을 예측하는 방법이다. 여러 단계의 조건을 갖는 문제와 그 조건과 그에 따른 해결방법을 트리 형태로 나타낸 것. 가장 큰 조건이 트리의 뿌리를 만들고, 세부 조건이 트리의 각 가지를 만들며, 해결 방안은 트리의 잎(leaf) 노드로 나타낸다²⁷⁾.



[그림 3-10] 결정 트리 예식

3.2.5 XGBoost (extreme gradient boosting) 기법

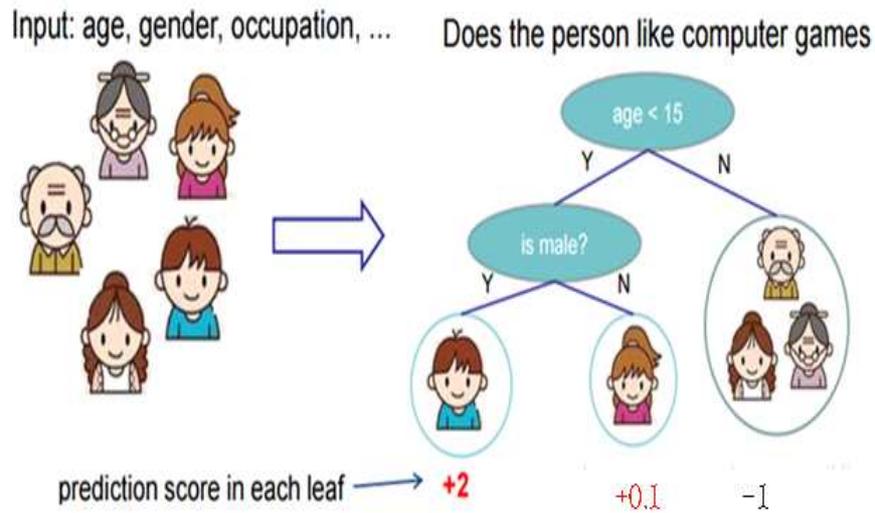
XGBoost는 선형 모델이나 트리 기반 모델에서의 과적합 문제를 해결하고, 규모가 큰 데이터 셋의 안정성과 훈련 속도를 향상시키기 위한 목적으로 Tianqi Chen과 Carlos Guestrin 이 소개한 방법이다²⁸⁾. 기본적으로 부스팅이라는 기술을 사용한다. 여러 개 약한 분류기를 묶어서 정확도를 예측하는 기법이다.

아래 [그림 3-11] 은 일반 트리 분석이다. 남자아이의 예측 점수는 2로 나타나

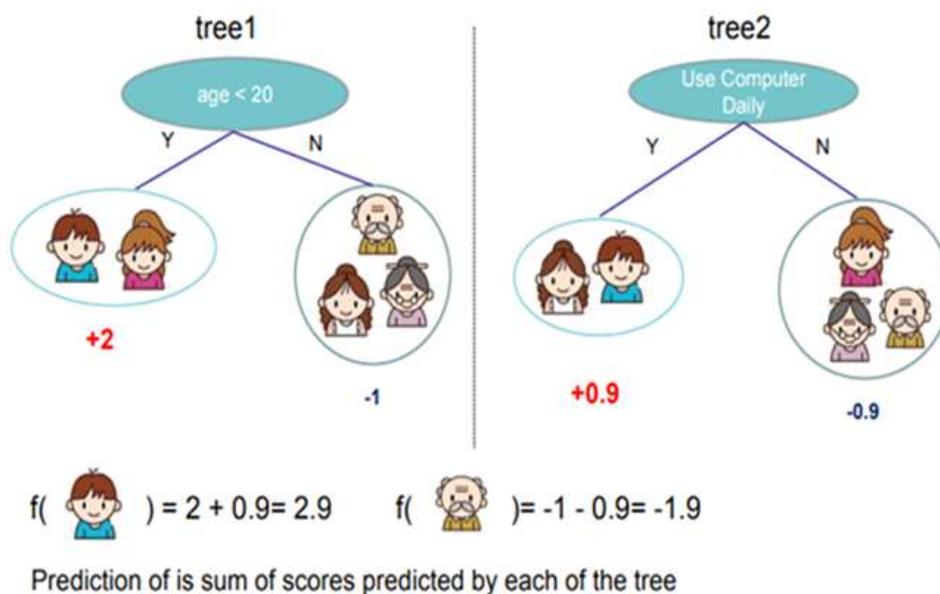
27) [네이버 지식백과] 결정 트리

28) 하지은(2017). RandomForest와 XGBoost를 활용한 민원 카테고리 및 담당 부서 자동분류 성능 비교 p5

며 할아버지 예측 점수는 -1로 나타난다. [그림 3-12] 의 XGBoost 분류기로 통한 데이터 분석은 남자아이가 예측 점수는 2.9로 나타났으며 할아버지 예측 점수는 -1.9로 나타나서 더 쉽게 분류할 수 있다.



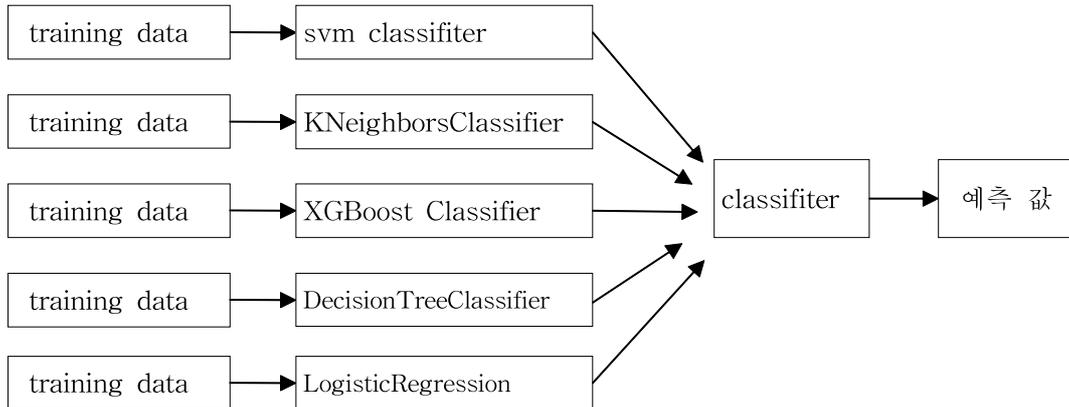
[그림 3-11] 일반 트리 분석



[그림 3-12] XGBoost 분류기로 데이터 분석

3.2.6 앙상블 Stacking 기법

앙상블 Stacking 기법은 간단하게 설명하면 학습된 결과 데이터를 다시 한번 학습해서 예측 결과를 보여주는 방식이다. Stacking 기법은 [그림 3-13] 과 같이 진행한다.



[그림 3-13] Stacking 기법 흐름도

3.3 머신러닝 모델 성능 평가 방법

머신러닝 모델 평가 방법은 아래 <표3-2>와 같이 혼동행렬(confusion matrix)과 기댓값(expected value)을 이용해서 모델을 평가한다.

	Positive 양성	Negative 음성
Positive 예	TP(true positive 진양성)	FP(false positive 위양성)
Negative 아니요	FN(false negative 위음성)	TN(true negative 진음성)

<표 3-2> 머신러닝 평가 혼동행렬

$$\text{Precision(정밀도)} = \frac{TP}{(TP+FP)}$$

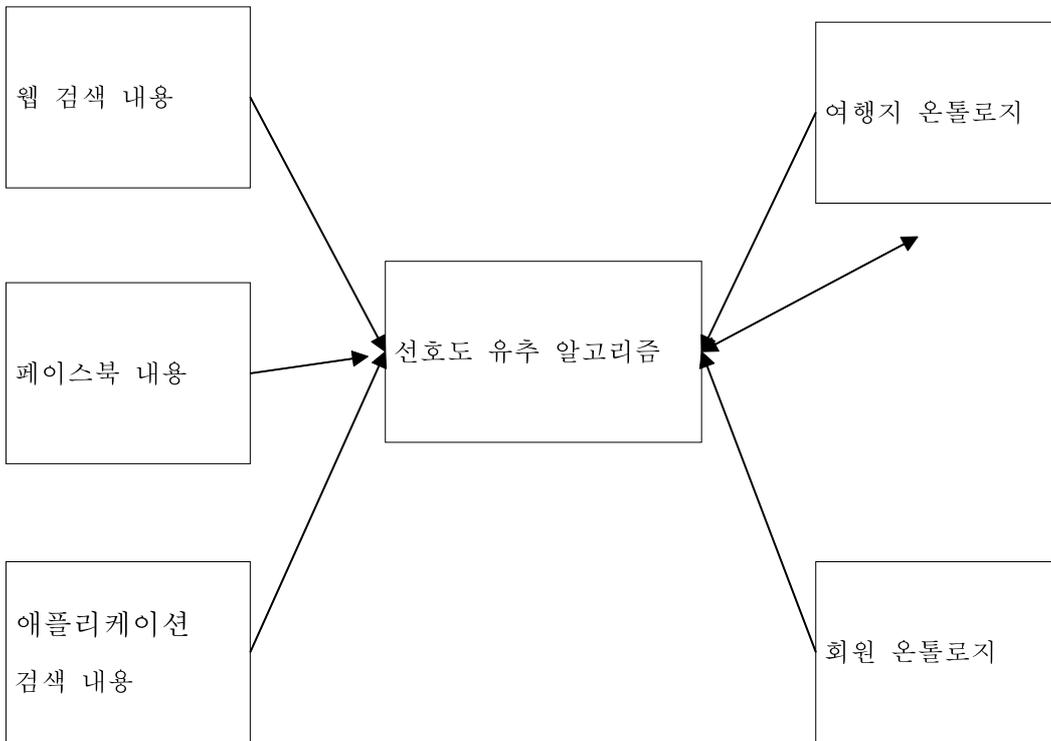
$$\text{Recall} = \frac{TP}{(TP+FN)}$$

$$\text{Accuracy(정확도)} = \frac{(TP+TN)}{(TP+FP+FN+TN)}^{29)}$$

3.4 관광지 추천 시스템 선행 연구

3.4.1 개인화 관광지 추천 시스템

(2015. 박연진, 송경아, 황재원, 창병모) 온톨로지 기반의 개인화된 여행 추천 시스템의 구현에서 사용자의 웹 검색 내용, 페이스북 검색 내용, 애플리케이션 검색 내용을 수집한 후에 선호도 유추 알고리즘으로 분석해서 관광지를 추천한다. 아래 [그림 3-14] 와 같이 진행한다³⁰⁾.



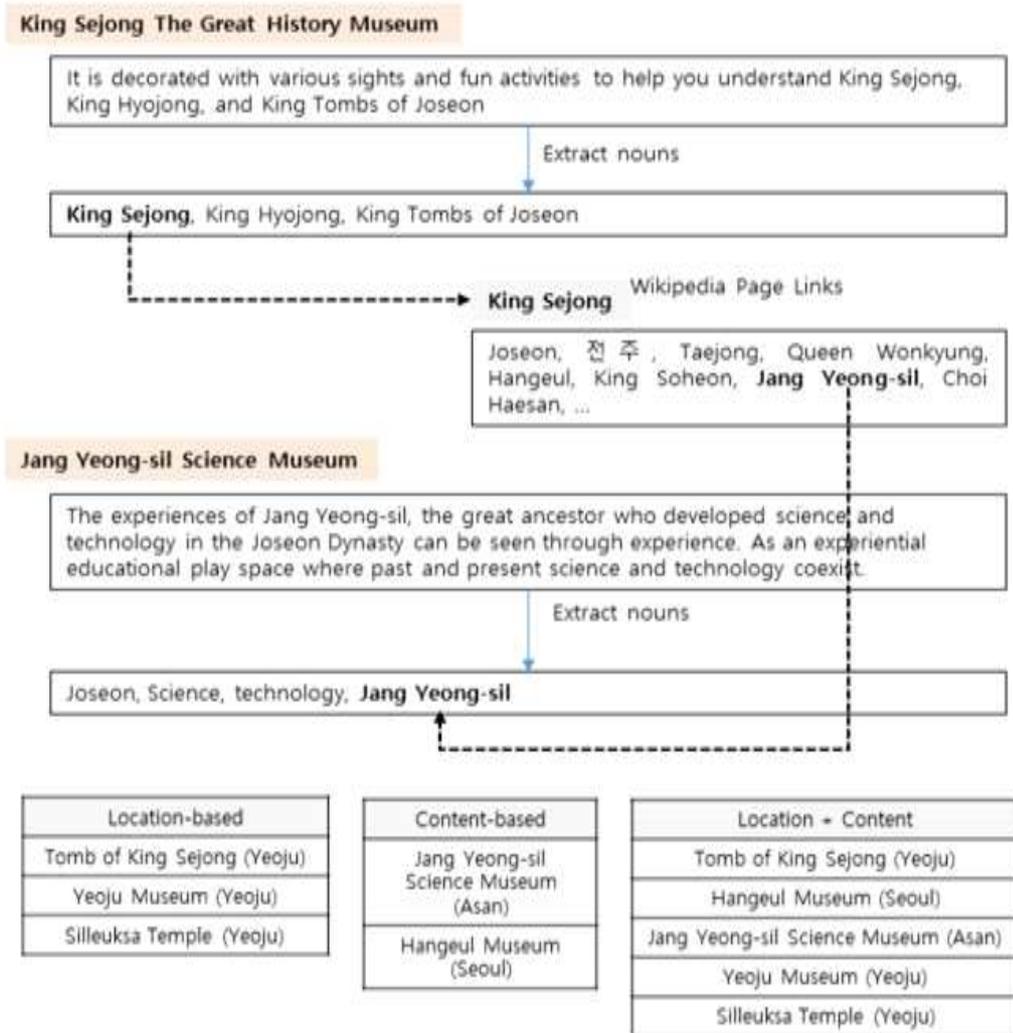
[그림 3-14] 개인화 추천 시스템

29) 신재욱(2019). IoT센서를 활용한 머신러닝(SVM)기반 실시간 운동자세 측정시스템,pp 23-24.

30) (2015. 박연진, 송경아, 황재원, 창병모) 온톨로지 기반의 개인화된 여행 추천 시스템

3.4.2 빅데이터를 이용한 관광지 추천 시스템

안진현 임동혁 (2020) 이벤트와 관련된 주변 관광지 자동 추천 알고리즘 개발 연구에서는 한국관광공사 LOD(Linked Open Data), 위키피디아, 국어사전 등에 사용하여 위치와 내용을 모두 고려해서 관광지의 설명내용의 명사들을 검색한 뒤 다른 관광지의 명사들과 비교를 하여 일치한 명사가 많이 있을수록 내용이 관련이 있다고 판단한 후에 각 관광지의 위도와 경도로 거리를 계산한 뒤 사용자한테 관광지를 추천한다³¹⁾.

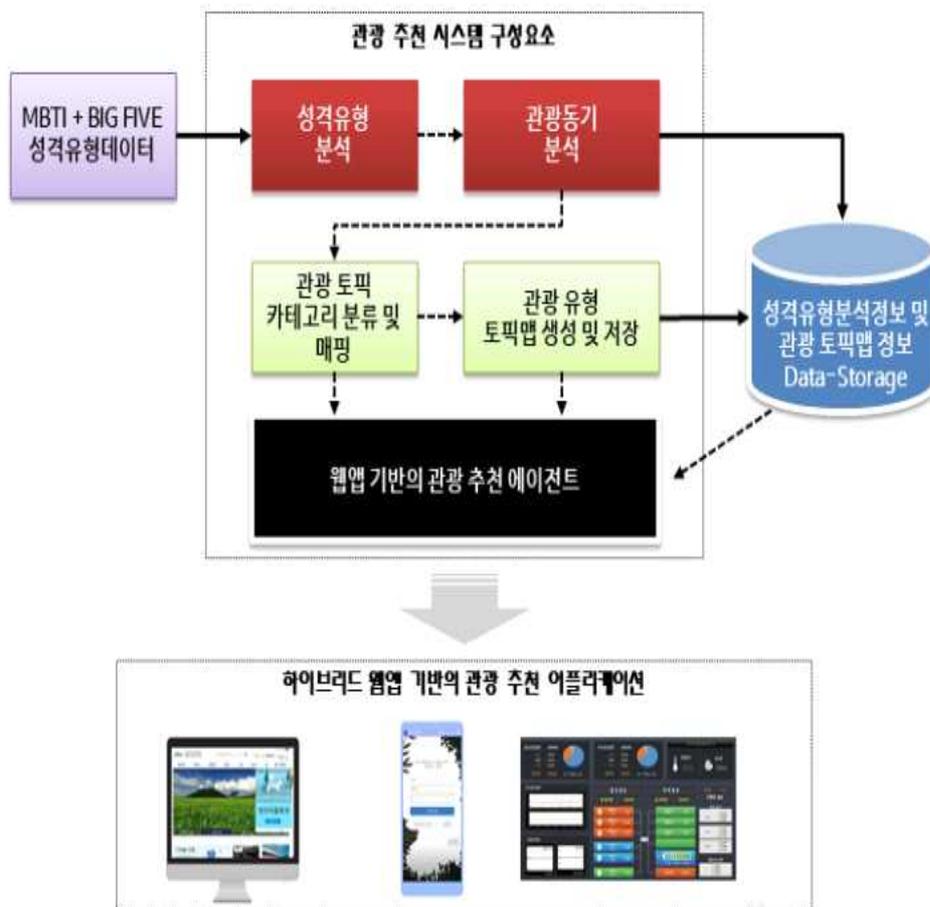


[그림 3-15] 빅데이터 기반 추천

31) 안진현 임동혁 (2020). 이벤트와 관련된 주변 관광지 자동 추천 알고리즘 개발 연구.

3.4.3 성격유형에 따른 관광지 추천에 관한 연구

류기환, 윤수현, 문석재(2020) MBTI와 Big Five 모델에 근거한 성격유형에 따른 토픽맵을 이용한 관광지 추천에 관한 연구에서 MBIT + Big Five을 활용한 하이브리드 웹앱 기반의 관광 추천 시스템을 제안한다. 이는 성격유형 데이터를 수집하여 성격유형과 관광 동기를 분석한 후, 토픽맵 기술을 적용하여 관광 토픽 카테고리를 분류 매핑하여 성격유형 결과에 연관되는 관광 유형을 추천한다.



[그림 3-16] Construction of a tourism recommendation system using a hybrid web app-based topic map³²⁾

³²⁾ 류기환, 윤수현, 문석재(2020) MBTI와 Big Five 모델에 근거한 성격유형에 따른 토픽맵을 이용한 관광지 추천에 관한 연구

3.5 관광지 추천 시스템 문제점

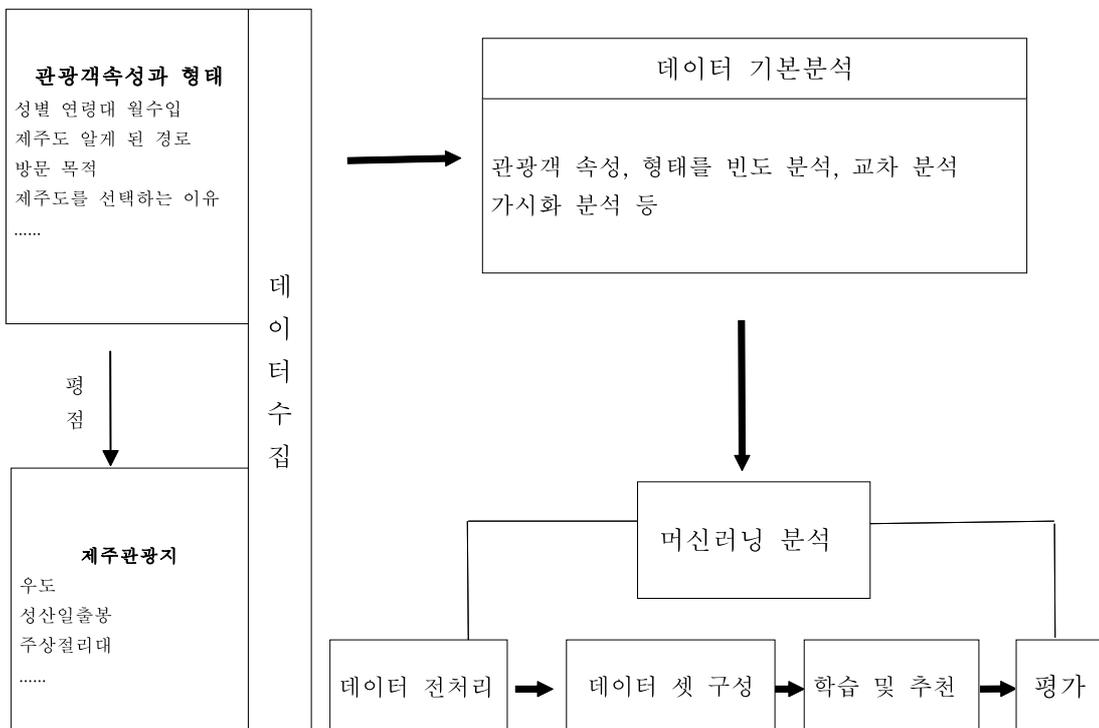
- 1) 위치 기반 추천은 가까운 관광지 순서로 추천함으로써 선호하지 않는 관광지를 추천할 수 있는 문제점이 있다.
- 2) 내용 기반 추천은 내용 위주로 추천하므로 거리를 고려하지 않고 관광지를 추천하는 문제점이 있다.
- 3) 개인화 기반 및 성격유형 기반 추천은 관광객의 취향과 여행지 간의 유사점을 추천함으로써 많은 데이터가 필요하고, 특정 관광객 개인의 취향만을 반영하기 때문에 동행하는 사람들에게는 적합하지 않을 수도 있다. 또한, 관광객의 관광 패턴 변화에 따라 새로운 데이터를 조사할 때마다 패턴 규칙 또한 변경이 불가피해서 이로 인한 시스템의 확장성 및 유연성 문제가 발생할 수 있다.
- 4) 빅데이터 기반 추천은 다중의 관광객이 선호하는 여행지를 우선으로 추천하게 되어 각 관광객의 특성에 맞는 관광지 추천이 제한되는 문제점이 있다.

위에 제시된 기존 관광지 추천 시스템의 문제점들을 종합해볼 때 개별관광객에게 알맞은 취향과 편의를 모두 고려한 추천은 어려울 것이기에 이를 보완해 줄 수 있는 머신러닝 시스템을 제안하다. 또한 머신러닝은 새로운 데이터가 입력되었을 때 별다른 규칙 변경이나 프로그래밍 없이 최선의 결과물을 도출해낼 수 있는 장점이 있다.

IV. 제안하는 방법

본 연구에서는 이미 제주를 방문했던 관광객의 속성과 형태, 그리고 관광지에 대한 만족도 평가를 토대로 머신러닝 알고리즘으로 선호 예측 모델을 만들고 이를 통해 머신러닝 기반 관광지 추천 시스템을 제안한다.

아래 [그림 4-1] 처럼 이미 제주도에 방문했던 중국 개별관광객의 속성과 형태를 기재해서 제주 관광지를 평점을 하는 방법으로 데이터를 수집한다. 수집한 데이터를 빈도 분석, 교차분석, 가시화 분석 등 데이터 기본분석을 한 후 머신러닝 분석에 필요한 데이터 전처리, 데이터 셋 구성, 학습 및 추천, 평가를 진행한 관광지 추천 방법이다.



[그림 4-1] 관광지 추천 방법

4.1 데이터 수집

4.1.1 설문지 구성

본 연구에 사용하는 데이터는 설문지 기법으로 수집했다. 이미 제주에 방문했던 관광객데이터를 수집하여 분석하고 관광지 선호 예측 모델을 만들고 사전에 정보를 얻고자 하는 다른 관광객에게 관광 도움을 주고자 한다. 또한 제주도 관광 이미지를 향상시키는 데 있어서 도움이 될 전망을 기대하며 설문지를 구성했다.

방문 관광객 속성 3항목 질문, 방문 형태 9항목 질문, 관광지 6항목 질문 및 기타 항목 질문 5항목으로 총 23개 항목 설문지로 구성했다. 설문지 구성은 <표 4-1> 와 같다.

구성 별	구성내용	질문 수
방문 관광객 속성	성별, 연령대, 월수입	3항목
방문 관광객 형태	제주도를 알게 된 경로, 다른 나라 방문 여부, 방문 목적, 동행자, 제주도 선택이유, 개별 관광 선택이유, 스케줄 준비 여부, 교통수단, 방문 횟수.	9항목
관광지	자연경관, 체험, 박물관, 테마파크, 공연, 해변	6항목
기타질문	며칠 일정, 만족도, 재방문, 중국어 정보제공 만족도, 개선해야 할 점	5항목

<표 4-1> 설문지 구성표

4.1.2 설문조사 실시 내용

설문지조사 대상은 이전에 제주도에 방문했던 중국 개별관광객, 조사 방법은 온라인 설문 도구를 이용해서 조사했으며 조사 범위는 2019년 이후에-2020년 2월 기간에 한 여행사 관광서비스를 이용한 중국 개별관광객이며 조사 기간은

2020년 8월 27일- 9월 10일 총 15일로 소유되었으며 총 652부를 배부하여 632부 회수되었으며 회수된 설문지 중 내용이 충실하지 않은 10부를 제외한 622부를 데이터 분석에 사용했다. 설문조사 실시 내용 아래 <표 4-2> 와 같다.

대상	제주도 방문했던 중국 개별관광객
조사방법	온라인 설문 도구 이용한 자기 기입식
조사기간	2020년 8월 27일~ 9월 10일
조사범위	2019년 이후에-2020년 2월 이전에 한 여행사 관광서비스를 이용한 중국 관광객
배부 수	652부

<표 4-2> 설문조사 실시 내용

4.2 기본 데이터 분석

4.2.1 관광객 속성 (인구 통계학)

관광객의 인구통계학적 특성을 파악하기 위해 빈도 분석(frequency analysis)을 사용한 결과는 다음<표 4-3>과 같다.

- (1) 성별에서는 남성 84명(13.50%), 여성 538명(86.50%)으로 여성이 남성보다 많은 것으로 나타났다.
- (2) 연령대에서는 10대 14명(2.25%), 20대 364명(58.52%), 30대 130명(20.90%), 40대 71명(11.41%), 50대 이상 43명(6.91%)으로 20대에서 가장 많은 비율을 보였다.
- (3) 월수입에서는 3,000위안 이하 20명(3.21%), 3,001-5,000위안 100명(16.08%), 5,001-8,000위안 211명(33.92%), 8,001-12,000위안 98명(15.76%), 12,000위안 이상 193명(31.03%)으로 나타났다.

구분	내용	빈도	퍼센트
성별	남	84	13.50
	여	538	86.50
연령	10대	14	2.25
	20대	364	58.52
	30대	130	20.90
	40대	71	11.41
	50대 이상	43	6.91
월 평균 소득	3000위안 이하	20	3.21
	3001~5000위안	100	16.08
	5001~8000위안	211	33.92
	8001~12000위안	98	15.76
	12000위안 이상	193	31.03

〈표 4-3〉 관광객 속성 통계분석(인구통계학)

4.2.2 관광객 형태 분석

관광객 형태를 파악하기 위해 빈도 분석(frequency analysis)을 사용한 결과는 다음<표 4-4>과 같다.

- (1) 제주도를 알게 된 경로는 TV 131명(21.06%), 친구 102명(16.40%), 여행사 18명(2.89%), 인터넷, 잡지 197명(31.67%), 기타 174명(27.97%)으로 나타났다.
- (2) 다른 나라 방문 여부는 있다. 428명(68.81%), 없다 194명(31.19%)으로 나타났다.
- (3) 제주도를 선택하게 된 이유는 중국에서 가깝다 321명(51.61%), 제주도 경치가 좋다 226명(36.33%), 한국문화 37명(5.95%), 친구추천 38명(6.11%)으로 나타났다.
- (4) 제주도 방문 목적은 관광 479명(77.01%), 쇼핑 56명(9.00%), 음식 70명

- (11.25%), 의료 9명(1.45%), 기타 8명(1.29%)으로 나타났다.
- (5) 여행 동반자는 혼자 22명(3.54%), 가정 237명(38.10%), 친구 354명(56.91%), 동료 9명(1.45%)으로 나타났다.
- (6) 개별 관광 선택된 이유는 모르는 사람이 같이 관광하기 싫다 10명(1.61%), 자유로움 610명 (98.07%), 기타 2명 (0.32%)으로 나타났다.
- (7) 관광 일정을 계획하였는지는 쫓다 571명(91.80%), 짜지 않았다 51명(8.20%)으로 나타났다.
- (8) 주로 교통수단 뭘 이용했는지는 가이드 전용차 231명 (37.14%), 택시 68명 (10.93%), 관광버스 250명(40.19%), 일반 버스 73명(11.74%)으로 나타났다.
- (9) 제주도 방문 횟수는 1번 483명(77.65%), 2번 58명(9.32%), 3번 65명(10.45%), 4번 4명(0.64%), 5번 이상 12명(1.93%)으로 나타났다.

구분	내용	빈도	퍼센트
제주도를 알게 된 경로	TV	131	21.06
	친구 알림	102	16.40
	여행사	18	2.89
	인터넷, 잡지	197	31.67
	기타	174	27.97
다른 나라 방문 여부	있다.	428	68.81
	없다	194	31.19
제주도를 선택하는 이유	중국에 가깝다.	321	51.61
	제주도 경치가 좋다.	226	36.33
	한국 문화	37	5.95
	친구 추천	38	6.11
제주도 방문 목적	관광	479	77.01

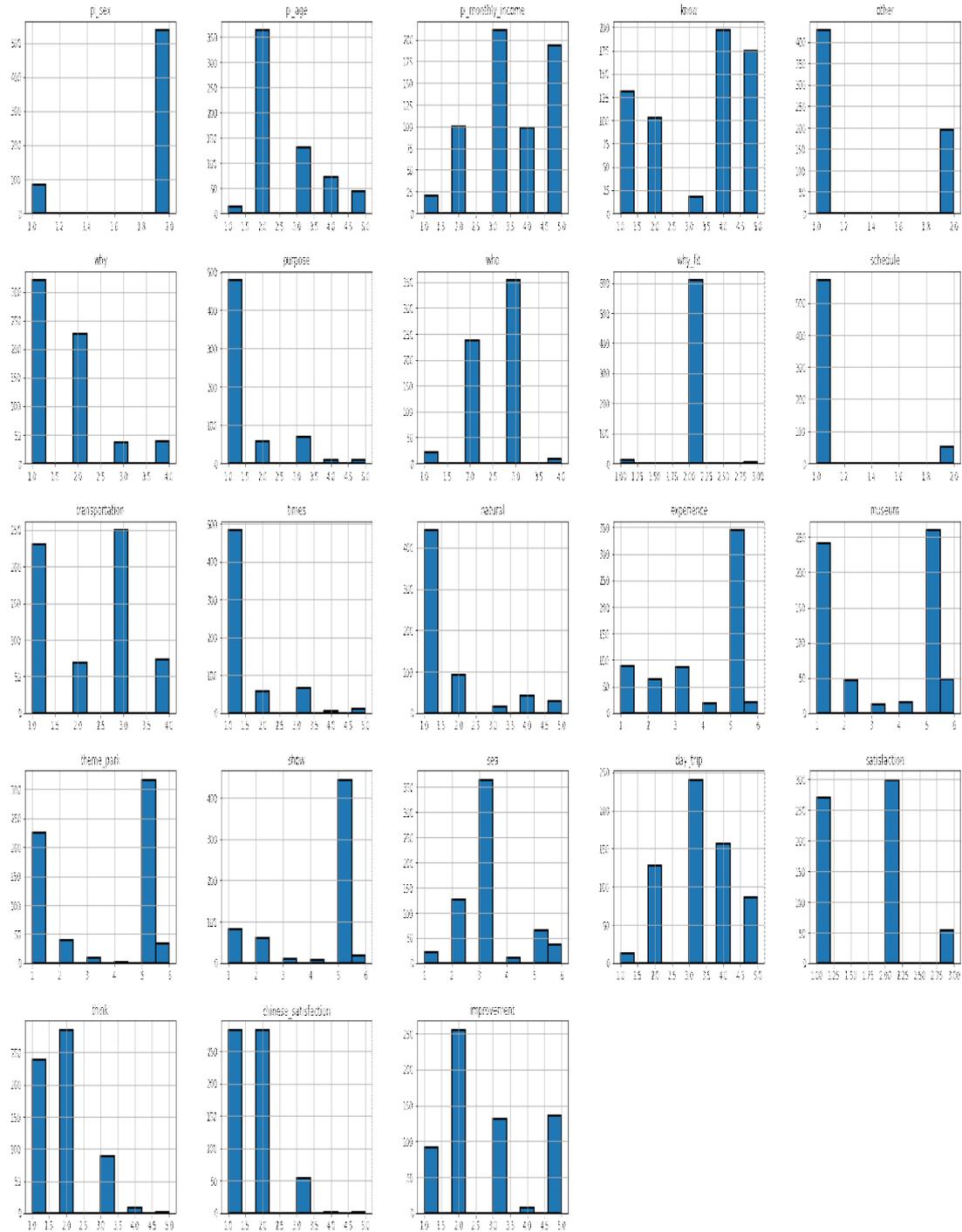
	쇼핑	56	9.00
	음식	70	11.25
	의료	9	1.45
	기타	8	1.29
여행 동반자	혼자	22	3.54
	가정	237	38.10
	친구	354	56.91
	동료	9	1.45
개별 관광 선택된 이유	모르는 사람이 같이 관 광하기 싫다	10	1.61
	자유로움	610	98.07
	기타	2	0.32
관광 일정을 계획하였는지	짚다	571	91.80
	짜지 않았다	51	8.20
주로 교통수단 별 이용했는지	가이드 전용차	231	37.14
	택시	68	10.93
	관광버스	250	40.19
	일반 버스	73	11.74
제주도 방문 횟수	1번	483	77.65
	2번	58	9.32
	3번	65	10.45
	4번	4	0.64
	5번	12	1.93

〈표 4-4〉 관광객 형태 통계분석

4.2.3 데이터 가시화 및 분석

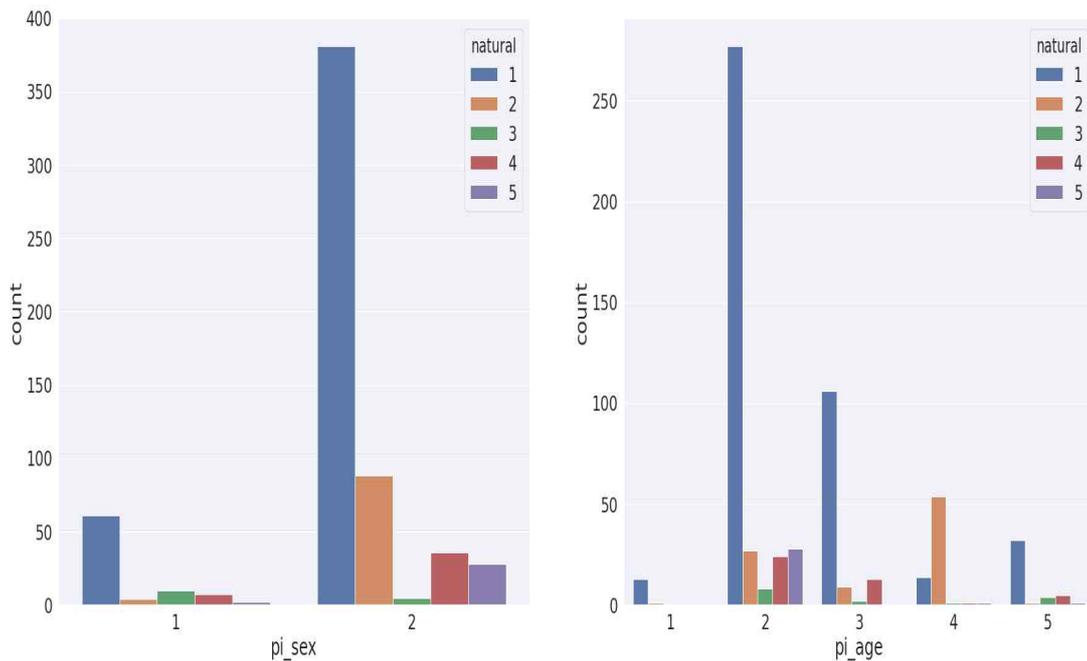
아래 [그림 4-2] 는 파이썬 matplotlib.pyplot 라이브러리를 이용한 관광객 설문조사 질문 답변 빈도 그림을 나타냈다. 질문 답변을 한 연령대를 살펴보면 비율이 20대에서 가장 높게 나타나 주로 젊은 층이 관광 많이 하는 것을 알 수 있다. 또한 개별 관광을 선호하는 이유는 자유로움 때문인 것이 100% 가까이 되어 관광객들은 관광의 자유로움

을 추구하는 것으로 알 수 있다. 다음으로 90% 넘는 비율로 대부분 개별관광객은 관광하기 전에 스케줄을 계획하는 특성 있다.



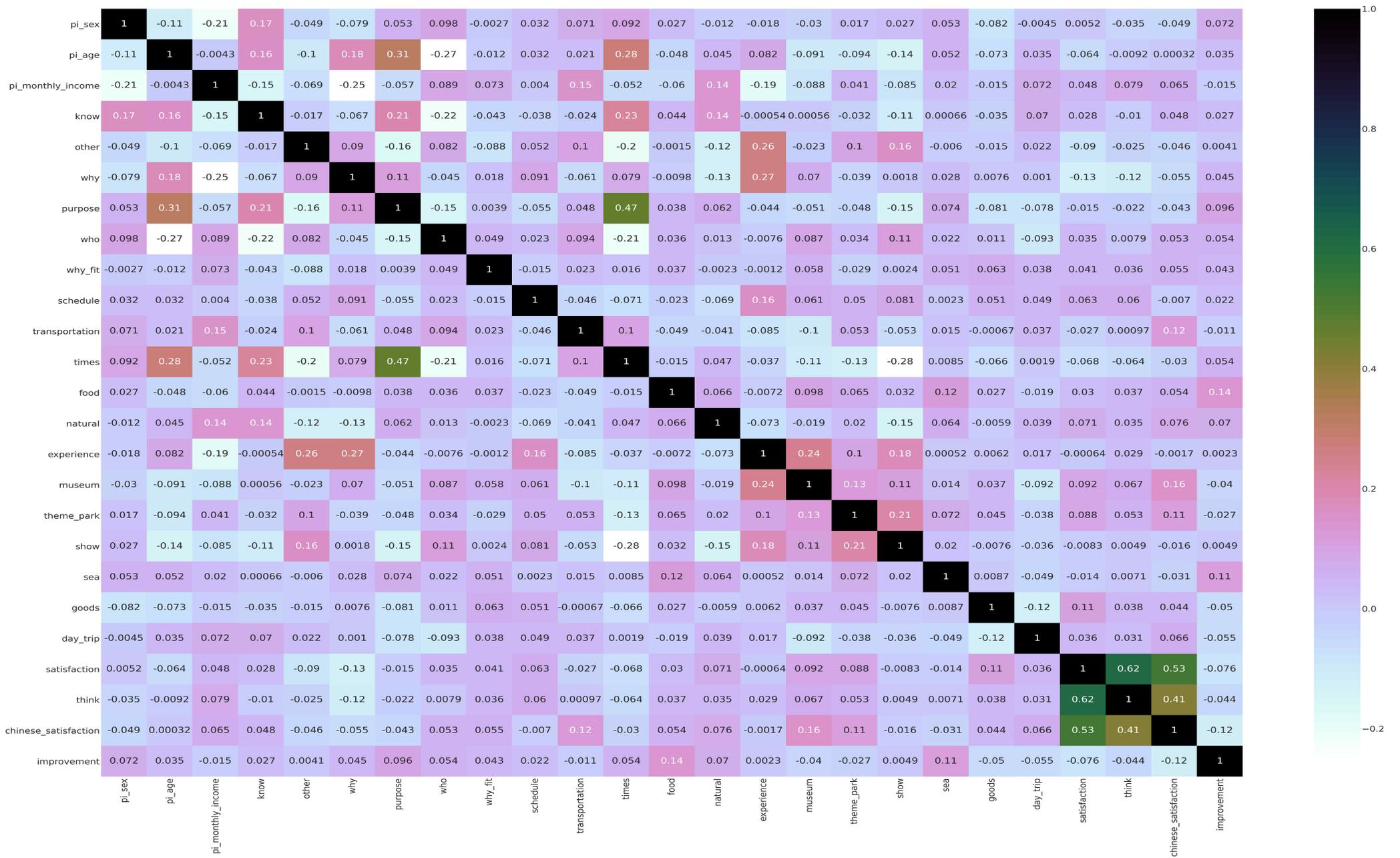
[그림 4-2] 질문 답변 빈도도

아래 [그림 4-3] 은 파이썬 matplotlib.pyplot 라이브러리를 이용한 관광객 설문 조사 성별별, 나이별 자연 관광지 선택 상황을 나타냈다. 성별별 자연 관광지 선택에서 남성과 여성 대부분이 우도를 선호하는 것을 알 수 있다. 다음으로 나이별 자연 관광지 선택에서 10대, 20대, 30대, 50대는 우도를 선호하는 것을 알 수 있고 40대는 성산일출봉을 선호하는 것을 알 수 있다.



[그림 4-3] 성별별, 나이별 자연 관광지 선택 빈도

아래 [그림 4-4] 는 파이썬 seaborn 라이브러리를 이용한 관광객 설문 조사 데이터 항목 간의 상관관계를 표현하다. 재방문이 만족도랑 긴밀한 관계를 확인할 수 있다. 예를 들어 관광에 대한 만족도가 높았다면 재방문 의사가 높기 때문에 두 피쳐 간의 상관관계 계수가 높게 나타난다. 상관관계 계수가 높을수록 두 피쳐의 비례가 결과에 영향을 미치는 것을 확인할 수 있었다.



[그림 4-4] 각 항목 간 상관관계

4.3 머신러닝을 이용한 관광지 추천

4.3.1 구현 환경

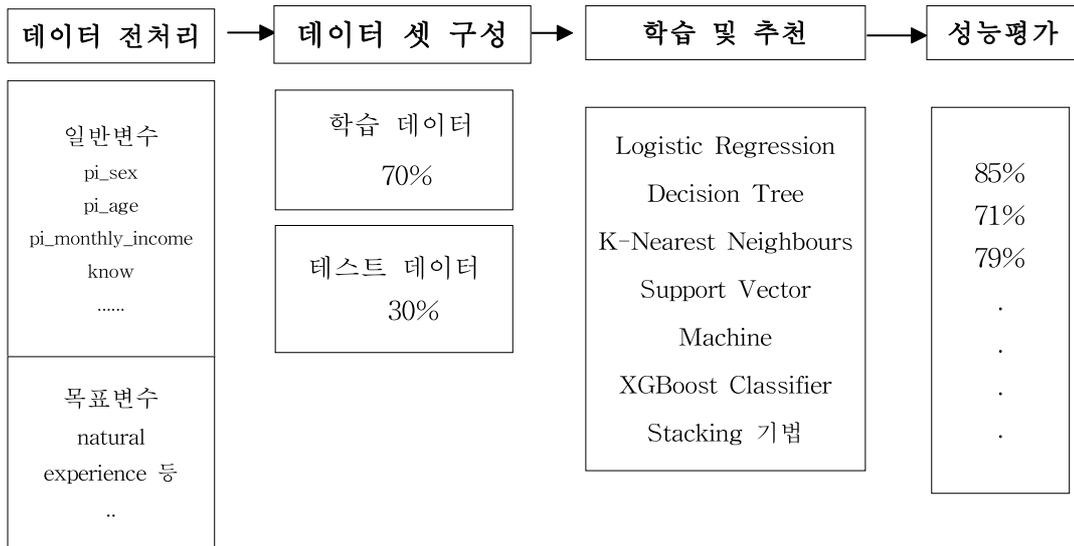
아래 <표 4-5>와 같이 개발 환경은 Kaggle 온라인 분석환경을 이용했다. Settings 언어는 Python을 사용했다. 응용 라이브러리에서는 파일을 읽기 위해서 pandas 라이브러리를 이용했고 XGBoost 알고리즘으로 데이터 학습을 하기 위해 xgboost 라이브러리를 사용했다. 또한 Logistic Regression, Decision Tree, K-Nearest Neighbours, Support Vector Machine (SVM) 등 알고리즘으로 데이터 분석을 하기 위해 sklearn 라이브러리를 사용했다.

항목	내용	설명
개발 플랫폼	Kaggle	온라인 개발 환경
사용언어	Python	머신러닝 개발 언어
응용 라이브러리	pandas	파일 읽기
	xgboost	XGBoost 알고리즘로 분석
	sklearn	Logistic Regression, Decision Tree, K-Nearest Neighbours, Support Vector Machine (SVM)
응용 데이터 셋	설문지 622부 자료를 정리된 csv 파일	이전에 한 여행사의 관광서비스를 이용한 중국 개별관광객

<표 4-5> 구현 환경

4.3.2 머신러닝을 이용한 처리 과정

본 논문에서 머신러닝 분석하기 위해서 데이터 전처리, 데이터 셋 구성, 학습 및 추천, 성능평가 및 추천하고 성능평가를 통해 총 네 가지 흐름으로 진행한다. 분석 흐름이 [그림 4-5] 와 같다.



[그림 4-5] 머신러닝을 이용한 처리 과정

4.3.3 데이터 속성

설문조사 데이터 중에 관광객 속성, 행태는 영어 약어로 표기했다.

한국어	영어 약어	한국어	영어 약어
일반변수			
성별	pi_sex	나이	pi_age
월수입	pi_monthly_income	제주도 알게 된 경로	know
다른 나라 방문 여부	other	제주도를 선택한 이유	why
방문 목적	purpose	여행 동반자	who
개별관광을 선택된 이유	why_fit	관광 일정을 계획하였는지	schedule
주로 교통수단 뭘 이용했는지	transportation	제주도 방문 횟수	times
목표변수			
좋아했던 자연 관광지 뭐였는지	natural	좋아했던 체험 뭐였는지	experience

<표 4-6> 관광객 속성, 행태 영어 약어로 표기

4.3.4 학습 데이터 셋 구성

머신러닝 분석을 진행하기 위해 데이터가 학습 데이터하고 테스트 데이터로 나눈다. 아래 <표 4-7>처럼 본 논문에서 학습데이터는 총 데이터의 70% 지정했고 테스트 데이터는 총 데이터의 30%로 지정했다. 622개 데이터 중에 435개 학습데이터로 187개 테스트 데이터로 나뉘었다.

전체 데이터 셋 개수	622개	100%
학습 데이터 개수	435개	70%
테스트 데이터 개수	187개	30%

<표 4-7> 데이터 셋 구성

일반 변수들하고 목표 변수 natural로 학습데이터 및 테스트 데이터를 세팅한 코드 아래 [그림 4-6] 과 같다.

```
train_X = train[['pi_sex', 'pi_age', 'pi_monthly_income', 'know', 'other', 'why', 'purpose', 'who', 'schedule', 'times']];
train_y=train.natural # output of our training data
test_X= test[['pi_sex', 'pi_age', 'pi_monthly_income', 'know', 'other', 'why', 'purpose', 'who', 'schedule', 'times']] # t
test_y =test.natural #output value of test data
```

[그림 4-6] 학습 데이터 테스트 데이터 세팅 코드

4.3.5 머신러닝 알고리즘에 따른 실험 결과

최적 관광지 선호 예측 모델을 찾기 위해 여러 가지의 알고리즘 모델로 같은 데이터를 분석한 다음, 랜덤으로 10번의 데이터를 추출하여 똑같은 조작을 하고 매번 분석의 정확도를 기록하여 각 알고리즘 모델의 평균 정확도를 비교해서 평가한다.

4.3.5.1 Logistic Regression

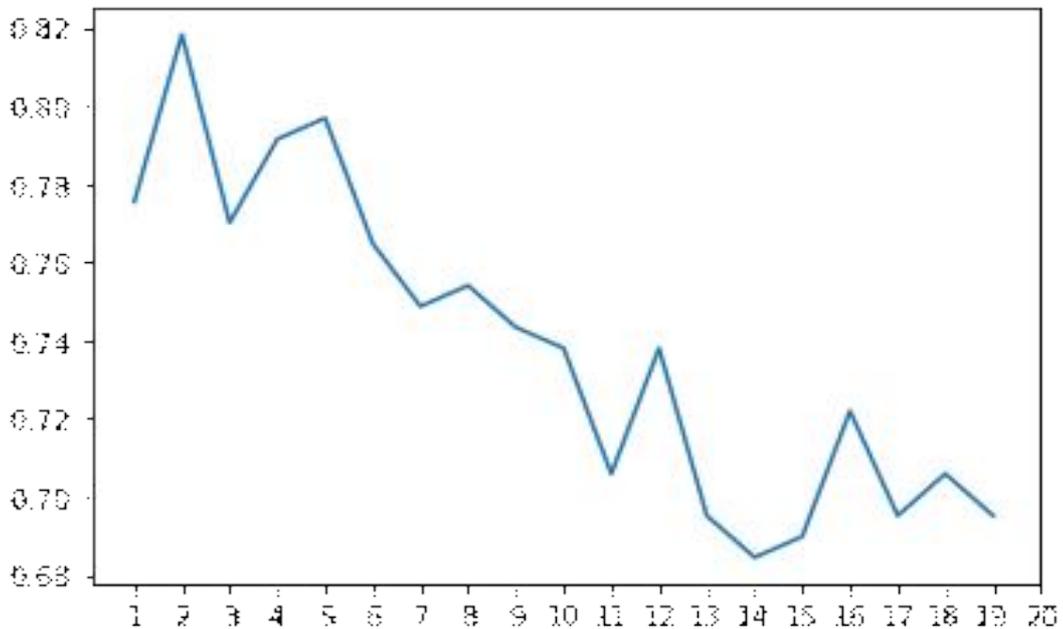
10번 랜덤 데이터를 Logistic Regression 모델로 통한 평균 정확도는 81.05%로 나타났다.

1	2	3	4	5	6	7	8	9	10	평균 정확도
79.1	80.2	77.0	80.2	80.7	79.7	83.4	81.8	83.4	85.0	81.05

<표 4-8> Logistic Regression 모델 추천 정확도

4.3.5.2 Decision Tree

Decision Tree 모델은 여러 파라미터 중에서 본 논문 데이터가 max_depth에 따라 정확도 결과가 크게 달라서 최적 max_depth 값을 구하는 것부터 진행했다. 아래 [그림 4-7] 을 보면 max_depth 2일 때 최적 정확도가 제일 높은 것으로 나타났다. 매번 랜덤 데이터를 바꾸면 최적 정확도의 max_depth도 변경된다.



[그림 4-7] Decision Tree 모델 분석 max_depth도

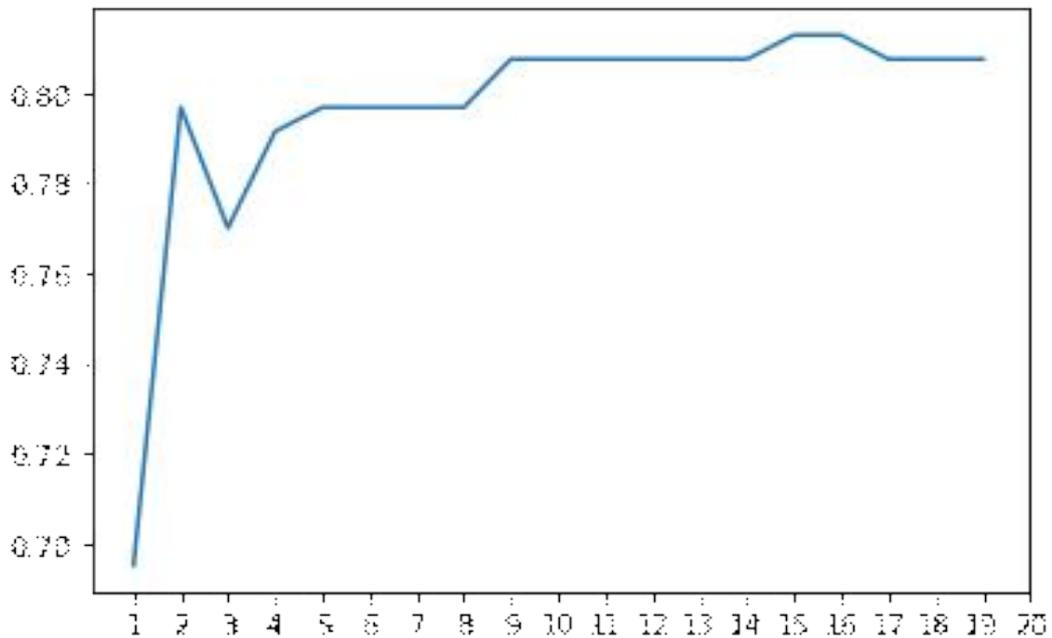
아래<표 4-9>는 위에 같은 10번 랜덤 데이터를 최적 max_depth를 구한 후 Decision Tree 모델 통한 평균 정확도는 82.25%로 나타났다.

횟수	1	2	3	4	5	6	7	8	9	10
max_depth	3	4	4	4	5	5	5	4	4	4
정확도	80.2	80.7	80.2	84.0	79.7	80.2	84.0	83.4	84.5	85.6
평균 정확도	82.25									

<표 4-9> Decision Tree 모델 추천 정확도

4.3.5.3 K-Nearest Neighbours

K-Nearest Neighbours로 분석하면 파라미터 n_neighbors에 따라 결과가 많이 달라져서 최적 n_neighbors부터 구해야 한다. [그림4-8]을 보면 n_neighbors는 16일 때 최적 정확도가 제일 높은 것으로 나타났다. 매번 랜덤 데이터를 바꾸면 최적 정확도의 n_neighbors도 변경된다.



[그림 4-8] K-Nearest Neighbours 모델 분석 n_neighbors도

아래<표 4-10>는 위에 같은 10번의 랜덤 데이터를 최적 n_neighbors를 구한 후 K-Nearest Neighbours 모델 통한 평균 정확도는 83.36%로 나타났다.

횟수	1	2	3	4	5	6	7	8	9	10
n_neighbors	9	4	2	2	5	16	4	6	4	12
정확도	81.8	80.7	80.7	85.6	83.4	81.8	84.5	82.9	85.6	86.6
평균 정확도	83.36									

<표 4-10> K-Nearest Neighbours 모델 추천 정확도

4.3.5.4 Support Vector Machine (SVM)

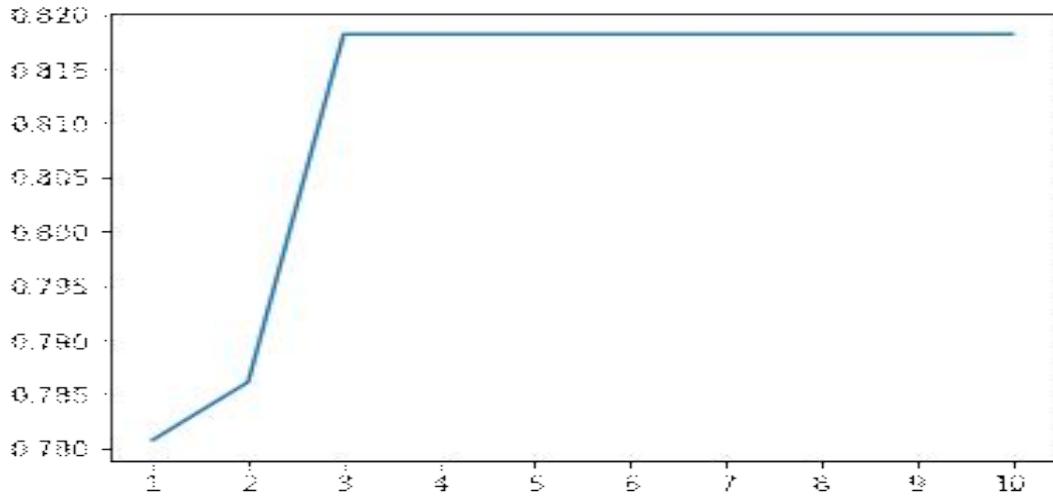
SVM 모델을 통한 분석은 파라미터에 커널(kernel)에 따라 결과가 달라진다. 위에 같은 10번 랜덤 데이터로 분석한 결과는 <표 4-11> 을 보면 rbf 커널의 평균 정확도가 82.84%로 모든 커널 중에 제일 좋은 성능을 보였다.

kernel	1	2	3	4	5	6	7	8	9	10	평균 정확도
linear	80.7	80.2	77.5	80.2	80.2	80.7	84.5	81.8	83.4	87.2	81.64
poly	78.1	80.7	78.6	82.9	80.2	75.9	82.3	84.0	84.0	83.4	81.01
sigmoid	73.3	72.2	70.6	74.9	72.7	69.5	74.3	71.1	73.3	73.8	72.57
rbf	82.4	81.3	79.7	85.0	82.4	81.8	86.1	79.1	85.0	85.6	82.84

<표 4-11> Support Vector Machine (SVM) 모델 추천 정확도

4.3.5.5 XGBoost

XGBoost 분류기 모델은 여러 파라미터 중에서 본 논문 데이터가 max_depth에 따라 정확도 결과가 크게 달라서 최적 max_depth부터 구하는 것을 진행했다. 아래 [그림 4-9] 를 보면 max_depth 3일 때 최적 정확도가 제일 높은 것으로 나타났다. 매번 랜덤 데이터를 바꾸면 최적 정확도의 max_depth도 변경된다.



[그림 4-9] XGBoost 분류기 모델 분석 max_depth도

아래 <표 4-12>는 위에 같은 10번 랜덤 데이터를 최적 max_depth를 구해서 세팅한 후 XGBoost 분류기 모델 통한 평균 정확도는 83.59%로 나타났다.

횟수	1	2	3	4	5	6	7	8	9	10
max_depth	4	6	4	4	4	3	5	7	4	4
정확도	81.8	82.4	80.2	83.4	82.9	82.4	85.0	84.5	86.1	87.2
평균 정확도	83.59									

<표 4-12> XGBoost 분류기 모델 추천 정확도

4.3.5.6 Stacking

위에 같은 10번 랜덤 데이터를 Stacking 기법 모델로 통한 평균 정확도는 82.95%로 나타났다.

1	2	3	4	5	6	7	8	9	10	평균 정확도
82.4	79.7	79.7	84.5	82.4	82.9	84.0	80.7	85.6	87.2	82.91

<표 4-13> Stacking 기법 모델 추천 정확도

V 실험 결과

중국 개별관광객에게 관광지 알맞은 추천하기 위해 모델 성능평가는 평균 정확도를 선택했다. 평균 정확도를 구하기 위해 같은 데이터를 지정한 상황에서 6개 모델로 분석한다. 이런 분석은 랜덤 데이터로 10번 진행한다.

결과는 아래 <표 4-14> 을 보면 자연 관광지에서 XGBoost 분류기 모델로 분석한 평균 정확도가 83.59%로 나오고 1순위를 기록했다. 체험 관광지에서도 XGBoost 분류기 모델은 77.01%의 평균 정확도로 1순위를 기록했다. 따라서 XGBoost 분류기 모델은 중국 개별관광객에게 관광지 추천에 있어서 우수한 성능이 보여주고 있다.

머신러닝 모델	자연 관광지 (%)	체험 관광지(%)
Logistic Regression 모델	81.05	65.78
Decision Tree 모델	82.25	71.12
K-Nearest Neighbours 모델	83.36	74.87
Support Vector Machine 모델	82.84	76.85
XGBoost 분류기 모델	83.59	77.01
Stacking 기법 모델	82.91	70.05

<표 4-14 > 알고리즘별 추천 정확도

VI 결론

1. 연구의 요약

본 논문은 각 관광객에게 맞는 취향과 편의를 모두 고려한 관광지 추천이 어려운 점을 보완해주기 위해 제주도에 이미 방문했던 중국 개별관광객의 설문 조사 데이터를 토대로 관광객의 속성과 형태와 관광지 간의 관계를 머신러닝 알고리즘으로 분석하여 최적의 선호 예측 모델을 만들어 머신러닝 기반의 관광지 추천 시스템을 제안했다. 이를 제안하기 위해 중국 관광객을 고객으로 하는 한 여행을 통해 제주도에 방문했던 관광객의 설문조사 데이터를 수집했고, 데이터 기본적인 분석에서 개별관광객과 형태의 빈도 분석, 개별관광객과 형태의 상관관계 분석을 시행했다.

또한, 머신러닝 기반 관광지 추천 모델로서 Logistic Regression 모델 분석, Decision Tree 모델 분석, K-Nearest Neighbours 모델 분석, Support Vector Machine (SVM) 모델 분석, XGBoost 분류기 모델 분석, Stacking 기법 모델 분석을 시행했다.

분석 결과에서 XGBoost 분류기가 다른 모델보다 높은 정확도를 나타내어 보여 주었다. 자연 관광지 추천에서는 83.59%의 평균 정확도 결과를 얻을 수 있었고, 체험 관광지 추천에서는 77.01%의 평균 정확도를 얻을 수 있었다.

2. 연구의 한계 및 향후 연구

첫째, 본 논문은 중국 개별관광객을 대상으로 진행한 연구이며, 향후 연구에 전체 중국 관광객 대상으로 연구할 예정이다. 둘째, 총 622명의 중국 개별관광객의 설문 데이터를 분석하였으나, 전체 개별관광객을 대표한다고 하기에는 무리가 있고, 누락된 경우의 수가 발생할 수 있으므로 향후 연구에서는 더 많은 데이터를 수집할 예정이다. 셋째, 코로나로 인해 제주에 방문하는 중국 관광객이 없어서 설문지 배부방식은 온라인 조사를 시행하였다. 조사한 데이터에서 여성의 답변 비율이 남성보다 많은 데 따른 데이터 불균형이 우려되므로 향후 연구에서는 남녀 답변 비율의 균형을 맞출 수 있도록 노력할 것이다.

[첨부자료]

제주 관광 관련 설문 조사

안녕하십니까? 제주에 방문해주셔서 감사합니다.

본 설문지는 제주도 중국인 개별관광객에게 편의성 및 만족도를 높이고 더 좋은 관광환경 및 서비스를 제공하기 위한 것입니다. 각 설문 항목에 대하여 진솔하게 답변해 주시면 큰 도움이 되겠습니다. 감사합니다.

제주대학교 대학원 컴퓨터공학과

지도교수: 변 영 철

연구자: 진 준 걸

인적사항

1. 성별: ① 남 ② 여
2. 연령대: ① 10대 ② 20대 ③ 30대 ④ 40대 ⑤ 50대 이상
3. 월수입 얼마:
① 3,000위안 이하 ② 3,001-5,000위안 ③ 5,001-8,000위안 ④ 8,001-12,000위안
⑤ 12,000위안 이상

방문 후 질문

1. 제주도를 어떻게 알게 되었습니까?
① TV ② 지인 ③ 여행사 ④ 인터넷 ⑤ 기타 ()
2. 한국 이외에 다른 나라를 방문하신 적이 있으신가요?
① 있다. ② 없다
3. 많은 관광지 중에 제주를 방문하시게 된 이유는 무엇입니까?
① 중국에 가깝다 ② 제주도 경치가 좋다 ③ 한국(제주) 문화 ④ 친구 추천
4. 어떤 목적으로 제주도에 방문하셨습니다?
① 관광 ② 쇼핑 ③ 음식 ④ 의료 ⑤ 기타 ()
5. 누구랑 같이 오셨습니까?
① 혼자 ② 가족 ③ 친구 ④ 동료
6. 단체관광이 아닌 개별 관광 선택된 이유 무엇인가요?
① 모르는 사람 같이 관광 싫다 ② 자유롭게 마음대로 옮기다 ③기타()
7. 관광계획을 짜고 오셨나요?
① 예 ② 아니요

8. 주로 어떤 교통수단을 이용하셨습니까?

- ① 가이드 카 ② 택시 투어 ③ 관광버스 ④ 일반 버스

9. 제주도는 총 몇 번 방문하셨습니까?

- ① 1번 ② 2번 ③ 3번 ④ 4번 ⑤ 5번 이상

10. 지난 여행에서 제일 좋았던 자연 관광지는 무엇인가요?

- ① 우도 ② 성산일출봉 ③ 외돌개 ④ 주상절리대 ⑤ 기타 ()

11. 지난 여행에서 제일 좋았던 체험은 무엇인가요?

- ① 서귀포잠수함 ② 김치 체험 ③ 한복체험 ④ 마린파크 돌고래체험 ⑤ 체험 없음 ⑥ 기타 ()

12. 지난 여행에서 제일 좋았던 박물관은 무엇인가요?

- ① 박물관은 살아있다. ② 테디베어 박물관 ③ 헬로키티 박물관 ④ 자동차 박물관 ⑤ 안 갔음 ⑥ 기타 ()

13. 지난 여행에서 제일 좋았던 테마파크는 무엇인가요?

- ① 에코랜드 ② 아쿠아플라넷 ③ 신화월드 ④ 코코몽 ⑤ 안 갔음 ⑥ 기타 ()

14. 지난 여행에서 좋았던 공연은 무엇인가요?

- ① 난타 ② 히어로 ③ 그림자 쇼 ④ 서커스월드 ⑤ 안 갔음 ⑥ 기타 ()

15. 지난 여행에서 좋았던 해변은 무엇인가요?

- ① 이호해변 ② 애월해변 ③ 월정리해변 ④ 협재해변 ⑤ 안 갔음 ⑥ 기타 ()

16. 지난 관광 일정은 며칠인가요?

- ① 2일 이하 ② 3일 ③ 4일 ④ 5일 ⑤ 5일 이상

17. 제주도 관광에 만족하십니까?

- ① 아주 그렇다 ② 그렇다 ③ 그저 그렇다 ④ 그렇지 않다

18. 제주도에 다시 방문하고 싶으신가요?

- ① 많이 가고 싶다 ② 가고 싶다 ③ 보통 ④ 안가고 싶다 ⑤ 절대 안 간다

19. 제주도 관광했을 때 중국어 안내 정보에 대해 어떻게 생각하시나요?

- ① 매우 만족 ② 만족 ③ 보통 ④ 불만족 ⑤ 매우 불만족

20. 제주도 관광에 개선해야 할 점으로 어떤 것이 있을까요?

- ① 중국어 안내가 부족하다 ② 교통이 불편하다 ③ 오락시설이 부족하다 ④ 서비스가 친절하지 않다 ⑤ 기타()

濟州島 旅游關聯 調查問卷

您好！感謝您訪問濟州島

這份問卷調查是爲了提供給中國游客更好的便利性,更高的滿意度,建設更好的旅游环境而進行的調查問卷，希望每一个回答都認真，您的回答對我們來說會有很大的幫助。謝謝您。

濟州大學 大學院 電腦工程系
指導教授：Yung-Cheol Byun
研究員：秦俊杰

人事項問題

- 1.性別: ① 男 ②女
- 2.年齡段: ① 10代 ② 20代 ③ 30代 ④ 40代 ⑤ 50代以上
- 3.月收入是多少?
① 3000元以下 ② 3001-5000元 ③ 5001-8000元 ④ 8001-12000元
⑤ 12000元以上

一般問題

- 1.您是怎么知道濟州島的呢?
① TV ② 朋友 ③ 旅行社 ④ 網絡*雜誌 ⑤ 其他 ()
2. 您去過除了濟州島以外的韓國的其他的國家嗎?
① 有 ② 沒有
3. 爲什么選擇了濟州島?
① 離中國近 ② 濟州島景色好 ③ 韓國 (濟州) 文化 ④ 朋友推荐
- 4.什么目的來的濟州島?
① 旅游 ② 購物 ③ 美食 ④ 医療 ⑤ 其他 ()
- 5.和誰一起來的濟州島呢?
① 自己 ② 家庭 ③ 朋友 ④ 同事
- 6.爲什么選擇了自由行?
① 不喜歡和不認識的人一起旅游 ② 自由行更自由沒有約束 ③其他()
- 7.您提前做好行程准備了嗎?
① 有 ② 沒有
- 8.主要用了什么交通工具旅游?
① 導游包車② 出租車 ③ 日游旅游車 ④ 公交車
- 9.您是第几次來濟州島呢?
① 1次 ② 2次 ③ 3次 ④4次 ⑤5次 以上

10. 之前濟州旅游時您最喜歡那个自然景區呢？

①牛島 ② 城山日出峰 ③獨立岩 ④ 柱狀節理帶 ⑤ 其他 ()

11. 之前濟州旅游時您最喜歡哪个體驗觀光呢？

①西歸浦潛水艇 ② 泡菜體驗 ③韓服體驗 ④ 海豚體驗 ⑤沒有參加體驗 ⑥其他 ()

12. 之前濟州旅游時您最喜歡的博物館是哪个好呢？

①泰迪熊博物館 ② 活生生博物館 ③凱蒂貓博物館 ④ 汽車博物館 ⑤沒有去 ⑥其他 ()

13. 之前濟州旅游時您最喜歡的主題公園是哪个好呢？

①森林小火車 ② 海洋館 ③神話世界 ④ 可可蒙儿 ⑤沒有去主題公園 ⑥其他 ()

14. 之前濟州旅游時您最喜歡的演出是哪个好呢？

①亂打秀 ② 塗鴉秀 ③影子秀 ④ 雜技表演 ⑤沒有看演出 ⑥其他 ()

15. 之前濟州旅游時您最喜歡的海邊是哪个好呢？

①梨湖海邊 ② 涯月海邊 ③月汀里海邊 ④ 挾才海邊 ⑤沒有去 ⑥其他 ()

16. 之前濟州旅游時旅游時是几天日程呢？

① 2天或以下 ② 3天 ③ 4天 ④ 5天 ⑤ 5天以上

17. 您對這次濟州之旅滿意嗎？

①很滿意 ②滿意 ③一般 ④不滿意 ⑤很不滿意

28. 想再來一次濟州島嗎？

① 很想 ② 想 ③一般 ④不想 ⑤絕對不想

19. 在濟州島旅游中，濟州島中國語觀光介紹信息提供是否滿意？

①很滿足 ②滿足 ③一般 ④不滿足 ⑤很不滿足

20. 您覺得濟州島旅游方面需要改進的有哪些呢？

① 中國語信息不足 ② 交通不便利 ③ 娛樂設施不足 ④ 旅游服務不好 ⑤ 其他 ()

참고문헌

- [1]한국관광공사(2020). 2019외래 관광객 조사 최종 보고서
- [2]김중열(2014). 국내 관광산업의 관광수지 결정요인에 관한 연구, 부산대학교 경제 통상대학원 석사학위논문,pp.1-2.
- [3]중국통신원 (2019).〈중국 정보 소비발전태세 및 전망보고서〉
- [4]중국산업정보망(2020). <http://www.chyxx.com/industry/202003/845172.html>
- [5]한국관광공사(2019). 2019년 한국 외래 관광객 실태 조사 최종 보고서
- [6]최학현·오세형·이주현·김정희(2008). 실버세대의 사용편의성을 위한 Digital(On-line) TV 인터페이스 디자인, 『디지털디자인학연구』,8(1).p.384
- [7]한국관광공사(2014). 2013년 한국 외래 관광객 실태조사보고서.
- [8]제주도청 홈페이지(2020) 관광객 입도 현황 표
- [9]중국문화관광부(2020) 2019년 전국여행사통계보고서
http://zwgk.mct.gov.cn/auto255/202008/t20200824_874310.html
- [10]김사현(2003). 비 제도권 개별관광자의 속성과 영향 pp.131-148.
- [11](오정 2018) 제주방문 중국인 개별관광객의 숙박시설 선택속성과 만족도에 관한 연구
- [12]한국관광공사(2020) 2019외래 관광객 조사 최종 보고서
- [13]한국관광공사(2008): www.visitkorea.or.kr
- [14]김사현(2003). 비 제도권 개별관광자의 속성과 영향. pp.1-3.
- [15] 김도훈, 김영규(2017). 순수 관광목적 가진 중국인 개별 자유 여행객의 한국 관광지 선택속성중요도-만족도 연구

- [16] (조현제·이필규, 2014) 추천 시스템 기반 초등학생 인공지능 교육프로그램 개발 및 적용, p8.
- [17][네이버 지식백과] 추천 시스템
- [18]류슬기(2019) 음악 가사의 분석을 통한 사용자기반의 음악 추천 시스템 제안, pp 4-5.
- [19][위키백과] 추천 시스템 Content-based filtering
- [20][위키백과] 추천 시스템 Collaborative filtering
- [21]김하연(2018). 사용자 기반 협업 필터링 알고리즘의 이웃 확장 연구,pp 8-11.
- [22]김하연(2018). 사용자 기반 협업 필터링 알고리즘의 이웃 확장 연구,pp 5-8.
- [23] <https://blog.naver.com/rmby7484/221973700822>
- [24] [인사이트 캠퍼스] 지도학습의 예시 : KNN(K-Nearest Neighbors) 알고리즘|작성자
판인사이트
- [25] [위키백과] 로지스틱 회귀
- [26][네이버 지식백과] 서포트 벡터 머신 [support vector machine] (생화학백과)
- [27][네이버 지식백과] 결정 트리
- [28]하지은(2017). RandomForest와 XGBoost를 활용한 민원 카테고리 및 담당 부서 자동분류 성능 비교 p5
- [29]신재욱(2019). IoT센서를 활용한 머신러닝(SVM)기반 실시간 운동자세 측정시스템,pp 23-24.
- [30] (2015. 박연진, 송경아, 황재원, 창병모) 온톨로지 기반의 개인화된 여행 추천 시스템
- [31] 안진현 임동혁 (2020) 이벤트와 관련된 주변 관광지 자동 추천 알고리즘 개발 연구
- [32]류기환, 윤수현, 문석재(2020) MBTI와 Big Five 모델에 근거한 성격유형에 따른 토픽 맵을 이용한 관광지 추천에 관한 연구

ABSTRACT

Machine learning-based tourist attraction recommendation reflecting the behavior of individual Chinese tourists

QIN JUNJIE

Department of Computer Engineering
The Graduate School of Jeju National University

Jeju Island is regarded as a very attractive place for Chinese tourists. Among all the foreign tourists visited Jeju Island in the past few years, China ranks first. Due to the diplomatic conflicts between China and Korea, Chinese government banned the group tourism heading to Korea. As a result of this, Chinese tourists visiting Jeju Island have mainly become individual tourists rather than group tourists so far. Most of the individual tourists utilize the websites or applications to search the information on tourist attractions, food information, shopping information, etc. Before travel, tourists usually choose their favorite scenic spots and plan their schedule. Therefore, a suitable recommendation system is helpful to select tourist attractions. The proper recommendation system of tourist attractions leaves a good memory for tourists and improves the intention of revisiting. However, the wrong recommendation system will not only causes a waste of money and time but also leads to terrible travel memories. Therefore, if the Jeju tour is undervalued, the visiting intention will naturally decrease as well as the

revisit to Jeju. So, it is meaningful to study the tourist attractions recommendation system which is suitable for tourists' preferences.

The existing recommendation systems have two main methods, a collaborative filtering approach that models and measures the similarity of an item or user, and recommends content-based filtering by analyzing the content itself filtering approach.

When recommending tourist destinations, preferred tourist destinations varies according to various attributes and types such as gender, age group, monthly income, the route to contact Jeju Island tour, the purpose of visit, travel companion, and a number of visits to Jeju. Factors related to visits are created through data analysis to make more suitable tourist attractions.

Therefore, in this study, a preference prediction model is created using a machine learning algorithm based on the attributes and forms of tourists who have already visited Jeju, and the satisfaction evaluation for tourist destinations, and through this, a machine learning-based tourist destination recommendation system is proposed.

In this paper, data obtained from individual Chinese tourists who visited Jeju Island, and by analyzing the data, a logistic regression model, a decision tree model, and K-nearest neighbors. Model, support vector machine (SVM) model and XGBoost classifier were used to experiment. As a result of the experiment, the tourist site recommendation using the XGBoost classifier showed an average accuracy of 83.59%, showing better performance than other models.