<table>
<tr><td rowspan="8">메타데이블
정보
(다중기입가능)</td><td>분야</td><td>데이터 유형[1)]</td><td>구축 데이터량</td><td>원천데이터
형식[2)]</td><td>라벨링 형식[3)]</td><td>라벨링 유형[4)]</td></tr>
<tr><td>관광</td><td>텍스트</td><td>16,000set</td><td>csv</td><td>csv</td><td>로그데이터
(텍스트)</td></tr>
<tr><td>데이터 출처[5)]</td><td>데이터 구축년도</td><td>구축기관(총괄)</td><td>가공기관</td><td>검수기관</td><td></td></tr>
<tr><td>자체 수집</td><td>2022년</td><td>㈜데이터웨이</td><td>㈜지디에스컨설팅그룹</td><td>㈜데이터웨이</td><td></td></tr>
<tr><td rowspan="2">데이터
문의처</td><td>기관명</td><td>문의담당자명</td><td>전화번호
(유선전화번호기입)</td><td>메일주소</td><td></td></tr>
<tr><td>㈜데이터웨이</td><td>김정남</td><td>02-2205-4500</td><td>33823698@data-way.co.kr</td><td></td></tr>
<tr><td>데이터 소개</td><td colspan="5">수도권, 동부권, 서부권, 제주 및 도서지역 각 권역별로 4천세트 씩,
총 16,000세트의 여행로그 데이터를 구축</td></tr>
<tr><td>주요키워드</td><td colspan="5">여행로그</td></tr>
<tr><td colspan="2">카테고리 정의서</td><td colspan="5">3-005_국내여행로그데이터수집_카테고리정의서.xlsx</td></tr>
</table>

<table>
<tr><td rowspan="8">메타데이블
정보
(다중기입가능)</td><td>분야</td><td>데이터 유형[1)]</td><td>구축 데이터량</td><td>원천데이터
형식[2)]</td><td>라벨링 형식[3)]</td><td>라벨링 유형[4)]</td></tr>
<tr><td>관광</td><td>이미지</td><td>726,522장</td><td>jpg</td><td>csv</td><td>사진정보
(텍스트)</td></tr>
<tr><td>데이터 출처[5)]</td><td>데이터 구축년도</td><td>구축기관(총괄)</td><td>가공기관</td><td>검수기관</td><td></td></tr>
<tr><td>자체 수집</td><td>2022년</td><td>㈜데이터웨이</td><td>㈜지디에스컨설팅그룹</td><td>㈜데이터웨이</td><td></td></tr>
<tr><td rowspan="2">데이터
문의처</td><td>기관명</td><td>문의담당자명</td><td>전화번호
(유선전화번호기입)</td><td>메일주소</td><td></td></tr>
<tr><td>㈜데이터웨이</td><td>김정남</td><td>02-2205-4500</td><td>33823698@data-way.co.kr</td><td></td></tr>
<tr><td>데이터 소개</td><td colspan="5">수도권, 동부권, 서부권, 제주 및 도서지역 각 권역별로 관광사진 데이터 80,000장 이상 씩, 총 726,522장 광광사진 데이터를 수집</td></tr>
<tr><td>주요키워드</td><td colspan="5">관광사진</td></tr>
<tr><td colspan="2">카테고리 정의서</td><td colspan="5">3-005_국내여행로그데이터수집_카테고리정의서.xlsx</td></tr>
</table>

1) 텍스트, 오디오, 이미지, 비디오,
2) txt, jpg,.....
3) json, csv,....
4) 내용요약(텍스트), 번역(자연어), 질의응답(자연어), 바운딩박스(이미지/동영상), 키포인트(이미지/동영상), 세그멘테이션(이미지/동영상), 전자(음성) .....
5) 4대 언론기사, 자체 수집,.,.,,

<table>
<tr><td rowspan="2">데이터셋 명</td><td>국문</td><td>국내 여행로그 데이터</td></tr>
<tr><td>영문</td><td>domestic travel log data</td></tr>
<tr><td colspan="2">구축목적</td><td>○ 여행자의 이동패턴과 소비내역, 활동 내역 등 데이터 수집
○ 관광업계 자체적으로 수집하기 어려운 양질의 AI데이터 제공
○ AI기술을 활용한 관광산업 혁신 생태계 구축
○ AI기술 기반의 개인화된 서비스로 관광객들의 경험 향상</td></tr>
<tr><td colspan="2">활용서비스</td><td>
<table>
<tr><th>학습 모델</th><th>알고리즘</th><th>성능지표</th><th>선정 사유</th></tr>
<tr><td>여행객 정보 기반 고지출 여행객 예측 모델</td><td>Pycaret</td><td>F1-score 0.70이상</td><td>- Data Leakage를 막기 위해 여행객의 사전 정보, 페르소나, 소득 수준, 호텔 예약 정보 등 여행 출발 이전부터 알 수 있는 정보를 정제한 후 선정
- EDA를 통해 각 변수별 특성을 파악하고 소비 지출에 영향을 주는 변수에 무엇이 있는지 1차적으로 확인
- 2D Tensor Data의 분류 예측 문제는 일반적으로 트리 기반 부스팅 모형이 가장 좋은 성능을 보이나 데이터 전체 데이터 개수가 크지 않은 경우 Over fitting의 문제를 고려해야 함
- 앙상블 이외의 모형이 더 좋은 성능을 보일 가능성을 배제할 수 없기 때문에 데이터 전처리 이후 Pycaret 알고리즘을 통해 Validation Data Set에 가장 높은 성능을 보이는 모델을 선정하고 추가 작업하여 최종 모델 선정</td></tr>
<tr><td>여행객 선호도기반 여행 장소 추천 알고리즘</td><td>Essemble model</td><td>Recall@10 0.25</td><td>- 추천시스템은 전통적으로 협력필터링, 컨텐츠 기반 시스템 그리고 이 둘의 장점을 합친 하이브리드 모델이 존재
- 최근 인공지능 분야의 비약적인 발전과 더불어 오토인코더와 같은 딥러닝 모델들이 추천시스템에 적용
- 그러나 본 개발 모델의 경우 추천 성능 못지않게 이후 확장 가능성 및 모델 결과에 대한 분석 그 자체가 중요함
- 따라서 기존 추천시스템에서 사용되던 딥러닝 기반 모델보다 사용자 정보를 넣었을 때 선호도를 예측하는 Regression 기반의 모델을 사용하여 추천 장소를 선정하는 방식을 선택
- 일반적으로 2D Tensor 데이터에서 가장 좋은 성능을 보이는 Random Forest, Cat Boost, LGBM, XG Boost등의 모델을 후보 모델로 선정함</td></tr>
</table>
</td></tr>
<tr><td colspan="2">소개</td><td>o 국내 여행로그 데이터 구축 필요성</td></tr>
</table>

- 코로나19 확산 이후 언택트 및 디지털 관광으로의 전환이 가속화되고 있음
- 인구구조와 여행패턴의 변화로 개인 맞춤형 관광 서비스가 본격화되고 있음
- 관광산업 혁신을 위해 AI 기술을 적극 도입해야 하며 이를 위한 학습용 데이터 필요

ㅇ 데이터 구축 내용
- 국내를 수도권, 동부권, 서부권, 제주 및 도서지역 등 4가지 권역으로 나누어 여행객을 모집하고, 스마트폰 전용앱을 통해 데이터를 수집
- 수도권, 동부권, 서부권, 제주 및 도서지역 각 권역별로 4천세트 씩, 총 16,000세트의 여행로그 데이터를 구축
- 여행동선 데이터, 소비내역 데이터, 활동기록 데이터, 여행지 사진 등을 수집
- 수집된 데이터를 정제한 후, 촬영사진 블러링, 영수증 Key-In 등의 가공작업 수행
- 고지출 여행객 예측모델 : F1-Score 80.07% 달성 (목표 70%)
- 여행장소 추천모델 : Recall@10 0.3745 달성 (목표 0.25)

ㅇ 시범 AI 모델 개발
- 여행자 정보 기반 고지출 여행객 예측 모델 : 여행객들의 사전조사 정보를 토대로 여행지에서 지출을 많이 하는 여행객들을 예측
- 여행자 선호도 기반 여행장소 추천모델 : 성별, 연령대, 소득 등의 여행객 정보와 여행지역(시도/시군구) 정보를 입력하면 10개의 여행지를 추천

데이터셋 통계 (구축 규모 및 분포)

1. 데이터 구축 규모

구분		구축실적
[3-005-277] 수도권	여행자 정보 (여행자 패널 데이터)	4,000 SET
	동선 정보 (GPS 데이터)	4,000 SET
	활동정보 (여행기록 데이터)	4,000 SET
	활동정보 (여행지 사진 데이터)	135,183 장
	소비 내역 (소비내역 데이터)	4,000 SET
	POI 데이터	1 Set
[3-005-278] 동부권	여행자 정보 (여행자 패널 데이터)	4,000 SET
	동선 정보 (GPS 데이터)	4,000 SET
	활동정보 (여행기록 데이터)	4,000 SET
	활동정보 (여행지 사진 데이터)	160,237 장
	소비 내역 (소비내역 데이터)	4,000 SET
	POI 데이터	1 Set
[3-005-279] 서부권	여행자 정보 (여행자 패널 데이터)	4,000 SET
	동선 정보 (GPS 데이터)	4,000 SET
	활동정보 (여행기록 데이터)	4,000 SET
	활동정보 (여행지 사진 데이터)	161,444 장
	소비 내역 (소비내역 데이터)	4,000 SET
	POI 데이터	1 Set
[3-005-280] 제주도 및 도사지역	여행자 정보 (여행자 패널 데이터)	4,000 SET
	동선 정보 (GPS 데이터)	4,000 SET
	활동정보 (여행기록 데이터)	4,000 SET
	활동정보 (여행지 사진 데이터)	269,658 장
	소비 내역 (소비내역 데이터)	4,000 SET
	POI 데이터	1 Set

2. 데이터 분포

		수도권		동부권		서부권		제주/도서	
성별	남	1,639	41%	1,563	39%	1,524	38%	1,488	37%
	여	2,361	59%	2,437	61%	2,476	62%	2,512	63%
연령별	20대	1,383	35%	1,335	33%	1,382	35%	1,398	35%
	30대	1,421	36%	1,321	33%	1,376	34%	1,366	34%
	40대	633	16%	652	16%	613	15%	627	16%
	50대↑	563	14%	692	17%	629	16%	609	15%
여행 기간별	당일	2,192	55%	1,792	45%	1,895	47%	536	13%
	1박 2일	1,252	31%	1,532	38%	1,551	39%	926	23%
	2박 3일 이상	556	14%	676	17%	554	14%	2,538	63%

수도권	서울	경인	계
	1,375	2,625	4,000

동부권	강원	대구	경북	부산	울산	경남	계
	1,561	218	961	558	171	531	4,000

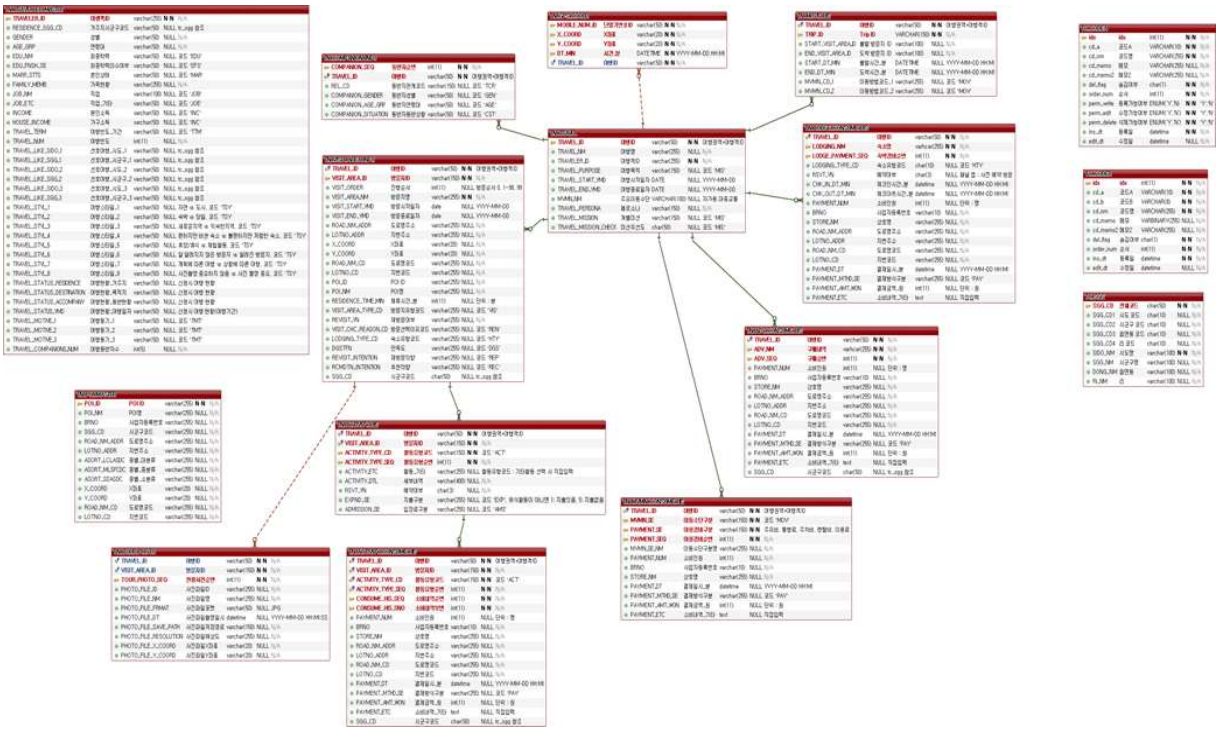
서부권	대전	충남/세종	충북	광주	전남	전북	계
	338	1,109	636	142	963	812	4,000

제주도 및 도서지역	제주	도서지역	계
	3,094	906	4,000

[3-005_국내여행로그데이터수집_데이터분포.png]

데이터셋 구성

1. 데이터 구축 ERD



[3-005_국내여행로그데이터수집_ERD.png]

2. 데이터 구성

데이터구분	데이터	데이터 명	수량
[3-005-277] 수도권	여행로그 데이터	tc_codea_코드A.csv	각csv 파일별 4,000Set 구성
		tc_codeb_코드B.csv	
		tc_sgg_시군구코드.csv	
		tn_activity_consume_his_활동소비내역_A.csv	
		tn_activity_his_활동내역_A.csv	
		tn_adv_consume_his_사전소비내역_A.csv	
		tn_companion_info_동반자정보_A.csv	
		tn_lodge_consume_his_숙박소비내역_A.csv	
		tn_move_his_이동내역_A.csv	
		tn_mvmn_consume_his_이동수단소비내역_A.csv	
		tn_tour_photo_관광사진_A.csv	
		tn_traveller_master_여행객 Master_A.csv	
		tn_travel_여행_A.csv	
		tn_visit_area_info_방문지정보_A.csv	
		tn_poi_master_POIMaster.csv	POI Master
	gps_Data	n_gps_coord_*.csv [* = 여행객 ID]	4,000개
	photo	여행객ID + 순번. jpg	135,183개
[3-005-278] 동부권	여행로그 데이터	tc_codea_코드A.csv	각csv 파일별 4,000Set 구성
		tc_codeb_코드B.csv	
		tc_sgg_시군구코드.csv	
		tn_activity_consume_his_활동소비내역_B.csv	
		tn_activity_his_활동내역_B.csv	
		tn_adv_consume_his_사전소비내역_B.csv	
		tn_companion_info_동반자정보_B.csv	
		tn_lodge_consume_his_숙박소비내역_B.csv	
		tn_move_his_이동내역_B.csv	
		tn_mvmn_consume_his_이동수단소비내역_B.csv	
		tn_tour_photo_관광사진_B.csv	
		tn_traveller_master_여행객 Master_B.csv	
		tn_travel_여행_B.csv	
		tn_visit_area_info_방문지정보_B.csv	
		tn_poi_master_POIMaster.csv	POI Master
	gps_Data	n_gps_coord_*.csv [* = 여행객 ID]	4,000개
	photo	여행객ID + 순번. jpg	160,237개
[3-005-279] 서부권	여행로그 데이터	tc_codea_코드A.csv	각csv 파일별 4,000Set 구성
		tc_codeb_코드B.csv	
		tc_sgg_시군구코드.csv	
		tn_activity_consume_his_활동소비내역_C.csv	
		tn_activity_his_활동내역_C.csv	
		tn_adv_consume_his_사전소비내역_C.csv	
		tn_companion_info_동반자정보_C.csv	
		tn_lodge_consume_his_숙박소비내역_C.csv	
		tn_move_his_이동내역_C.csv	
		tn_mvmn_consume_his_이동수단소비내역C.csv	
		tn_tour_photo_관광사진_C.csv	

		tn_traveller_master_여행객 Master_C.csv	
		tn_travel_여행_C.csv	
		tn_visit_area_info_방문지정보_C.csv	
		tn_poi_master_POIMaster.csv	POI Master
	gps_Data	n_gps_coord_*.csv [* = 여행객 ID]	4,000개
	photo	여행객ID + 순번. jpg	161,444개
[3-005-280] 제주도 및 도사지역	여행로그 데이터	tc_codea_코드A.csv	각csv 파일별 4,000Set 구성
		tc_codeb_코드B.csv	
		tc_sgg_시군구코드.csv	
		tn_activity_consume_his_활동소비내역_D.csv	
		tn_activity_his_활동내역_D.csv	
		tn_adv_consume_his_사전소비내역_D.csv	
		tn_companion_info_동반자정보_D.csv	
		tn_lodge_consume_his_숙박소비내역_D.csv	
		tn_move_his_이동내역_D.csv	
		tn_mvmn_consume_his_이동수단소비내역_D.csv	
		tn_tour_photo_관광사진_D.csv	
		tn_traveller_master_여행객 Master_D.csv	
		tn_travel_여행_D.csv	
		tn_visit_area_info_방문지정보_D.csv	
		tn_poi_master_POIMaster.csv	POI Master
	gps_Data	n_gps_coord_*.csv [* = 여행객 ID]	4,000개
	photo	여행객ID + 순번. jpg	269,658개

데이터셋 구축 수행기관 담당자						
	주관기관	기관명	책임자명	전화번호 (유선전화번호기입)	메일주소	담당업무
		㈜데이터웨이	김정남	02-2205-4500	33823698@data-way.co.kr	실무책임자
	참여기관	기관명	담당업무	기관명	담당업무	
		㈜케이스탯 리서치	여행자운영 데이터수집	㈜에이드리븐	여행자 모집/관리	
		㈜지디에스 컨설팅그룹	데이터가공	㈜티지360	여행자 모집/관리	
		㈜올포랜드	공간데이터 정제/가공	와이비에스 에듀	데이터가공	
		고려대학교 산학협력단	검증용AI 알고리즘	㈜데이터웨이	사업관리 데이터검수	