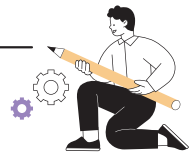
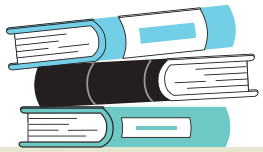
2024
공공부문
데이터 분석·활용
우수사례집

행정안전부 NIA 한국지능정보사회진흥원

Contents





교통
안전

PART 1

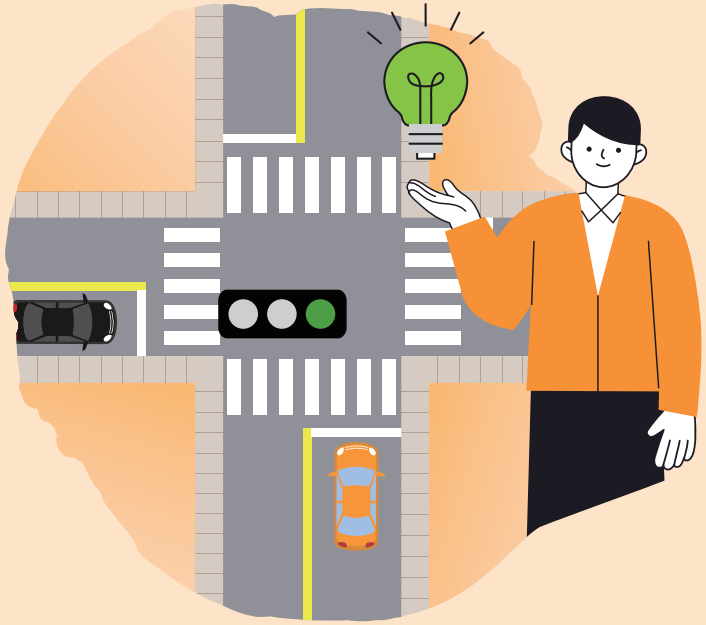
교통안전

PART 1-1
교통안전

우리 도시 교차로는 언제나 그린라이트

모빌리티 빅데이터 기반 교통신호 최적화 사례

진주시청



추진목적/배경

부적절한 교통신호 운영은 교통혼잡의 주요 원인 중 하나로, 국가 수송 경쟁력 저하, 자동차 배기가스로 인한 환경오염 등 다양한 사회적 문제를 초래한다. 그러나 대부분의 지자체는 많은 비용이 소요되는 고정식 교통 검지 인프라를 설치하고 운영하기에는 예산상 제약이 있어 신호 운영 개선에 어려움을 겪고 있다. 더욱이 고정식 교통 검지기는 투입된 비용에 비해 수집되는 교통정보가 매우 제한적이다.

이러한 문제를 해결하기 위해 진주시청은 저예산-고효율의 모빌리티 빅데이터를 활용하여 교통신호의 효율적인 운영을 지원할 수 있는 데이터셋과 모빌리티 빅데이터 기반의 Infra-free 공용 신호 운영 솔루션 개발을 추진하였다. 이를 통해 교통신호 운영 분석 및 서비스 향상에 대한 기회를 제공하고자 했다. 막대한 초기 예산이 필요한 교통 검지 인프라의 구축 없이, 교통신호 운영에 최적화된 회전별·방향별·차종별 교통량, 대기행렬, 정지 수 등을 수집, 생성, 분석하는 기술을 이용하여 신호 운영 개선을 수행하였다.

분석 사전 준비

– 활용 데이터

데이터명	형태	내용	출처	기준 년도	내·외부 데이터
차량GPS데이터 (내비게이션)	TXT (4TB)	차량의 위치 및 이동을 추적할 수 있는 포인트 궤적 데이터	한국교통연구원	2021	외부
관측교통량 (방범용CCTV기반)	CSV (5GB)	방범용CCTV 기반 교통량 계측자료(위치정보 및 방향정보 포함)	진주시청	2021	내부
관측교통량 (스마트교차로 기반)	CSV (3GB)	스마트 교차로 기반 회전교통량 (위치정보 및 방향정보 포함)	진주시청	2021	내부
신호운영DB	XLSX (10MB)	교차로별 신호운영 TOD Table자료	진주시청	2021	내부
고해상도 교통 네트워크	SHAPE, DB(30MB)	노드-링크 및 기타 부가정보를 포함하는 공간정보	한국교통연구원	2021	외부

차량GPS 궤적데이터와 고해상도 교통네트워크 데이터는 한국교통연구원에서 국가통합교통체계효율화법에 의거한 KTDB 사업을 통해 보유한 자료를 활용하였다. 차량GPS 데이터는 단말기 장비의 식별ID와 차량번호를 제외하고 출발지와 도착지 위치정보의 일부를 삭제하여 개인정보 노출을 방지하였다.

분석과정

분석 프로세스 도식화

– 분석 환경

1. 분석 인프라 : 한국교통연구원 내 PC 이용
2. 분석 환경 : Transit-7F(교차로 용량 분석 프로그램 HCS, 교통 시뮬레이션 분석 프로그램 COSSIM 포함 제품), Synchro, Passer 등

– 데이터 수집

- 모빌리티 빅데이터 기반의 통합 교통 신호운영 데이터 셋 구축을 통해 기존 신호운영의 최적화와 모니터링 문제를 해결하고자 함

1. 사용 데이터 출처 : 기관 자체 생성 데이터 및 사업참여 기관 데이터
2. 사용 데이터 형식 : csv, xlsx, shape 등

– 데이터 전처리

1. 데이터 파일 단위 통일화
   - 출발 시간을 기준으로 데이터를 정렬
   - 1일 단위로 가공 및 적재

2. 차량 GPS 데이터 이상치 판단 및 오차구간 보정(정교화)
   - 음영 구간(지하차도, 터널, 고가 밑 등), 고층 빌딩 주변, 신호 대기 상태에서 GPS 수신 불안정
   - 재구조화 알고리즘 적용하여 차량 궤적 데이터 정교화
3. 출도착 궤적 분리
   - 1일 단위 연속 수집되는 차량 GPS 데이터를 출발지와 도착지 기준으로 통행 분리
   - 출도착 궤적분리 모듈 활용
4. 차량 GPS 데이터와 교통 분석용 네트워크 맵 매칭 및 경로 생성
   - 기초 교통 및 통행 지표 DB는 교통 분석용 네트워크 단위로 구축
   - 차량 GPS 데이터는 기초 교통 및 통행 지표 구축 시 주요 기초 데이터로 활용
   - 지표 생성 단위와 동일한 교통 분석용 네트워크 기준으로 맵 매칭 후 링크 기준 경로 데이터 생성
   - 차량 GPS 데이터 위치 정보를 기준으로 각 포인트의 진행 방향각 계산 (방향각은 진북 기준 시계방향)
   - 포인트 주변 도로 네트워크(링크) 검색 후 포인트와 링크의 최단 거리, 링크에서의 방위각, 링크를 따라 이동한 거리 정보 계산

   → 최대 비용을 나타내는 경로 선정 및 적재
5. 통합 병합
   - 사용자가 휴게소에서 내비게이션 종료 또는 시작 시 차량 GPS 데이터 통행 분리

     → 시종점으로 적합하지 않은 구간은 연속된 통행 유지 처리
   - 휴게소-휴게소, 고속도로-고속도로, 도시고속도로-고속도로 조합 시 병합
   - 병합 시 시간 정보 및 구분 코드 추가하여 데이터 분석 및 검증에 활용

- 모델링

▶ 교통운영을 위해 한국교통연구원에서 개발한 5대 기술*

* 교통운영 DB를 기반으로 한 신호운영 및 모니터링 DB 구축 핵심 기술

① 교차로 교통량 산정 기술

→ 관측이 불가능하거나 누락된 교차로에 대한 접근로별/회전방향별/차종별/시간대별 교통량 산정

② 궤적 재구성 기술

→ 차량의 운행 형태를 고려한 합리적 이상치 제거 및 보정기술

③ 궤적 기반 교차로 교통 변수 계측 기술

→ 회전방향별 교통변수(속도, 대기행렬길이, 정지수, 정지지체 등) 계측 기술

④ 궤적 기반 신호 운영 상태 모니터링 기술

→ 운영 중인 교차로의 신호 현시 계획하의 실제 차량 주행 상태 모니터링 기술

⑤ 교통 분석용 GIS 맵

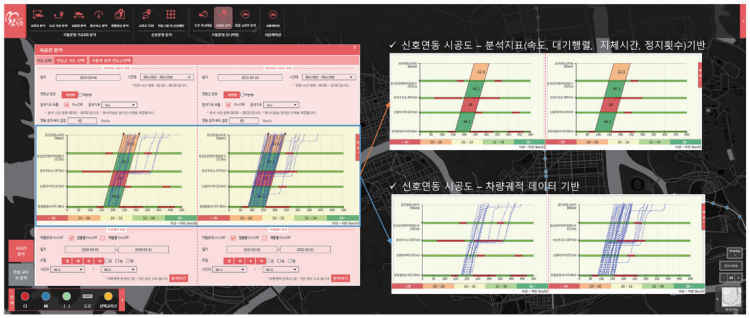
→ 교차로 분석에 최적화된 GIS 맵 구성 기술

- 결과 구현

▶ 모빌리티 빅데이터 기반의 Infra-free 공용 신호운영 솔루션

- 진주시 전역의 교통현황을 종합적으로 모니터링하고 문제지역을 감지 및 대응하기 위한 기반환경 구축
- 구축한 신호운영 데이터 셋을 관리자가 보다 효율적으로 모니터링하고 분석할 수 있도록 신호운영 모니터링·분석 온라인 서비스 구축
- 신호운영 모니터링·분석 온라인 서비스는 교통운영 기초 DB분석 기능, 신호운영현황 분석 기능, 교통현황 모니터링 기능으로 구분

〈교통운영 기초DB분석 기능〉



〈신호운영현황 분석 기능〉

〈교통현황 모니터링 기능〉

정책활용/기대효과

본 사업에서는 구축한 신호운영 데이터셋 분석기능을 활용하여 진주시 주요 축인 동진로, 대신로, 월아산로에 위치한 50개 교차로를 테스트베드로 선정하였다. 이 테스트베드는 기하구조와 신호 운영 측면에서 복합적인 문제들이 얽혀 있었으며, 본 사업은 신호운영 개선에 중점을 두고 진행되었다. 신호 운영 개선을 위해 통행 행태 기반 SA 그룹 설정, 교차로 교통량을 활용한 TOD 개선, 연동 체계 최적화를 수행하였고 '23년 12월 2~3주에 현장적용을 수행하였다. 현장적용 결과 동진로와 대신로의 평균속도는 약 0.3~1.3km/h 증가하였으며, 정지횟수는 0.1~0.7회 감소하여 신호운영 개선효과가 있음을 확인할 수 있었다.

[사업 대상지역 위치도]

[시청사거리 → 공단광장교차로방면]

[공단광장교차로 → 시청사거리방면]

〈테스트 베드 선정〉

〈진주시 신호운영 개선을 위한 현장적용〉

또한 구축한 신호운영 데이터 셋과 기능을 활용하여 진주시를 대상으로 신호운영 개선을 수행한 결과 데이터 수집 커버리지를 확대하고 교통혼잡을 완화하며, 교통혼잡 비용을 절감하는 효과를 거두었다. 데이터 수집 커버리지는 기존 2.9%에서 100%로 확대되었으며, 평균속도는 약 1.2%~4.7% 증가했다. 마지막으로 평균속도 증가에 따라 교통혼잡비용이 약 4.8% 감소한 것으로 분석되었다.

성과지표		개선 전	개선 후	비고
데이터 수집 커버리지 확대		- 13개 교차로(2.9%)	- 425개 교차로(100%)	- 교차로별·차종별·시간대별 회전교차로 교통량 데이터 100%구축
교통혼잡완화	평균속도	- 동진로: 33.2km/h - 대신로: 33.1km/h	- 동진로: 34.1km/h - 대신로: 33.7km/h	- 조사구간 평균속도 1.2%~4.7% 증가
	평균정지횟수	- 동진로: 3.7회 - 대신로: 4.2회	- 동진로: 3.6회 - 대신로: 3.7회	- 조사구간 정지횟수 2.7%~18.9 감소
	평균제어지체	- 75.6초/대	- 63.8초/대	- 공단광장교차로(CI) 평균 15.25 초/대 감소
교통혼잡 비용절감		- 동진로: 4,540만원/일 - 대신로: 3,937만원/일	- 동진로: 4,281만원/일 - 대신로: 3,783만원/일	- 동진로: 교통혼잡비용 5.7% 감소 - 대신로: 교통혼잡비용 3.9% 감소 - 총 4.8% 감소

본 사업에서 구축한 모빌리티 빅데이터 기반 Infra-free 교통 신호운영 분석 공용 솔루션의 기대효과는 다음과 같다. 첫째, 교통정보 수집을 위한 인프라 설치를 최소화할 수 있다. 둘째, 개별 차량의 궤적 데이터를 기반으로 기존 현장 조사 및 검지기로 산출 불가능한 핵심 지표를 도출할 수 있다. 셋째, 관리자는 효율적인 신호 운영 현황 모니터링과 신속한 대응이 가능하다. 넷째, 공공과 민간의 데이터 융합을 통해 가치 있는 신규 데이터를 생산할 수 있다. 다섯째, 교통신호 운영 개선에 한계를 겪고 있는 지자체를 지원하여 공공의 편익과 형평성을 증대시킬 수 있다.

PART 1-2
교통안전

교통 빅데이터 분석을 통한 버스전용차로 제도개선으로 국민 이동 편의 증진

한국도로공사

추진목적/배경

한국도로공사는 고속도로 교통정체 개선을 위해 노선신설과 정체개선 사업을 꾸준히 수행하고 있으며, 국민 중심의 서비스를 제공하기 위한 노력을 이어가고 있다.

경부선은 1994년 버스전용차로 최초 시행 후 2008년 관련 제도가 개선되었지만, 이후 교통량 증가로 구간 개선에 대한 논의가 지속적으로 이루어져 왔다. 또한 영동선의 경우 2018년 평창동계올림픽을 계기로 설치되었으나, 일반차로의 정체를 가중시켜 운영 필요성에 대한 검토가 필요했다.

경부선을 이용하는 차량은 23년 기준 연평균 472백만대이고 주말이나 휴일에 영동선을 이용하는 차량은 연평균 60백만대이다.

버스전용차로 운영으로 인해 최좌측차로가 버스전용으로 지정되면서 일반차량의 정체가 가중되고 있으며, 버스전용차로에 대한 불만 민원이 연간 5천여건 발생한 것으로 확인되었다.

이에 국민 불만 해결과 고속도로 기능의 회복을 위해 한국도로공사는 최근 7년간의 객관적인 교통데이터를 활용하여 버스전용차로 제도 운영의 적정성에 대해 분석하게 되었다.

분석 사전 준비

– 활용 데이터

1. 분석 인프라

데이터명	형태	내용	출처	기준년도	내·외부 데이터
VDS, AVC, DSRC 등의 교통데이터	xlsx	교통량, 속도, 차종 등	한국도로공사	2018 ~2024	내부
CCTV 영상데이터	JPG	버스전용차로 이용차량 비율	한국도로공사	2023	내부

고속도로에 설치된 지점·구간별 VDS(차량검지기), AVC(교통량 조사장비), DSRC(단거리전용통신) 등을 활용하여 통행속도와 교통량 데이터를 확보하였다. 또한 CCTV 및 교통량 조사를 통해 차량이용패턴 정보를 수집하였다.

분석과정

분석 프로세스 도식화

- 분석 환경

1. 분석 인프라 : 기관 내 PC 이용, 교통분석지원시스템(자체), 영업빅데이터 시스템(자체)
2. 분석 환경 : Excel 등

- 데이터 수집

사용 데이터 출처 : 기관 보유 데이터(VDS, AVC, DSRC 등)

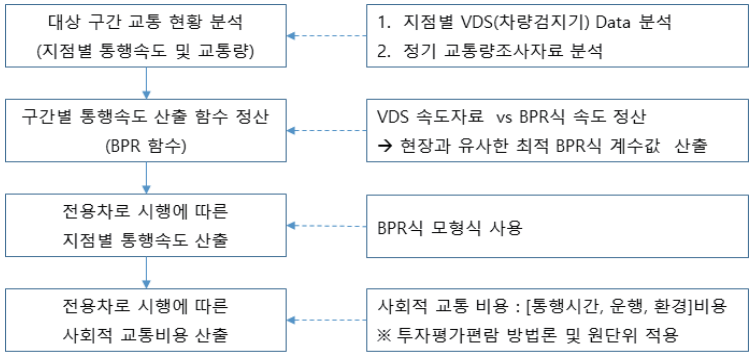
사용 데이터 형식 : xlsx

- 데이터 전처리

1. 결측값 및 오류 점검
   - VDS 자료 중 평균속도가 표출되지 않거나, 전후 구간과의 차이가 크게 발생하는 구간의 경우 오류로 판단
2. 이상치 점검
   - 구간별 속도편차가 50 이상 크게 나타나거나 일부 구간에 대해 나타나지 않는 경우 제외

-모델링

1. 데이터 처리 프로세스



가. BPR 모형식* 파라미터 정산

* BPR식 : 링크 통행시간이 자유통행시간을 기준으로, 용량대비 교통량 비율에 비례하여 증가한다는 이론에 근거한 함수

- 대상구간의 VDS 지점 검지자료 수집
- 평균속도 시공도 작성
- BPR 정산 (용량 초기속도, 알파(α), 베타(β))
  - VDS 교통량 입력시 VDS속도와 RMSE*가 최소가 되도록 BPR 모형 파라메터 정산 (구간별 RMSE 산정)

* RMSE(평균 제곱근 오차, Root Mean Squared Error) : 값의 차이가 얼마나 있는지 알려주는 척도로 사용, 값이 작게 나오는 것을 목표로 함

나. 대상구간 버스전용차로 시행유무에 따른 통행속도 산출

- 교통량 조사자료(2024년) 교통량 및 차종구성 비율 수집
- 버스전용차로 수요 산정
  - '버스전용차로 설치운영지침 연구용역(경찰청, 2020)' 방법 준용
- BPR식을 활용한 구간별 차로별(일반차로, 버스전용차로) 통행속도 산정

다. 대상구간 버스전용차로 시행유무에 따른 교통비용 산정

- 도로투자평가편람상의 비용항목 및 산출방법 사용
- 버스전용차로 시행 및 미시행시 통행시간비용, 운행비용, 환경비용 산정

2. 구간별 교통데이터 분석(운영기준 적정여부)

▶ 첨두 3시간 평균 교통량 중 일반차량 대비 버스교통량 비율

- 경부선: 9.0%로 기준치 5.6% 상회
- 영동선 : 5.8%로 기준치 7.9% 하회(기준미달)

버스전용차로 운영 기준이상 기준미달

□ **경부선 평일**(L=39.7km) · 운영기준 7.9% 이상, 설치기준 5.6%이상 (단위 : 버스/일반차량 교통량, 대/시)

구분	대상구간	차로수	2018년		2019년		2020년		2021년		2022년		2023년	
			부산	서울	부산	서울	부산	서울	부산	서울	부산	서울	부산	서울
운영구간 (39.7km)	양 재 금토JCT	5	873 / 3,792 (20.0%)	920 / 4,228 (21.8%)	822 / 4,307 (18.7%)	728 / 4,733 (15.2%)	138 / 5,119 (2.7%)	75 / 6,079 (1.2%)	95 / 5,826 (1.6%)	92 / 5,281 (1.7%)	796 / 4,912 (16.2%)	772 / 4,707 (16.4%)	895 / 4,789 (18.7%)	1,073 / 5,132 (20.9%)
	금토JCT 판교JCT	5	1,066 / 3,585 (29.7%)	950 / 3,959 (24.0%)	973 / 3,554 (27.4%)	1,255 / 3,476 (36.1%)	255 / 4,212 (6.4%)	178 / 4,424 (4.0%)	184 / 4,191 (4.4%)	153 / 3,922 (3.9%)	743 / 3,955 (18.8%)	689 / 3,821 (18.0%)	891 / 3,938 (22.6%)	1,039 / 3,811 (27.3%)
	판교JCT 신갈JCT	4	1,084 / 5,202 (20.8%)	1,232 / 5,119 (24.1%)	1,705 / 4,763 (35.8%)	1,540 / 4,957 (31.1%)	95 / 5,710 (1.7%)	69 / 5,941 (1.2%)	185 / 4,234 (4.4%)	230 / 6,057 (3.8%)	970 / 4,992 (19.4%)	900 / 5,216 (17.3%)	1,161 / 5,476 (21.2%)	1,001 / 5,936 (16.9%)
	신갈JCT 수원신갈	5	1,431 / 5,404 (26.5%)	734 / 4,786 (15.3%)	1,187 / 4,504 (26.4%)	839 / 4,510 (18.6%)	248 / 5,237 (4.7%)	160 / 5,612 (2.9%)	639 / 5,071 (12.6%)	509 / 5,562 (9.2%)	767 / 6,233 (12.3%)	702 / 5,400 (12.8%)	731 / 5,915 (12.4%)	664 / 5,427 (12.2%)
	수원신갈 기 흥	4	1,161 / 5,672 (20.5%)	1,150 / 6,169 (18.6%)	911 / 5,132 (17.8%)	931 / 5,042 (18.5%)	935 / 5,092 (18.5%)	914 / 5,213 (17.5%)	682 / 4,903 (13.9%)	563 / 5,083 (11.1%)	913 / 5,308 (17.1%)	742 / 5,340 (13.9%)	661 / 5,400 (12.1%)	902 / 4,900 (18.4%)
	기 흥 동탄JCT	4	851 / 4,823 (17.6%)	691 / 4,380 (15.8%)	466 / 4,134 (11.3%)	466 / 4,443 (10.5%)	464 / 4,404 (10.5%)	464 / 4,117 (11.3%)	540 / 4,652 (11.6%)	517 / 5,042 (10.3%)	718 / 5,114 (14.0%)	692 / 4,939 (14.0%)	737 / 5,417 (14.7%)	688 / 4,513 (15.2%)
	동탄JCT 오 산	4	517 / 4,075 (10.9%)	525 / 5,068 (10.4%)	685 / 4,726 (14.3%)	841 / 4,583 (18.4%)	257 / 4,907 (5.6%)	340 / 5,009 (6.7%)	396 / 5,434 (7.3%)	475 / 5,187 (9.2%)	581 / 5,224 (11.1%)	581 / 5,583 (10.4%)	637 / 5,582 (11.3%)	680 / 6,706 (10.1%)
미운영구간 (38.1km)	오 산 남사JCT	4	448 / 4,892 (9.2%)	362 / 5,114 (7.1%)	304 / 5,002 (6.1%)	371 / 4,727 (7.8%)	372 / 4,747 (7.8%)	306 / 5,022 (6.1%)	349 / 5,382 (6.5%)	385 / 5,116 (7.5%)	509 / 5,469 (9.7%)	645 / 5,461 (11.8%)	468 / 4,926 (9.5%)	556 / 4,947 (11.2%)
	남사JCT 안성JCT	4									571 / 5,229 (10.9%)	439 / 5,437 (8.2%)	536 / 5,700 (9.4%)	677 / 5,726 (11.8%)
구간	안성JCT 안 성	4	706 / 5,341 (13.2%)	576 / 4,121 (14.0%)	317 / 5,603 (5.7%)	344 / 5,393 (6.7%)	344 / 6,015 (5.7%)	319 / 5,825 (5.5%)	330 / 5,982 (5.5%)	301 / 5,936 (5.1%)	394 / 5,651 (7.0%)	427 / 6,233 (6.8%)	461 / 5,804 (7.9%)	417 / 5,770 (7.1%)
	안 성 북천안	4	770 / 5,041 (15.3%)	679 / 5,009 (13.6%)	630 / 5,201 (12.1%)	615 / 4,894 (12.6%)	242 / 5,234 (4.6%)	290 / 5,365 (5.4%)	311 / 5,796 (5.4%)	281 / 5,216 (5.4%)	340 / 4,984 (7.0%)	364 / 6,025 (6.0%)	466 / 5,221 (8.9%)	472 / 5,564 (8.5%)
	북천안 천 안	4	650 / 5,173 (12.6%)	583 / 4,574 (12.7%)	611 / 5,319 (11.5%)	653 / 5,284 (12.4%)	286 / 5,498 (5.2%)	315 / 5,511 (5.7%)	315 / 5,576 (5.6%)	299 / 5,467 (5.5%)	409 / 5,127 (8.0%)	338 / 5,593 (6.0%)	432 / 5,174 (8.3%)	401 / 5,637 (7.0%)

* '23년 경부선 오산IC~양재IC 버스교통량 비율(평균 16.8%) 기준 이상 (기준 7.9% 이상)

** '23년 경부선 천안IC~오산IC 버스교통량 비율(평균 9.0%) 기준 이상 (기준 5.6% 이상)

□ **영동선 주말**(L=26.9km) · 운영기준 7.9% 이상

버스전용차로 운영 기준이상 기준미달

검토구간	차로수	2018년		2019년		2020년		2021년		2022년		2023년	
		강릉	인천	강릉	인천	강릉	인천	강릉	인천	강릉	인천	강릉	인천
신갈JCT ~ 마 성	5	171 / 4,926 (3.5%)	274 / 4,783 (5.8%)	243 / 4,948 (4.9%)	281 / 4,983 (5.6%)	212 / 4,148 (5.1%)	153 / 4,556 (3.4%)	196 / 5,219 (3.8%)	207 / 5,593 (3.7%)	290 / 4,980 (5.8%)	297 / 6,007 (4.9%)	383 / 4,984 (7.7%)	288 / 5,281 (5.5%)
마 성 ~ 서용인JCT	5	358 / 4,703 (7.6%)	324 / 4,225 (7.7%)	340 / 4,389 (7.7%)	256 / 4,543 (5.6%)	122 / 4,461 (2.7%)	99 / 4,390 (2.3%)	187 / 5,037 (3.7%)	135 / 4,926 (2.7%)	317 / 5,034 (6.3%)	268 / 5,398 (5.0%)	270 / 4,594 (5.9%)	274 / 4,626 (5.9%)
서용인JCT ~ 용 인	5									265 / 5,176 (5.1%)	227 / 5,304 (4.3%)	310 / 4,552 (6.8%)	326 / 5,344 (6.1%)
용 인 ~ 양 지	4	398 / 5,065 (7.8%)	331 / 4,492 (7.4%)	173 / 3,630 (4.8%)	183 / 4,004 (4.5%)	104 / 3,936 (2.6%)	113 / 4,013 (2.8%)	220 / 4,507 (4.9%)	221 / 5,151 (4.3%)	233 / 4,689 (5.1%)	205 / 4,803 (4.3%)	324 / 5,021 (6.5%)	364 / 5,345 (6.8%)
양 지 ~ 덕 평	4	301 / 4,104 (7.3%)	275 / 3,636 (7.6%)	160 / 3,486 (4.6%)	202 / 3,903 (5.2%)	88 / 3,562 (2.5%)	92 / 3,688 (2.5%)	85 / 4,031 (2.1%)	106 / 3,807 (2.8%)	164 / 4,243 (3.9%)	202 / 3,918 (5.2%)	173 / 4,151 (4.2%)	216 / 4,101 (5.3%)
덕 평 ~ 호법JCT	4	276 / 3,707 (7.4%)	265 / 3,443 (7.7%)	139 / 3,629 (3.8%)	160 / 4,335 (3.7%)	94 / 3,665 (2.6%)	97 / 3,799 (2.6%)	94 / 4,199 (2.2%)	113 / 3,908 (2.9%)	157 / 4,099 (3.8%)	169 / 4,103 (4.1%)	173 / 4,152 (4.2%)	227 / 4,320 (5.3%)

* '23년 신갈JCT~호법JCT 버스교통량 비율(평균 5.8%) 기준 미달 (기준 7.9% 이상)

3. BPR 모형식의 통행시간 값을 활용한 구간별 통행속도 산정

- 경부선 버스전용차로 구간 연장 시, 영동선 버스전용차로 구간 폐지 시에 대한 평균 통행속도 분석

$$T = T_0 \times [1 + \alpha (V/C)^{\beta}]$$

T : 링크의 통행시간(시간)

T0 : 이상적인 상태에서의 링크 통행시간(시간)

V : 교통량(승용차)(대/시)

C : 용량(승용차)(대/시)

▶ BPR 모형식 파라미터 정산

- BPR함수식의 계수값(T_0 , α, β)을 정산하여 VDS 통행속도와의 오차 최소화

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (V_{m_i} - V_{o_i})^2}{n}}$$

여기서, n = 구간수

V_m = BPR 정산속도

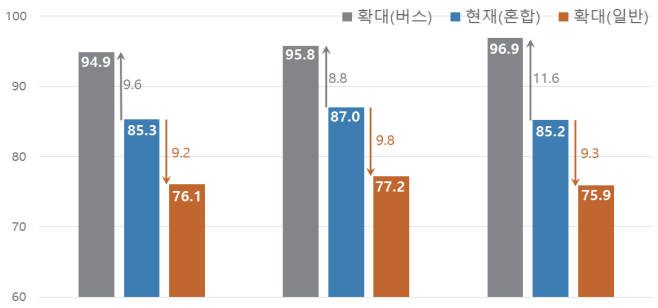
V_o = VDS 속도

▶ 통행속도 분석결과

- 경부선 평일 버스전용차로 확대 시

→ 버스 평균 통행속도 8.8~11.6km/h까지 향상

→ 일반차량 평균 평행속도 9.2~9.8km/h까지 감소



• 영동선 주말 버스전용차로 폐지 시

→ 버스 평균 통행속도 11.0km/h까지 감소

→ 일반차량 평균 평행속도 20.2km/h 증가하여 62.3km/h에서 82.5km/h 까지 상승

구분	버스		일반 폐지	폐지	
	버스	일반		버스	일반
인천	92.5km/hr (20.0분)	61.1km/hr (30.9분)	80.1km/hr (22.2분)	-7.6km/hr (1.8분)	23.8km/hr (-8.7분)
강릉	94.4km/hr (20.0분)	63.4km/hr (29.8분)	84.9km/hr (23.6분)	-14.3km/hr (3.6분)	16.7km/hr (-6.2분)
평균	93.5km/hr (20.2분)	62.3km/hr (30.4분)	82.5km/hr (22.9분)	-11.0km/hr (2.7분)	20.2km/hr (-7.4분)

4. 버스전용차로 이용차량 비율 분석

▶ CCTV 영상을 통한 주말기간 6개월의 상습 정체 시간 분석

• 토요일 : 마성-용인 강릉방향 분석

• 일요일 : 양지-덕평 인천방향 분석

→ 경부선은 승합차 비율이 버스대비 0.5배로 나타남

→ 영동선은 승합차 비율이 버스대비 1.8배로 측정되어 대중교통 활성화라는 버스전용차로 취지에 맞지 않는 것으로 나타남

(단위 : 대/시)

구 분		교통량	버스 (비율)	승합차 (비율)	합계	승합차 /버스
영동선	평균시간 교통량	4,332	93(35%)	170(65%)	263	1.8배
경부선	평균시간 교통량	5,146	265(64%)	152(36%)	417	0.5배

정책활용/기대효과

한국도로공사는 교통데이터를 면밀하게 분석하고 객관적인 결과를 바탕으로 국회, 경찰청, 국토교통부, 감사원, 버스연합 등 관계기관을 적극적으로 설득하였다. 이에 경찰청은 버스전용차로 운영과 관련한 관계기관 협의체를 개최하고 면밀한 검토를 시행했다.

그 결과, 「고속도로 버스전용차로 시행 고시」가 개정(2024. 06. 01.) 되었다.

이에 따라 경부선 버스전용차로는 16년 만에 평일 구간을 확대하여, 기존 오산에서 양재까지 운영중인 구간(39.7km)을 안성에서 양재까지 연장(58.1km) 운영하게 되었다. 반면 영동선의 버스전용차로는 7년 만에 폐지되었다.

경부선 버스전용차로 확대로 대중교통 이용자의 평일 출퇴근 시간이 평균 33분 단축되었으며, 시점부 교통사고가 월평균 13건에서 4건으로 69% 감소하였다.

또한 영동선의 버스전용차로 폐지로 인한 사회적 편익을 분석한 결과 연간 323억 원의 절감 효과가 발생한 것으로 나타났다. 신갈~호법구간의 통행속도는 시속 15km 증가하였으며, 월평균 교통사고 발생 건수는 16건에서 3건으로 81% 감소하였다. 최대 정체길이는 13km에서 5km로 8km 줄었으며 정체 지속시간도 11시간에서 4시간으로 7시간 단축되었다.

주요 언론은 경부선 및 영동선 버스전용차로 개선에 대한 높은 관심을 보이며 긍정적인 효과를 기대한다는 내용의 보도를 150회 이상 내보냈다.

버스전용차로 개선을 위해 빅데이터 분석을 수행하고 관계기관을 설득하는 과정에서 어려움도 있었지만, 국민 편익 증진을 목표로 업무를 추진한 결과 버스전용차로 개선이라는 중요한 과제를 해결할 수 있었다.

PART 1-3
교통안전

선박운항데이터를 활용한 GIS기반의 교통혼잡도 예측 모델개발

해양교통 혼잡 모델 : 해양교통량 예측 모델로 해양사고 막는다

한국해양교통안전공단

추진목적/배경

우리나라 관할 해역은 약 433,000KM^{2}로 국토 면적의 4.3배에 달하며, 여객 및 화물 운송, 어업, 레저 등 다양한 활동이 이루어지고 있다. 그러나 기기 손상 사고를 제외한 해양사고의 치사율은 9.8%로, 도로교통사고보다 약 7배 높아 다른 교통수단에 비해 더욱 철저한 안전관리가 요구된다.

도로교통에서는 교통 혼잡 구간, 사고 다발 지점, 낙석·결빙 위험 지역 등 다양한 안전 정보를 여러 수단을 통해 제공하여 운전자의 안전 운전을 유도하고 있다. 반면, 해양교통에서는 현재 몇 척의 선박이 얼마나 운항하고 있는지조차 정확히 파악할 수 없는 실정이었다.

이를 해결하기 위해 한국해양교통안전공단은 각 기관에서 개별적으로 관리하던 선박 운항 데이터를 통합·분석하여 선박 운항 특성과 해역 점유율을 고려한 해양교통량 예측 모델 개발을 추진하였다. 특히, 각기 다른 선박위치발신장치 데이터를 통합·분석하여 해양교통량을 예측한 사례는 국내에서 처음이다.

분석 사전 준비

– 활용 데이터

1. 분석 인프라

데이터명	형태	내용	출처	기준년도	내·외부 데이터
격자 정보	SHP	격자 ID, 위치, 수심, 해상 면적 등	해양수산부	2024	외부
AIS 정보	CSV	선박 식별 ID, 수신시각, 선박위치 등	해양수산부	2019~2024	외부
V-PASS 정보	CSV	선박 식별 ID, 수신시각, 선박위치 등	해양경찰청	2019~2024	외부
선박 정적 정보	CSV	선박 ID, 선종 코드, 선박길이, 선박톤수 등	한국해양교 안전공단	2019~2024	내부
해양기상·환경	SHP CSV	파고, 풍속, 수심 등	기상청, 해양수산부, 국립해양조사원	2019~2024	외부

활용된 데이터는 해역을 일정한 크기로 분할한 격자 정보, 선박 톤수와 선종 에 따라 다르게 수집되는 선박위치발신장치(AIS*, V-PASS**) 정보, 선박의 크기와 선종 등을 매칭 할 수 있는 선박 정적 정보와 해양기상·환경 정보이다.

* AIS(선박자동식별장치) : 선박의 명세, 위치, 속력, 침로 등의 정보를 송수신하는 시스템이며, 해상 안전을 강화하기 위해 국제적으로 요구되는 장비

** V-PASS : 어선의 위치를 자동으로 발신하는 무선설비 장치

선박위치발신장치 데이터는 해양수산부, 해양경찰청 등 데이터 보유 기관과 협의하여 TCP* 방식으로 실시간 연계하고 있으며, 풍속, 파고 등 해양기상 데이터는 국가기후데이터센터와 협의하여 SFTP** 방식으로 연계하고 있다. 또한, 선박제원 등 내부 데이터는 DB연결로 직접 연계하고, 해양격자 등 환경 데이터는 해양수산빅데이터플랫폼과 국립해양조사원의 개방海에서 제공하는 데이터를 활용하였다.

* TCP (Transmission Control Protocol) : 인터넷을 통해 데이터를 신뢰성 있게 전송하기 위한 프로토콜이며, 연결 지향형 서비스로 주로 실시간 통신이나 파일 전송에 사용

** SFTP (Secure File Transfer Protocol) : 파일 전송 프로토콜, 데이터 전송 중 암호화로 보안 강화

해양은 도로와 달리 정해진 항로가 없어 선박의 이동 방향이 자유롭고, 수심 및 조석 간만의 차에 따라 선박이 이동할 수 있는 공간이 제한되는 특성이 있다. 따라서 해상 교통에 대한 개념 정립이 선행되어야 한다. 또한 기상변화로 인해 선박 운항이 통제될 수 있으므로, 기상 여건이 해상교통량에 미치는 영향을 충분히 고려해야 한다. 이에 따라 해양교통 전문가 자문을 통해 해상교통량 예측에 활용된 변수의 개념을 명확히 정리하고, 외부 영향 인자 및 해상 교통혼잡도 산정 방법론 등을 면밀히 분석하였다.

아울러, 대용량 원시 데이터의 전처리 방안과 개인정보 보호 등에 대한 협의를 진행하였다. 공단 빅데이터 플랫폼에서 원시 데이터를 전처리하고, 개별 선박 위치를 식별할 수 없도록 격자별 집계 데이터를 분석에 활용하여 업무 효율을 높였다.

분석과정

– 분석 환경

1. 분석 인프라 : 한국해양교통안전공단 內 빅데이터 플랫폼 이용
   - GPU학습서버(2식) : CPU 8코어*2, 메모리 128GB*2
   - Storage(2식) : 저장공간 100TB*2
   - 빅데이터 시스템(17식) : CPU 8코어*14, 12코어*6 메모리 128GB*5, 64GB*12
2. 분석 환경 : python

– 데이터 수집

1. 사용 데이터 출처 : 기관 자체 생산 데이터 및 외부기관 데이터 (해양수산부, 해양경찰청, 국립해양조사원, 기상청)
2. 사용 데이터 형식 : shp, csv

– 데이터 전처리

1. 격자 데이터 전처리·가공
   - 해양수산부 격자 4단계* 기준 예측 영역 정의

     * 해양공간을 1분 30초 단위로 나눈 격자로 가로 약2.3km, 세로 약2.8km 크기
   - 전자해도 상 해안선을 기준으로 육·해상 구분
   - 기본수준면(Datum Level) 0m, 갯벌 영역 등을 고려하여 해상면적 산출
   - 해상면적 5%이하는 Masking 처리 후 최종 예측 격자 확정
2. 선박 데이터 전처리·가공
   - 선박위치발신장치 이중 설치 선박 중복 제거, 이상치 처리
   - 격자별·시간별 선박 교통량 집계
3. 해양환경 데이터 전처리·가공
   - 기상청 파랑예측 수치모델(RWW3)의 72시간 예측 데이터 활용
   - 유효, 비유효 값 처리 및 최종 학습, 입력 데이터 생성

– 모델링

1. 탐색적 데이터 분석(EDA)

▶ 격자 데이터 특성 분석
- 전체 격자(레벨 4 기준) 중 예측 격자 수 277,539개(67.8%)
- 격자 내 육지가 포함되지 않은 격자(넓은 해역) 97%, 육지가 포함된 격자(좁은 해역) 3% 차지
- 혼잡도 산정 계수 정의(넓은 해역 = 6L, 1.6L / 좁은 해역 = 3L, 2L) 및 적용

  * L : 선박 길이

▶ 선박 정적 데이터 특성 분석
- 우리나라 등록선박은 약 10만여척이며, 일별 운항 선박은 약 1.5만~4만척
- 평균 선박 길이는 AIS 장착 선박 92.0m, V-PASS 장착 선박 9.4m

▶ 선박 동적 데이터 특성 분석

- 학습에 사용되는 총 데이터 수는 약 8.7억만건이며, 그 중 일일 선박 운항데이터는 약 50~70만건
- 평균 선박길이 43.8m, 평균 선박 속도 8.1knot(4m/s)
- 춘계(5~6월)와 추계(9~10월)에 교통량이 높고 동계(1~2월)에 상대적으로 낮은 경향을 보이고, 고파도 강풍 등의 기상 여건이 교통혼잡에 높은 영향을 미침
- 대체로 주말/휴일(명절 등)에 교통량이 감소하는 경향을 보임
- 일별 변화 경향에서도 기상 여건(비, 안개 등)이 교통혼잡도에 큰 영향을 미치는 것으로 판단됨

▶ 국내 어항의 일 교통량 변화(+조위 및 파랑) 특성 분석

- 국가어항을 포함한 격자로 일반선보다 V-PASS를 장착한 어선이 지배적
- 대포항은 동해에 위치하여 조위에 따른 교통량 영향은 미비
- 새벽 출항 후 주간 조업 후 입항함에 따라 야간 교통량 감소
- 태풍, 풍랑특보로 파고가 높을 때에는 출항통제 영향으로 교통량 거의 없음

▶ 협수로 해역의 일 교통량 변화(+조위 및 파랑) 특성 분석

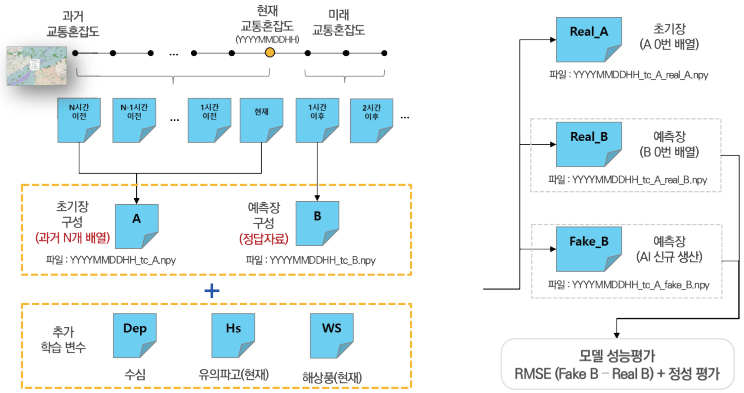
- AIS를 장착한 선박 보다 V-PASS를 장착한 어선의 비율이 높음
- 주간 교통량이 높고, 파고가 높아지는 시기 교통량은 거의 없음
- 조차가 약 6m에 달하는 해역으로 주간 저조 시 교통량이 일시적으로 감소하여 양봉 형태의 교통량 변화를 보임

2. 사용한 분석 모델

- 시계열에 강한 LSTM 모델*

  * LSTM : 순차적 데이터에서 장기 의존성을 학습할 수 있도록 설계된 순환 신경망(RNN) 구조로, 중요한 정보를 기억하며 불필요한 정보를 잊는 기능을 제공함

- 이미지 기반 학습에 강한 GAN 모델을 비교·분석하여, 더욱 정교하고 효율적으로 예측 결과를 생성하는 ddGAN 모델 최종 선정



- 검증 및 고도화

- 해상교통혼잡도 예측 모델 파라미터 조정
- 학습시간, RMSE, MSE, MAE, MAPE 등의 지표를 활용한 통계 검증 및 모델 평가를 통해 최적 모델 선정

  * RMSE(Root Mean Squared Error) : 예측값과 실제값의 차이를 제곱하여 평균한 후 제곱근을 취한 값

  * MSE(Mean Squared Error) : 예측값과 실제값의 차이를 제곱해 평균한 값

  * MAE(Mean Absolute Error) : 예측값과 실제값의 차이의 절대값을 평균한 값

  * MAPE(Mean Absolute Percentage Error) : 예측 오차를 실제값에 대한 비율로 나타낸 값

- 결과 구현

- 실시간 데이터를 활용하여 기준일부터 72시간까지의 해상교통량을 격자별 1시간 단위로 예측 → 예측결과를 4단계로 구분하여 시각화

정책활용/기대효과

한국해양교통안전공단은 개발된 해상교통혼잡 예측 모델을 해양교통안전정보시스템(MTIS)에 적용하여 '24년 1월부터 대국민과 유관기관에 제공하고 있다. 지도 상에서 교통혼잡 예측 결과를 4단계로 구분하여 시각화함으로써 혼잡해역을 한눈에 파악할 수 있으며, 예측 결과는 엑셀 파일로 다운로드 가능하다. 이를 통해 선박 운항자와 이용자는 혼잡 예상 해역을 우회하거나, 보다 안전한 위치를 선정하여 낚시 및 레저 활동을 계획하는 등 안전항해 계획 수립에 활용할 수 있다. 또한, 해양경찰 등 긴급 선박의 순찰경로 설정과 국가 해상교통망 관리 등 정책 의사결정을 지원하여 해양 안전관리의 핵심 기반이 될 것으로 기대한다.

공단은 '24년 7월부터는 교통혼잡 예측 결과를 선박모니터링시스템(VMS*)과 연계하여 여객선 항로 내 교통 혼잡도가 높은 구간의 정보를 사전에 제공하는 등 여객선의 안전 운항 관리에도 본 모델을 활용하고 있다.

* VMS : 선박의 실시간 위치, 여객선 안전 운항 상태 등을 모니터링 하는 시스템

더불어, 선박 크기와 종류 등을 고려한 개별 선박 이동 예측을 통해 해역별 혼잡 정확도를 보완하고, 교통정보 서비스의 품질을 지속적으로 향상해 나갈 계획이다.

PART 1-4
교통안전

장거리 통학생을 위한 통학버스노선 개설

의정부형 통학버스, 학생들의 통학과 안전을 책임지다

경기도 의정부시

추진목적/배경

의정부시의 동부와 서부 간 학군 불균형 문제로 인해 송산권역 학생들은 흥선권역의 7개 고등학교로 통학하는 과정에서 긴 통학 시간과 복잡한 환승 등 열악한 통학 환경에 직면해 있다. 송산권역의 장거리 통학률은 최대 36.7%에 달하며, 1,000명 이상의 중·고교 학생이 단일 학군으로 배치된 상황에서 학군 분리 및 조정은 단기간 내 해결이 불가능한 실정이다.

특히 고산지구 신도시 개발에 따른 인구 유입 증가로 장거리 통학 학생 수가 지속적으로 늘어날 것으로 예상되며, 이는 학생들의 안전과 학습권 보장에 중대한 영향을 미칠 것으로 우려된다.

현재 대중교통 이용 시 환승의 불편함과 과도한 통학 시간으로 인해 학생들의 학업 집중도와 안전이 저해되고 있다. 의정부시는 이러한 문제를 해결하기 위해 송산권역과 흥선권역 간 교통 현황과 학생 거주지 데이터를 기반으로 효율적이고 안전한 통학 체계를 설계하였으며, 이는 단순히 교통 여건을 개선하는 것을 넘어 학군 불균형 문제 해결의 출발점이 될 것이다.

분석 사전 준비

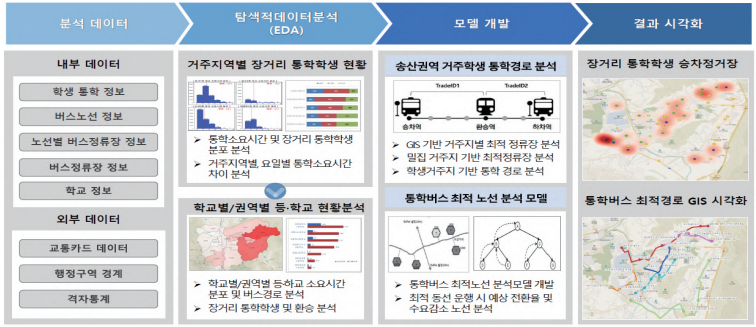
– 활용 데이터

데이터명	형태	내용	출처	기준년도	내·외부 데이터
버스노선 정보	csv	노선번호, ID, 거리 등	의정부시	2024	내부
노선별 버스정류장 정보	csv	순번, 좌표, 거리 등	의정부시	2024	내부
버스정류장 정보	csv	ID, 명칭, 번호, 좌표 등	의정부시	2024	내부
관내 학교 정보	csv	관내 학교 상세정보	의정부시	2024	내부
관내 학생 통학 정보	csv	학생 거주지, 통학방법, 버스 이용 희망 등	의정부 교육지원청	2024	내부
관내 교통카드 O/D 데이터	csv	관내 출발지/도착지 이동 데이터	한국교통 안전공단	2024	외부
학생 거주 격자 정보	shp	50m 격자별 학생 거주 인구 데이터	국토지리 정보원	2024	외부
의정부시 행정동 경계	shp	행정동 공간정보	국토교통부	2024	외부

분석과제 추진 과정에서 필요한 데이터를 확보하기 위해 의정부교육지원청과 한국교통안전공단 등 관련 기관과 협력하였다. 이 과정에서 데이터 안전성 확보를 위해 개인 식별이 가능한 모든 정보를 비식별화하고 가명처리 방식을 적용하였다.

의정부교육지원청을 통해 학생 설문조사를 실시하여 기존 통학 방식과 통학버스 수요를 파악하고, 학생들이 거주지에서 통학하는 거리가 얼마나 되는지 확인할 수 있었다. 또한 한국교통안전공단의 교통카드 O/D 데이터를 활용하여 관내 학생들의 통학 환경을 정량적으로 분석할 수 있는 기반을 마련하였다.

분석과정



(분석 프로세스 도식화)

– 분석 환경

1. 분석 인프라
   - CPU : i7-13700HX 프로세서
   - 메모리 : 32GB
   - 데이터 저장용량 : 512GB SSD
   - GPU : RTX 4060
2. 분석 환경 : python 3.11, anaconda 23.7.4, Jupyter notebook 6.5.4,

※ 주요 활용 라이브러리: pandas, geopandas, networkx, heapq, sklearn

– 데이터 수집

1. 사용 데이터 출처 : 기관 자체 생성 데이터 및 외부기관 보유 데이터
2. 사용 데이터 형식 : csv, shp

– 데이터 전처리

1. 분석 대상 필터링
   - 사용자자구분코드(03: 청소년)를 활용하여 학생 이용자 데이터 필터링

• 학생 사용자의 교통카드 데이터를 평일 기준 월 4회 이상 사용한 사례로 필터링하여 통학 패턴을 추출

2. 결측치, 오류, 이상치 점검

• 승·하차정류장, 승·하차시각 결측치 제거

• 승차정류장과 하차정류장이 동일한 오류 데이터 제거

• 승차 태그 시간보다 하차 태그 시간이 빠른 데이터 제거

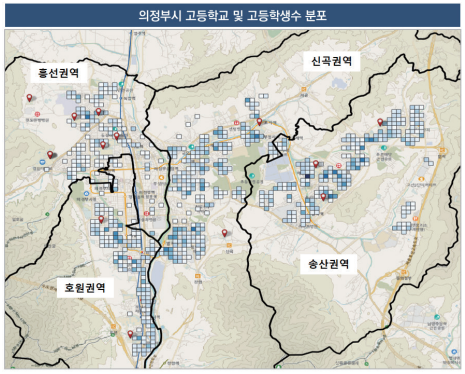
• 승차 시간과 하차 시간의 차이가 비정상적으로 작거나 큰 데이터 제거

– 모델링

1. 탐색적 데이터 분석(EDA)

▶ 의정부시 고등학교 및 고등학생수 분포

• 서부(흥선권역) 고등학교는 8개교, 고등학생은 1,439명인 반면, 동부(신곡권역, 송산권역) 고등학교는 6개교이며, 고등학생은 6,527명으로 동부권역에서 서부권역으로 등교하는 학생들이 다수 발생함을 확인

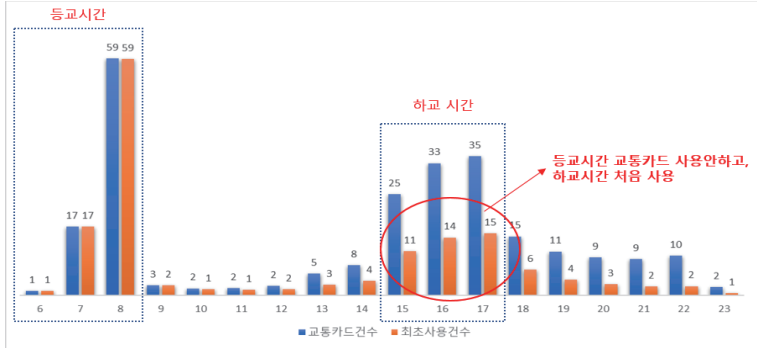


권역	행정동	고등학교 동별	고등학교 권역	고등학생 수 동별	고등학생 수 권역
흥선권역	가능동	3	8	485	1,439
	녹양동	2		454	
	흥선동	3		298	
	의정부1동	0		202	
호원권역	의정부2동	1	2	408	1,959
	호원1동	1		747	
	호원2동	0		804	
신곡권역	신곡1동	1	2	918	3,304
	신곡2동	1		1,319	
	자금동	0		568	
	장암동	0		499	
송산권역	송산1동	1	4	801	3,223
	송산2동	2		1,059	
	송산3동	1		1,363	
합계		16		9,925	

▶ 통학환경 분석

• 하교시간에만 대중교통을 이용하는 학생이 다수 존재

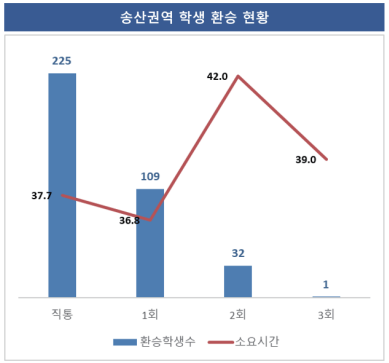
→ 자가용 및 사설버스 등 대체교통수단 이용 확인



시간대별 교통카드 사용건수 분석 결과

▶ 송산권역 학생 환승 현황분석

• 송산권역 학생들의 교통수단을 확인한 결과, 환승 학생은 약 39%이며, 마을버스에서 일반버스로 환승하는 학생들의 평균 통학 소요시간이 51.5분으로 가장 오래 걸림



※ 환승 횟수별 통학학생수 중복 산출

• 실제 송산3동 거주 학생의 통학경로는 일반버스를 이용하여 61분이 소요되며, 송산1동 거주학생은 장거리 도보 이동 후 경전철 또는 마을버스를 이용하여 통학함.

▶ 학생설문조사 기반 희망 이용률 및 기대수요 산출

• 설문응답률이 낮은 3개 학교 학생 수를 응답률 상위 4개 학교의 평균 학생수로 환산 후 산출

→ 7개 학교 학생 중 송산권역에 거주하고 통학버스를 희망하는 학생수 약 469명으로 추계

2. 사용한 분석 모델

• 통학버스 최적경로 탐색은 가능한 많은 학생을 태우면서 최소한의 정류장에 정차하는 경로 최적화 수행함

• 승차정류장 구역 분류를 위해 K-means 클러스터링* 기법 사용

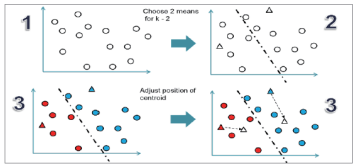
*데이터를 비슷한 특성을 가지는 여러 개의 클러스터로 그룹화하는 비지도 학습 알고리즘

• 통학버스 경로 탐색을 위해 백트레킹(backtracking)* 기법 사용

*가능한 모든 후보해를 탐색하면서 조건을 만족하는 해를 찾는 알고리즘

승차정류장 구역 분류

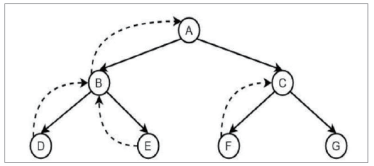
[K-means 클러스터링]



- K-means 클러스터링은 비지도 학습 알고리즘 중 하나로 데이터를 비슷한 특성을 가지는 여러 개의 클러스터로 그룹화하는 방법
- 거리 기반 k-means 클러스터링은 클러스터의 중심과 데이터 포인트 간의 거리를 기반으로 클러스터를 형성하는 방식으로 작동
  데이터 포인트들을 여러 클러스터로 그룹화할 때 각 클러스터의 중심을 선택하고, 각 데이터 포인트를 가장 가까운 중심에 할당하며 반복적으로 클러스터를 업데이트

통학버스 경로 탐색 알고리즘

[백트래킹(backtracking) 알고리즘]



$$\text{Maximize} \quad \sum_i s_i \cdot x_i$$

$\sum_i x_i \leq$ 최대 정유 정류장 수

$\sum_i s_i \cdot x_i \leq$ 서를 버스 최대 탑승 인원

- 백트래킹 알고리즘은 특정 문제의 해결을 위해, 가능한 모든 후보해를 탐색하면서 조건을 만족하는 해를 찾는 알고리즘 기법
- 문제의 가능한 해들을 나타내는 상태 공간 트리를 구성하며, 유효하지 않은 경로는 더 이상 탐색하지 않도록 가지치기 수행

– 검증 및 고도화

1. 모델 및 데이터 분석 결과 검증

• 교통카드 O/D 데이터와 모델 결과를 비교하여 실제 통학 경로와의 적합성(일치율)을 측정

• 결과 시각화를 통해 클러스터링 및 경로 탐색 결과의 신뢰성 제고

2. 고도화

• 고등학교에 진학하는 학생들의 시계열 데이터를 분석하여 향후 몇 년간의 학생 분포와 이동 수요를 예측하고 이를 모델에 반영

• 기존 교통카드 O/D 데이터와 학생 거주지 전수데이터를 결합하여 학생들의 실제 거주 위치와 주요 통학 목적지를 모델에 반영

- 결과 구현

- 의정부형 통학버스 6개 버스노선 개발
  - 2개의 차고지에서 출발하는 경로분석 알고리즘을 통해 최적의 통학버스 노선 도출

*제외된 부분은 직행버스 승차정거장으로 통학버스 대체효과가 낮아 제외함.

버스노선	학생수 (거주지기준)
1	44
2	86
3	60
4	91
5	58
6	42
합계	381

- 의정부형 통학버스 최종노선 및 예상 승차인원 결과 도출

노선1 : 송산1동 노선 — 예상 승차인원수 : 44 ~ 58명

정류소 ID	정류장 이름	OD 데이터	500m 학생수	700m 학생수
08312	정음마을고산2단지.고산종합사회복지관	3	6	10
08076	고산대광로제비앙.고산센트레빌아파트	12	22	31
08615	고산수자인디에스티지.아트포레	16	16	17
합계		31	44	58

※ 고산지구 신도시 입주로 승차학생 증가 예정

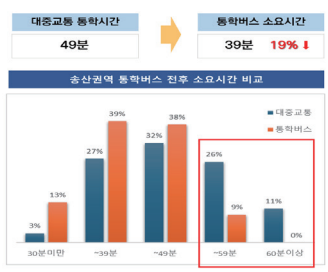
노선2 : 송산2동 노선(1) — 예상 승차인원수 : 86 ~ 104명

정류소 ID	정류장 이름	OD 데이터	500m 학생수	700m 학생수
08290	산들마을4단지	6	31	31
08272	산들마을1단지남문	1	10	16
08225	송산주공7단지	0	30	42
08246	송산주공1.9단지	8	15	15
합계		15	86	104

정책활용/기대효과

의정부시는 데이터 기반의 통학버스 노선을 도출함으로써 장거리 통학 학생들의 안전한 통학 환경을 조성하고 교통 여건을 개선하여 시민 불편을 해소하였다.

통학버스 운행을 통해 얻을 수 있는 기대효과는 다음과 같다. 첫째, 통학시간이 50분 이상 소요되는 학생 비율이 기존 37%에서 9%로 감소할 것이다. 둘째, 기존에 대중교통을 이용하던 학생의 통학시간이 통학버스 이용 시 평균 49분에서 39분으로 약 19% 단축되고, 최대 59%까지 감소하여 학생들의 학업 및 생활 여건이 획기적으로 개선될 것이다.



의정부형 통학버스는 '24년 8월 12일 개통식을 가진 후 13일부터 정식 운행하여 시민들로부터 큰 호응을 얻었다.

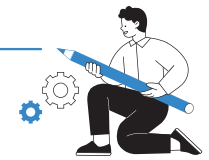
데이터 기반의 통학버스 운행은 교통 여건 개선뿐만 아니라 학군 불균형 문제 완화, 학생 안전성 확보, 시민 만족도 향상 등 다방면에서 긍정적인 효과를 창출할 것이다.

의정부시는 향후 학생 통학 수요와 교통 흐름을 지속적으로 분석하여 새로운 노선을 확대하고, 분석 결과를 더 나은 서비스 제공을 위한 기초 자료로 활용할 계획이다. 또한 지속 가능한 교통 정책을 추진하며 데이터 기반의 정책을 통해 시민들의 삶의 질을 향상시킬 수 있도록 노력할 것이다.

공공
행정

PART 2

공공행정



PART 2-1
공공행정

빅데이터 분석으로 국민 생활을 개선하고 업무 효율을 올리다

국민연금공단

추진목적/배경

우리 모두 행복한 사회를 희망하지만 사회적 약자의 빈곤과 그로 인한 비극은 계속해서 되풀이되고 있다. 대한민국의 노인 빈곤·자살률은 OECD 국가 중 1위이며 독거노인의 70%가 빈곤층에 속한다는 점은 심각한 사회 문제로 자리잡고 있다.

저소득층과 빈곤층 지원을 위한 우리나라의 대표적인 제도로는 기초수급자 생계급여 제도와 농어업인의 안정적 노후 소득 보장을 위한 연금보험료 국고지원 제도가 있다.

질병이나 부상으로 치료 또는 요양이 필요한 기초수급자가 생계지원을 받기 위해서는 의학평가와 근로능력평가를 통해 '근로능력 없음' 판정을 받아야 한다. 그러나 기존 근로능력평가 제도는 호전 가능성이 없는 대상자에게도 반복적인 평가를 요구하여 실익 없는 평가로 국민 불편을 초래한다는 비판을 받아왔다. 한편 국가는 소득이 낮은 농어업인에게 연금보험료의 50%를 국고로 지원하고 있으나, 공단의 지속적인 안내에도 불구하고 최근 5년간 신청률은 25%에 그쳐 저조한 실정이다.

이에 공단은 국민권익 향상과 불합리한 제도 개선을 위한 과제를 발굴하고, 빅데이터를 활용한 사회 문제 해결을 목표로 분석과제를 추진하였다.

분석 사전 준비

– 활용 데이터

1. 기초수급자 근로능력 평가 제도 개선을 위한 빅데이터 분석

데이터명	형태	내용	출처	기준 년도	내·외부 데이터
근로능력 평가결과	CSV	고착질환 여부, 의학적평가단계, 호전가능성, 평가결과 등	국민연금공단	2012~ 2021	내부

2. 농어업인 연금보험료 국가지원 미신청 해소를 위한 빅데이터 분석

데이터명	형태	내용	출처	기준 년도	내·외부 데이터
농어업 경영체 등록자	CSV	성별, 나이, 거주지 등	국민연금공단	1995~ 2021	내부
상담 이력	CSV	상담일자, 상담내용 등	국민연금공단	1995~ 2021	내부
가입 이력	CSV	가입기간, 가입종별, 보험료 납부액 등	국민연금공단	1995~ 2021	내부
반추납 이력	CSV	반납 이력, 추납 이력 등	국민연금공단	1995~ 2021	내부
국고지원 납부이력	CSV	국고지원 수납금액 및 기간	국민연금공단	1995~ 2021	내부
업종변경 이력	CSV	사업장 정보, 가입 기간 등	국민연금공단	1995~ 2021	내부

분석 과제별로 분석 대상을 정의하고, 분석 대상의 특성을 파악할 수 있는 데이터를 생성하기 위해 업무 시스템에서 활용되는 데이터와 정보 시스템(DW), 빅데이터 시스템에 축적된 데이터를 수집하였다. 여러 시스템에 분산되어 있는 업무 중심 데이터를 분석 중심의 데이터로 가공한 후, 최종 데이터 마트를 CSV 파일 형태로 분석 환경에 적재하였다.

분석 환경

1. 분석 인프라 : 기관 내 빅데이터 시스템 이용
2. 분석 환경 : R, Python, Presto 등

– 데이터 수집

1. 사용 데이터 출처 : 기관 자체 생성 데이터
2. 사용 데이터 형식 : csv

– 데이터 전처리

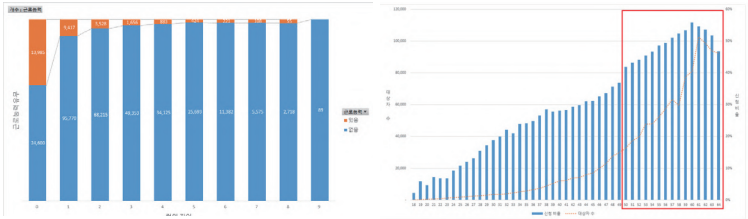
1. 결측값 및 오류 점검
   - 소득유형 코드 등의 결측값에 새로운 코드값 부여
   - 현재 삭제된 지역 코드(행정동·법정동) 삭제
2. 이상치 점검
   - 과세 소득 10만 원 미만자 결측 처리 및 중앙값 대체
   - 상담 횟수, 업종 변경 횟수 등에 대한 상단 범주화
3. 파생변수 생성
   - 최종 근로능력 평가결과를 목표변수로 정의(근로능력 '있음' 또는 '없음')
   - 최종 평가 결과를 제외한 '근로능력 없음' 횟수를 근로능력 없음 연속 횟수로 정의
   - 평가 이력의 길이를 평가 횟수로 정의
   - 근로능력 평가자의 최종 평가시 연령을 현재 연령으로 정의
   - 근로능력 평가자의 최초 평가시 연령을 최초 평가 시 연령으로 정의

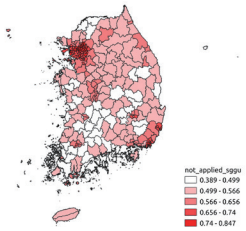
– 시각화

1. 활용 도구
   - R, Python, Excel, QGIS 등

2. 주요 시각화 분석 내용

- 근로능력 없음 연속 횟수와 최종 평가 결과 간 연관성 파악
- 연령별 보험료 지원 신청률 추이 시각화
- 지역별 보험료 지원 미신청률 분포 시각화





– 모델링

1. 사용한 분석 모델

▶ 기초수급자 근로능력 평가 제도 개선을 위한 빅데이터 분석

- 평가자의 일반적 특성과 평가 결과(근로능력 유무) 간 영향 요인 파악
  → 빈도 분석 및 교차 분석 실시
- 연구 가설의 채택 여부를 결정하기 위한 통계적 가설 검증
- 유사 집단별 효율적 유효 기간 연장 가능 대상 선정 및 규모 파악 (집단 수: 12,000개) → 시뮬레이션 방식 활용

▶ 농어업인 연금보험료 국가 지원 미신청 해소를 위한 빅데이터 분석

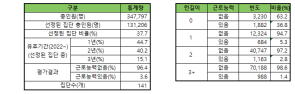
- 농어업인 국고지원 미신청 사유 파악 → 분석 관점별 영향 요인 분석
- 신청 가능성 예측 모형 개발(신청 영향 요인 반영) → RandomForest*, XGBoost** 모델 이용

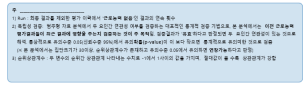
* RandomForest : 여러 개의 결정 트리를 결합하여 예측 정확도를 높이고 과적합을 줄이는 앙상블 학습 기법

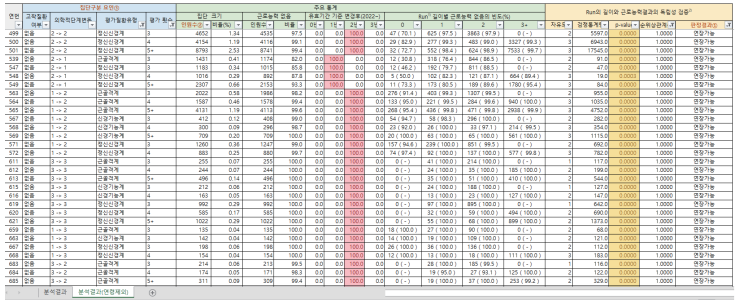
** XGBoost : Gradient Boosting 기법을 기반으로 한 효율적이고 강력한 머신러닝 모델로, 예측 성능을 극대화하고 과적합을 방지하는데 뛰어난 성능을 보임

2. 유효 기간 연장 대상 규모 산정을 위한 시뮬레이션 도구

- 엑셀 기반 도구 개발
- 업무 담당자의 용이한 시뮬레이션 수행 지원
- 합리적 평가 제도 개선을 위한 정책 수립 지원







3. 예측모형 성능평가

▶ 농어업인 연금보험료 국고지원 신청 가능성이 높은 대상자 예측모형

- 비교 대상 알고리즘
  → RandomForest, XGBoost
- 예측모형(분류) 성능 평가 지표
  → Accuracy*(정확도), AUC**(Area Under Curve)

* Accuracy : 전체 데이터 중에서 모델이 정확하게 예측한 비율을 나타내는 성능 지표

** AUC : 이진 분류 모델의 성능 평가 지표로, ROC 곡선 아래 면적을 나타내며 값이 1에 가까울수록 모델 성능이 우수함을 의미함

	Accuracy	AUC
RandomForest	0.9167	0.9170
XGboost	0.9183	0.9188

- 알고리즘 간 성능 비교 결과 → 정확도와 AUC측면에서 유의미한 차이 없음
- RandomForest 선정 사유 → 공단 분석 환경 고려
- 모형 최적화 → 변수 선택 및 변수 변환 등을 통한 조정

– 모델 검증 및 고도화

- 농어업인 연금보험료 국고지원 신청 가능성이 높은 대상자 예측모형 효과성 검증

▶ 예측모형 기반 안내 대상 선정 및 효과 분석

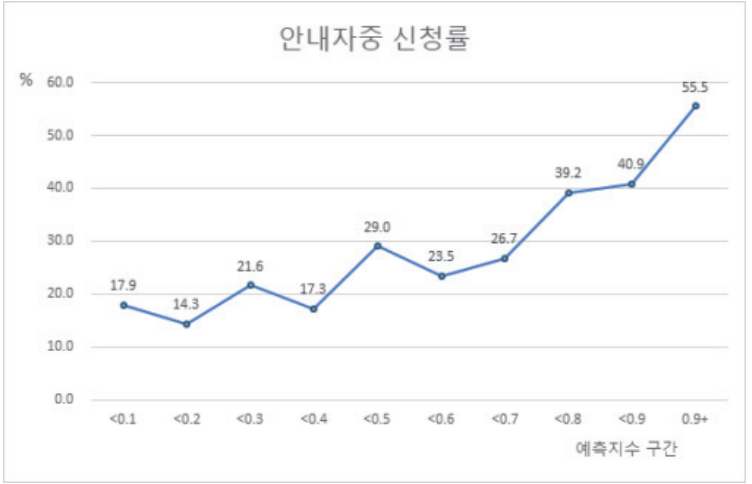
- 예측 확률을 활용한 안내 대상 명부 구축
- 신청 가능성 고순위 대상자 우선 안내 실시
- 전년 동기 대비 신청률 비교 분석

분석내용	활용 성과
• 농어업인 국고지원 미신청자에 대한 사유 파악 • 예측모델 개발을 통해 국고지원 신청 가능성이 높은 대상자 타겟팅 • 22.7월 말 기준 신청 대상 농어업인 예측지수 126만 건 제공	• 하반기 제도안내 대상자 132,825명에 예측지수를 활용하여 제도 안내 ➡ 전년 동기(9~10월) 신규 신청자 수 11,128명 대비 4,104명(36.9%) 증가

▶ 보험료 지원 신청 집단과 미신청 집단의 평균 예측 확률 비교

- 신청 집단의 예측 확률이 미신청 집단보다 높음
- 예측 확률(지수)이 높을수록 안내 대비 신청률이 증가하는 경향 확인

예측지수 구간	안내 대상 (A)	신청자 분포		대상자중 신청률 (D=B/A)	안내자 분포					
					신청		미신청		계	
		빈도 (B)	% (C)		빈도 (E)	% (F=E/I)	빈도 (G)	% (H=G/I)	빈도 (I)	%
–	1,687	206	5.1	12.2	0	–	0	–	0	0.0
〈0.1	62,695	1,347	33.3	2.1	253	17.9	1,162	82.1	1,415	46.7
〈0.2	13,433	452	11.2	3.4	56	14.3	336	85.7	392	12.9
〈0.3	7,576	256	6.3	3.4	49	21.6	178	78.4	227	7.5
〈0.4	5,235	191	4.7	3.6	23	17.3	110	82.7	133	4.4
〈0.5	4,123	178	4.4	4.3	31	29.0	76	71.0	107	3.5
〈0.6	3,745	174	4.3	4.6	19	23.5	62	76.5	81	2.7
〈0.7	3,553	158	3.9	4.4	24	26.7	66	73.3	90	3.0
〈0.8	3,883	206	5.1	5.3	47	39.2	73	60.8	120	4.0
〈0.9	4,475	269	6.7	6.0	67	40.9	97	59.1	164	5.4
0.9+	7,963	607	15.0	7.6	167	55.5	134	44.5	301	9.9
전체	118,368	4,044	100.0	3.4	736	24.3	2,294	75.7	3,030	100.0



정책활용/기대효과

기존 제도하에서 18~60세 기초수급자가 생계급여를 받기 위해서는 병원에서 진단서를 발급받고, 주민센터에 기초생활보장 신청 서류를 제출한 후, 국민연금공단에서 의학적 평가 및 근로능력 평가를 통해'근로능력 없음'판정을 받는 절차를 2~3년 주기로 반복해야 했다.

국민연금공단은 분석 결과를 바탕으로 복지부와 논의를 거쳐 '근로능력 평가 기준 등에 관한 고시'를 2022년 12월에 개정하였다.

연속 세 번 '근로능력 없음'인 경우 유효기간

호전 가능성	의학적 평가 결과 (단계)	유효기간		
		현행	개선	
고착	1	2년	3년	1년 연장
	2~4	3년	5년	2년 연장
비고착	2~4	2년	4년	2년 연장

제도 개정으로 근로능력 평가 유효 기간이 최소 1년에서 최대 4년까지 연장되었다. 이에 따라 제도 시행 첫해인 2024년에는 취약계층 2만 8천 명의 평가 부담 및 사회적 비용이 절감된 것으로 나타났다.

또한 농어업인 연금보험료 신청 가능성 예측모형을 활용하여 안내 대상자를 선정한 결과, 전년 동기 대비 4,104명의 농어업인이 신규로 국고 지원 혜택을 받게 되었다.

이처럼 국민연금공단은 국민의 데이터를 기반으로 불합리한 제도를 개선하고 주요 사업을 전개하는 등 국민 삶의 질 개선을 위해 노력하고 있다.

PART 2-2
공공행정

국가 보호지역, 최후의 4%를 지키자! 위성영상 기반 국립공원 변화탐지

국립공원공단

추진목적/배경

국립공원은 1헥타르(ha)당 14.3톤의 탄소를 흡수하며, 그 경제적 가치는 62조 원에 이르는 국가 보호지역의 핵심 지역이다. 따라서 국립공원을 보호하는 것은 기후 위기에 대응하는 가장 효과적인 방법 중 하나다.

그러나 최근 국립공원 내에서 기업형 농업 등으로 인한 산림 훼손이 점점 더 정교하고 광범위하게 이루어지고 있으며, 기후변화로 인한 산사태와 산불 등의 자연재해도 증가하고 있다. 하지만 국립공원은 면적이 넓고 접근이 어려운 특성상 불법 행위와 재난 상황을 신속하게 파악하는 데 어려움이 따른다.

이러한 문제를 해결하기 위해 공단은 위성영상을 활용해 산림 훼손 및 자연재해 발생 현황을 자동으로 탐지하는 시스템을 도입하였다. 이를 통해 기존 인력중심의 순찰 방식에 위성영상 기반 변화탐지를 적용하여 국립공원 관리 사각지대를 해소함으로써 보다 효율적이고 정밀한 관리 체계를 구축하고자 한다.

분석 사전 준비

– 활용 데이터

데이터명	형태	내용	출처	기준 년도	내·외부 데이터
위성영상 (SENTINEL-2A)	jp2	RGB 컬러영상	ESA(유럽우주국)	2021~2023	외부
위성영상 (SENTINEL-2A)	jp2	적외선(RED) 영상	ESA(유럽우주국)	2021~2023	외부
위성영상 (SENTINEL-2A)	jp2	근적외선(NIR) 영상	ESA(유럽우주국)	2021~2023	외부
공원경계	shp	공원경계	국립공원공단	2023	내부

본 과업에서는 국립공원의 특성을 반영하여 적절한 데이터 선정을 위한 세 가지 기준을 수립하였으며, 이에 부합하는 유럽우주국(ESA)의 SENTINEL-2A 위성영상을 활용하기로 하였다.

SENTINEL-2A 위성영상은 국립공원의 광범위한 산림 지역을 포괄적으로 촬영할 수 있으며, 3~5일 간격의 짧은 촬영 주기로 실시간에 가까운 현장 모니터링이 가능하다. 또한, 무료로 다운로드할 수 있어 데이터 확보 및 유지 관리에 대한 비용 부담이 거의 없는 장점이 있다.

위성영상은 Copernicus Browser(https://browser.dataspace.copernicus.eu/)에서 수집하였으며, 촬영 기간, 위성 종류, 운량, 촬영 전 기상 상태 등을 고려하여 지도상에서 원하는 범위를 설정한 후 검색하는 방식으로 영상을 확보하였다. 이후, jp2 형식의 RGB 영상, 적외선 영상, 근적외선 영상을 각각 추출한 뒤 GeoTIFF 형식으로 변환하였다.

아울러, 분석 대상지 선정을 위해 국립공원공단이 보유한 국립공원 경계 공간정보 데이터를 활용하였다.

분석과정

– 분석 환경

1. 분석 인프라 : 기관 내 PC 이용
2. 분석(개발) 환경 : python3.6
   - (영상수집) Sentinelsat 등
   - (UI 개발) PyQQt5 등
   - (영상처리) QGIS, Osgeo GDAL, Numpy, PIL 등
   - (결과표출) QGIS, Matplotlib 등
   - (기타) rasterio, datetime, os, shutil, sys 등

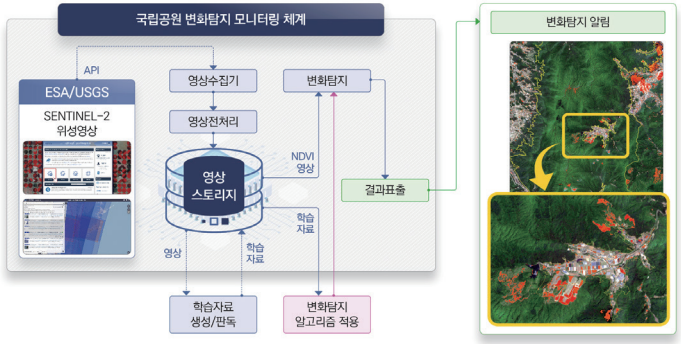
– 데이터 수집

1. 사용 데이터 출처 : 유럽우주국(ESA) 위성영상 다운로드
2. 사용 데이터 형식 : jp2, Geotiff

– 데이터 전처리

1. 변수 선정
   - 국립공원 경계 내부의 구름 여부를 변수로 선정
   - 위성영상 촬영 전 24시간 이내 강우 여부를 변수로 선정
   - 위성영상 내 노이즈 여부를 변수로 선정
2. 이상치 점검
   - 국립공원 경계 내부에 구름이나 노이즈가 포함된 위성영상 제외
   - AWS 자료를 참고하여 촬영 24시간 이전에 강우가 발생한 위성영상 제외
   - 국립공원 경계 및 고해상도 영상 자료를 활용하여 위성영상의 위치 정확도 점검
3. 이상치 처리
   - 지상기준점(GCP) 및 수치표고모델(DEM)을 활용하여 위성영상의 위치 보정 수행

– 모델링



국립공원 변화탐지 모니터링 체계 구현 프로세스

1. NDVI 기반 변화탐지 구현
   - SENTINEL-2 위성영상을 활용하여 자료 처리 및 분석 수행
   - NDVI 기반 학습 모델을 통해 변화 탐지를 수행하는 방식으로 구축
   - 모델 운용을 위한 GPU 및 관련 라이브러리 환경 구축

▶ 정규식생지수(NDVI, Normalized Difference Vegetation Index)
   - 식생 활력도를 측정하는 방법으로, 식물의 특성을 분석하는 데 활용됨.

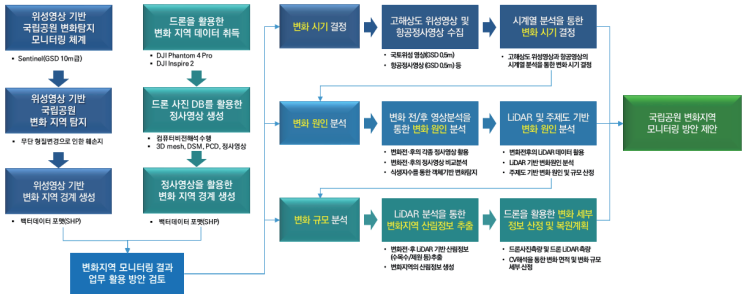
▶ 학습 자료 구축
   - 2020년 1월~2023년 4월까지 수집된 총 70개 위성영상(Scene) 활용
   - 월별 최대·최소 NDVI 값을 기준으로 총 103만여 개의 BIN(Binary File) 형식 데이터 생성

2. 변화 탐지 알고리즘 적용

▶ 이상치 탐지 기법(Outlier detection algorithm)
   - 데이터의 정상·비정상 범위를 구분하여 변화 여부 식별

– 검증 및 고도화

- 모델 검증을 위해 공원 내·외에서 발생한 변화 탐지 결과를 현장 조사 및 NDVI 분석을 통해 확인
- 변화의 원인, 시기, 규모 등 세부 정보를 파악하기 위해 공간정보 분석을 활용한 모니터링 방안 제안



변화탐지 검증 및 모니터링 프로세스

– 결과 구현

1. 플랫폼(변화 탐지 모니터링 플러그인) 개발
   - 사용성과 유지 관리 비용을 고려하여 QGIS 플러그인 기반으로 GUI 구성
   - SENTINEL-2 위성영상을 활용한 변화 탐지를 위해 영상 수집, 등록, 자료 처리 및 분석, 변화 탐지 결과 표시, 결과 알림 등의 특화된 기능 구현
   - 사용자 운용 환경을 Windows OS에서 구동되도록 설정하여 사용자 편의성 강화
2. 플랫폼(변화 탐지 모니터링 플러그인) 배포
   - 국립공원공단 치악산국립공원사무소 외부망 PC에 설치
   - 변화 탐지 분석 및 시각화 서비스 제공

정책활용/기대효과

국립공원공단은 2024년부터 치악산국립공원 전 지역(175.668㎢)을 대상으로 변화 탐지 모니터링 체계를 시범 운영하였다. 이를 통해 공원 내·외에서 발생한 총 28건의 자동 변화 탐지 결과를 현장 조사한 결과, 22건에서 의미 있는 변화를 확인하였다. 특히 불법 형질 변경 등 4건에 대해서는 고발 및 수사의뢰 등의 행정 조치를 취하며 실무 적용 가능성을 검증하였다.

기존에는 공단 직원이 지역담당 순찰제도에 따라 담당 구역(치악산 국립공원의 경우 1인당 평균 10.3㎢)을 순찰하였으며, 도보 및 차량을 이용해 개인의 주관적 판단에 따라 일부 지역만 점검하는 방식이었다. 이로 인해 공원 전역을 체계적으로 감시하는 데 한계가 있었다. 그러나 위성영상 기반의 자동 변화 탐지 기법을 도입하면서 변화가 감지된 특정 지역을 우선적으로 순찰할 수 있게 되었으며, 이를 통해 순찰 시간을 획기적으로 단축할 수 있었다. 또한, 울창한 산림과 가파른 경사로 인해 접근이 어려운 지역뿐만 아니라 사유지 및 보안 지역 등 기존에 확인이 어려웠던 곳까지 모니터링이 가능해지면서 순찰의 효율성이 극대화되었다.

본 과업의 궁극적인 목표는 불법 행위 적발 실적을 높이는 것이 아니라 사전 예방에 있다. 위성영상을 활용해 국립공원 전역을 지속적으로 모니터링하고 있음을 홍보함으로써, 공원 경계부에서 발생할 수 있는 인위적 산림 훼손 등의 불법 행위를 효과적으로 억제할 수 있을 것으로 기대된다.

또한 본 모니터링 체계는 QGIS 플러그인 형태로 개발되어 별도의 시스템 유지·관리 비용이 거의 발생하지 않으며, NDVI 학습 데이터 구축만으로 전국적으로 적용 및 확산이 가능하다. 이에 따라 광범위한 산림을 관리하는 중앙부처 및 지자체에서도 적극적으로 활용할 수 있을 것으로 기대된다.

"자연, 우리의 미래"는 국립공원공단의 대표 슬로건이다. 위성영상 기반 변화 탐지 모니터링 체계의 도입을 통해 미래 세대에게 국립공원의 아름다운 자연환경을 온전히 물려줄 수 있기를 기대한다.

변화탐지 모니터링 체계 현장적용 기대 효과

PART 2-3
공공행정

데이터 기반의 과학적 도시정비, 노후계획도시정비플랫폼

한국국토정보공사

추진목적/배경

노후 건축물 증가로 인한 주거 환경 악화, 안전 문제 발생, 도시 인프라 부족 등은 도시의 활력을 저하시킨다. 이러한 문제의 심각성은 2023년 4월 발생한 분당 정자교 붕괴 사고를 비롯해 아파트 단지 배관 부식, 도로 침하 등 크고 작은 사고들로 드러났다.

이에 정부는 2023년 12월, 「노후계획도시 정비 및 지원에 관한 특별법」(이하 '특별법')을 제정했다. 이 특별법은 노후계획도시를 광역적이고 체계적으로 정비하는 데 필요한 사항을 지원하여 도시기능과 주거환경을 개선하고, 미래도시로의 전환을 통해 국민 생활의 질을 향상시키려는 내용을 담고 있다.

현재 주민들이 거주하는 도시를 대규모로 재정비하는 것은 전 세계적으로도 유례를 찾기 어려운 일이다. 따라서 기존의 방식과는 다른 새로운 지원 체계가 필요하다.

한국국토정보공사는 디지털트윈(Digital Twin)을 기반으로 한 「노후계획도시정비플랫폼」을 통해 통합데이터 분석과 정책 집행 시뮬레이션을 제공하여 노후계획도시정비사업의 체계적 추진과 신속하고 정확한 의사결정을 지원하고자 했다.

분석 사전 준비

– 활용 데이터

데이터명	형태	내용	출처	기준 년도	내·외부 데이터
노후계획도시 지구단위계획	SHP, xlsx	지구단위계획구역 택지지구정보 등	택지정보시스템	2024	내부
인구 현황	SHP, TXT	인구수, 성별/연령별	통계지리정보서비스 (SGIS)	2022~ 2024	외부
가구 현황	SHP, TXT	가구수, 가구원수	통계지리정보서비스 (SGIS)	2022~ 2024	외부
종사자 현황	SHP, TXT	종사자수, 증감수	통계지리정보서비스 (SGIS)	2022~ 2024	외부
사업체 현황	SHP, TXT	사업체수, 증감수	통계지리정보서비스 (SGIS)	2022~ 2024	외부
학교 현황	xlsx	학생수, 학급수	학교알리미	2024	외부
도로 현황	SHP	도로 정보	도시계획정보시스템 (UPIS)	2024	외부
건축물 현황	DB	대지면적, 준공연도, 용도, 건폐율, 용적률 등	세움터 (건축물대장)	2024	내부

노후계획도시정비 기본계획 수립을 위해 정비사업 대상인 택지지구 경계를 기준으로 공공데이터 범위 내에서 기반시설 정보를 수집하였다.

분석과정

– 분석 환경

1. 분석 인프라 : 한국국토정보공사 플랫폼(LX플랫폼) 활용
2. 분석 환경 : WEB/WAS, Geoserver, 3DMap Server, 특화솔루션* 등

* 자동배치 시뮬레이션 모듈, 3D 가시화 모듈로 분석영역에 단지설계를 자동으로 수행하고 분석결과를 3차원 공간정보에서 표출하여 정보제공 및 의사결정을 지원하는 솔루션

– 데이터 수집

1. 사용 데이터 출처 : 기관 자체 생성 및 수집 데이터
2. 사용 데이터 형식 : SHP, TXT, XLSX, DB

– 데이터 전처리

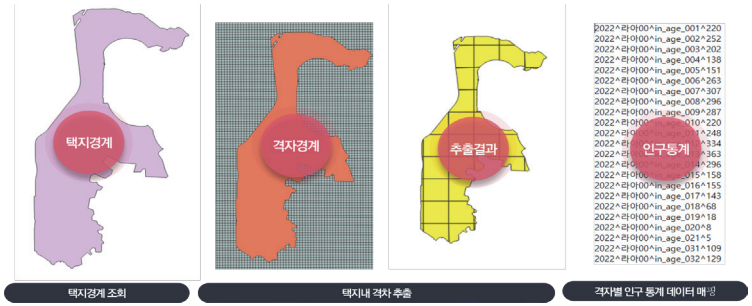
1. 격자 데이터 분석으로 택지별 영역 추출
   - 분석 대상 : 택지별 영역, ‘22년 인구 격자통계(통계청)
   - 격자 데이터 분석결과

지역명	면적	계획인구 (택지정보)	그리드크기 (격자)	추정 인구수 (분석결과)	비고
고양일산	15,735,711.0㎡	276,000명	100m	246,334명	
			500m	282,173명	
			1000m	324,723명	
군포산본	4,203,186.5㎡	167,896명	100m	116,573명	
			500m	145,350명	
			1000m	183,614명	
부천중동	5455778.4㎡	165,688명	100m	157,653명	
			500m	256,708명	
			1000m	319,144명	
안양평촌	5,105,904.4㎡	168,188명	100m	134,617명	
			500m	176,649명	
			1000m	239,441명	
성남분당	19,639,218.6㎡	390,320명	100m	340,640명	
			500m	391,021명	
			1000m	449,013명	

   - 연계 필요 데이터

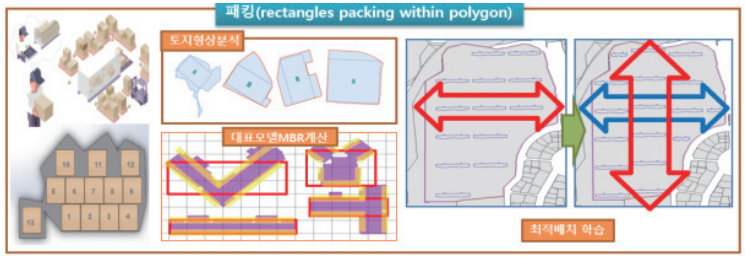
대상자료명		격자경계	자료형식	기준년도	대상지역
격자통계	인구	100m	TXT	2022	전국
	가구	100m	TXT		
	주택	100m	TXT		
	사업체	100m	TXT		
	종사자	100m	TXT		
격자경계		100m	SHP	2023	전국

2. 택지지구경계를 기준으로 데이터 추출

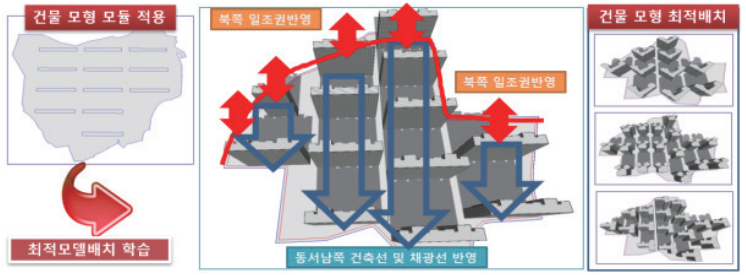


– 모델링

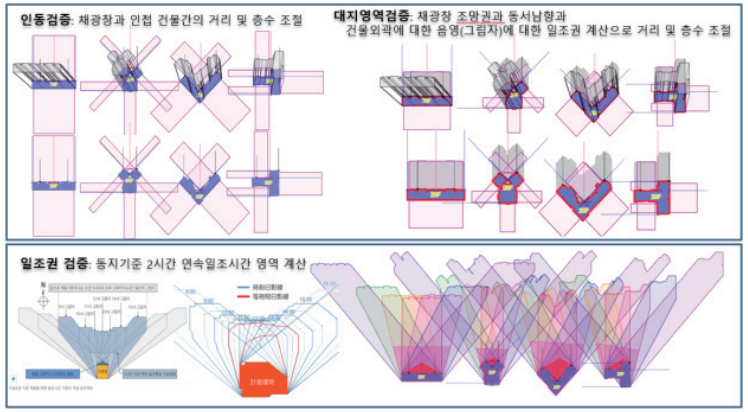
1. 패킹배치 학습 알고리즘을 적용한 부정형 사업영역 최대 동수 배치
   - 사업영역 대지 방향성 분석을 통한 건물 배치방향 계산
   - 다각형 토지 내에 최대한의 건물이 배치되도록 패킹배치 학습 적용
   - 행과 열에 대한 Moving배치 학습을 통해 최대 동수 재계산



2. 건물 모델별 최적배치
   - 각 타입에 지정된 MBR(사각형) 연산자를 사용하여 배치 위치를 참조로 건물 모델을 배치
   - 건물 모형에 맞춰 세로 그룹별로 건축후퇴선을 남쪽으로 최대한 이동시켜 배치
   - 건축후퇴선 포함 여부를 계산하여 사업영역 내 건물의 최대배치를 완료한 후, 건폐율 및 용적율 제한을 적용하고, 인동간격을 분석하여 빈 공간에 추가 건물을 배치

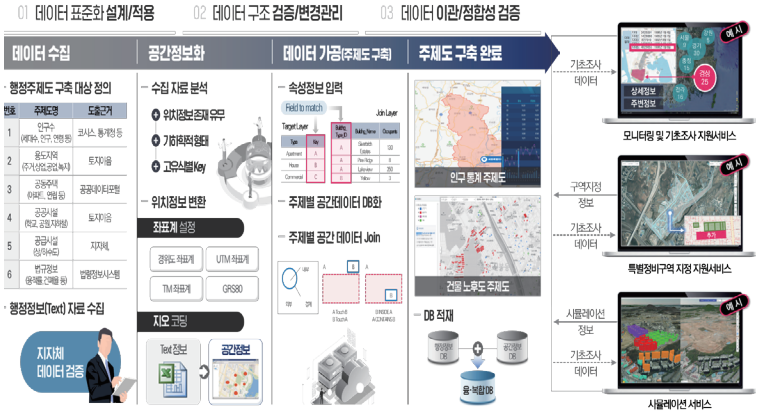


3. 건물배치 검증
   - 대지영역 검증: 북쪽은 일조사선(H/2), 동서남향은 채광창으로부터 조망선(H/2) 검증
   - 인동간격 검증: 각 동의 최적규모(층수)를 계산하여 채광창과의 조망선 중첩 여부 검증
   - 일조권 검증: 일조시간이 연속 2시간 이상 확보된 영역 분석
   - 검증 결과: 검증에 제한된 건물은 적색으로 표시

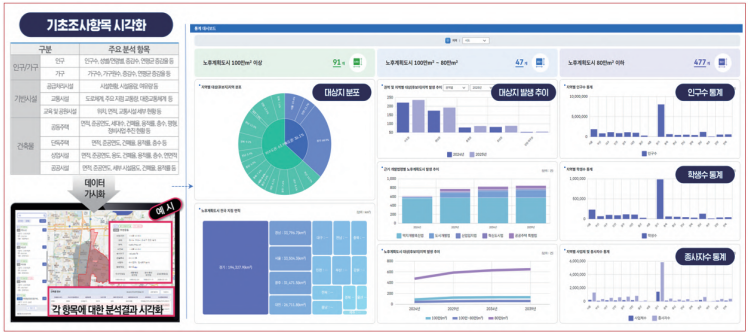


– 결과 구현

1. 기초조사 데이터셋 제공
   - 도시현황, 공간정보 등의 데이터 연계, 오류 검증 및 표준화 지원을 통해 신속한 정비기본계획 수립 지원



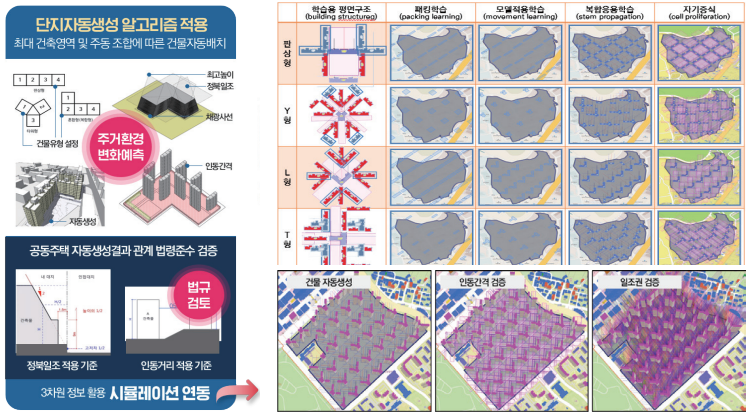
2. 데이터 통계 시각화
   - 특별정비예정구역에 대한 발생 추이 및 지역별 기초현황 통계 시각화



3. 특별정비예정구역 모니터링
   - 노후계획도시정비법에 따른 대상(후보)지 지도 기반 시각화 서비스
   - 후보지 상세정보와 기초조사 데이터 제공

4. 단지 자동배치 알고리즘 적용

• 특별정비예정구역의 토지형상, 법령, 규제 등을 적용한 자동배치 알고리즘으로 건물 자동배치와 최적 용적률 제공

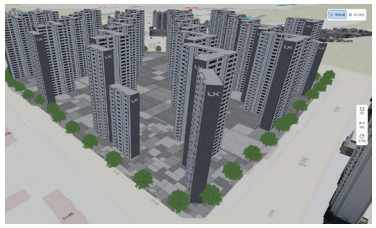


5. 디지털트윈을 활용한 시뮬레이션

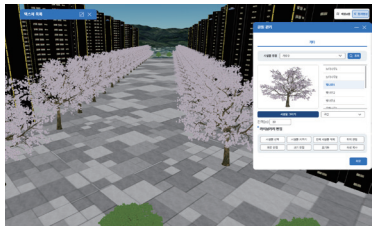
▶ 정비 전/후 비교 3D 시뮬레이션

• 자동배치 시뮬레이션 결과와 도시데이터를 연계한 정주여건 분석

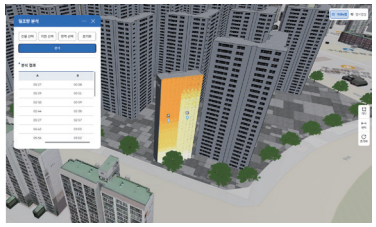
• 정비사업 전후의 일조량, 조망권 등을 비교·분석하여 정비사업 대상지역 검토



3D 가시화 및 텍스처링을 통합단지 생성



식재생성 기능을 통한 공원관리 기능



3D 가시화 및 텍스처링을 통합단지 생성

식재생성 기능을 통한 공원관리 기능

용적률 상향에 따른 도시밀도(경관, 조망권 등) 시뮬레이션 검증으로 주거환경 변화 예측

– 시스템 배포 및 운영

1. 노후계획도시정비플랫폼 서비스 개발

• 데이터 연계/수집→데이터 가공→단지 자동설계 모듈 적용→디지털트윈 시뮬레이션 저장→분석결과 시각화까지의 전반적인 프로세스를 구현할 수 있도록 시스템 구성

• 노후계획도시정비 포털을 포함하여, 모니터링, 특별정비구역 시뮬레이션, 단지 자동배치 알고리즘, 시각화 서비스 등을 패키징하여 배포

2. 노후계획도시정비플랫폼 운영 배포

• 한국국토정보공사에서 구축한 클라우드 기반 플랫폼에 배포

• LX플랫폼 기반으로 행정망 서비스를 구성하여 지속적으로 서비스를 개발하고 관리할 수 있도록 운영환경 구축

정책활용/기대효과

한국국토정보공사는 「노후계획도시정비플랫폼」의 구축을 통해 정비기간 단축과 사업 확산 촉진이라는 두 가지 주요 성과를 달성했다.

우선 특별법과 「도시및주거환경정비법」 등 관련법에 근거한 행정데이터 및 개방데이터를 수집하는 방법과 절차(기초조사 단계)가 종전 방식에 비해 약 1/3 수준으로 간소화되었다(약 3개월 → 약 1개월). 기존에는 데이터를 수집하기 위해 관련법을 분석하고 필요한 데이터 목록을 정리한 후, 데이터를 제공하는 기관 및 플랫폼에 개별적으로 접속해야 했다. 그러나 「노후계획도시정비플랫폼」을 통해 해당 정비구역을 지도에서 설정하면 관련 데이터가 일괄적으로 제공되며, 데이터 기반의 정비지역 기초분석 보고서도 자동으로 생성된다. 이를 통해 불필요한 행정력을 크게 줄이고 데이터에 기반한 정책 실행이 가능해졌다.

또한 노후계획도시의 예상 모습을 디지털트윈(Digital Twin)을 기반으로 한 3D 시뮬레이션을 통해 시각화함으로써, 특정 시간대와 위치에 따른 일조권, 조망권, 인동간격 등의 분석 결과를 한눈에 파악할 수 있게 되었다. 이렇게 시각화된 분석 결과를 토대로 주민, 조합, 시공사 등 이해관계자에게 노후계획도시정비사업에 대한 이해를 쉽고 빠르게 도울 수 있게 되었다. 그 결과, 기존의 노후계획도시정비 기본계획 수립 단계에서 정주환경 적정성 검증 등의 기간이 1/2 수준으로 단축되는 효과를 보였다.(약 8개월→약 4개월)

PART 2-4
공공행정

병역판정자료를 연계·활용한 병역면탈 징후 탐지

병무청

추진목적/배경

병역의무를 회피하거나 불법적으로 면제받으려는 병역면탈 행위가 지속적으로 증가하는 추세이다. 과거에는 외형적 특징을 이용한 병역면탈 시도가 많았지만 최근에는 정신질환을 위장하는 등 더욱 지능적이고 은밀한 방식으로 변화하고 있다.

특히 병역면탈은 다른 범죄와 달리 직접적인 피해자가 없고 제보 및 수작업 자료 분석에 의존하는 특성으로 인해 수사가 어렵다.

현재의 수사는 제보나 수사관의 경험에 의존하여 병역면탈 의심자를 선별하고 범죄 혐의점 여부를 확인하는 방식인데, 이는 시간이 많이 걸리고 업무 정확도와 인력 운용 측면에서도 비효율적이다. 따라서 보통 2개월이 소요되는 병역면탈 의심자 식별 과정에서 단순 반복 업무의 시간을 줄일 필요가 있었다.

이러한 문제를 해결하기 위해 병무청은 최근 5년간의 병역판정검사 데이터. 병역의무자 데이터 및 각종 외부데이터를 결합하여 병역면탈 수사를 위한 분류 모델을 개발했다. 이를 통해 수사 업무의 효율성을 높이고, 기획수사 정책 수립의 근거를 마련하였다.

분석 사전 준비

– 활용 데이터

데이터명	형태	내용	출처	기준 년도	내·외부 데이터
병역판정검사	csv	검사일자, 종류 등	병무청	2017-2022	내부
병역판정검사결과	csv	검사항목별 결과	병무청	2017-2022	내부
병역처분내역	csv	의무자의 병역처분정보	병무청	2017-2022	내부
검사부령조항	csv	검사항목별 부령, 치유기간 등	병무청	2017-2022	내부
의무자 정보	csv	의무자 정보 (생년, 학력, 지역 등)	병무청	2017-2022	내부
연기내역	csv	연기종류, 일자 등	병무청	1999-2023	내부
민원접수내역	csv	민원출원내내용	병무청	2008-2023	내부
진료이력	csv	병원명, 병역처분 등	국민건강보험공단	2023	외부
취업이력	csv	사업자명, 취득일자	근로복지공단	2023	외부
자격면허취득자료	csv	자격면허구분, 자격면허 등	산업인력공단	2023	외부

병무청이 보유한 의무자의 최근 5년('17~'22년) 병역판정검사 관련 정보, 병역의무 이행연기 및 각종 민원출원 내역 등을 기본 분석 대상으로 선정하였다. 또한 개인을 식별할 수 있는 병적번호(ID)는 개인정보 비식별화솔루션을 이용하여 비식별 처리하였다.

병역면탈 의심자의 특징을 식별하기 위한 독립변수(feature)로 진료내역 및 관련 질병, 범죄이력, 취업 정보, 자격·면허 여부 등을 설정하고, 이를 기반으로 학습 데이터셋을 구축하여 AI 기반 분류 모델을 개발하였다.

외부 데이터는 병역법에 따라 정신 질환 수검자의 진료 이력 등 6종의 행정 정보를 행정정보공동이용센터를 통해 수신받아 활용하였다. 다만, 법적 규제로 인해 최초 이상징후 탐색 단계에서는 외부 데이터를 결합하지 못하고 면탈의심군을 분리한 이후에야 결합이 가능했던 점은 한계라 생각된다.

분석과정

(분석 프로세스 도식화)

– 분석 환경

1. 분석 인프라 : 기관 내 PC 이용
2. 분석 환경 : Python, Jupyter notebook 사용

– 데이터 수집

1. 사용 데이터 출처 : 기관 자체 생성 데이터, 법무부, 건강보험공단, 산업인력공단, 근로복지공단
2. 사용 데이터 형식 : csv

– 데이터 전처리

1. 결측값 정비
   - 검사부령조항 데이터에서 치유개월이 N/A인 경우 0으로 대체

2. 데이터 제거
   - 각 데이터에서 중복 및 불필요 데이터항목 제거
3. 데이터 형식 통일화
   - 일자를 년/월/일 형태로 표준화 처리
   - Float형 숫자를 Integer 형태로 변환
   - 숫자 및 문자열 구분이 어려운 경우 문자열로 통일
   - 값이 여부(N/Y)인 경우 0/1로 변환
4. 파생변수 생성
   - 병역판정검사 및 민원 등이 의무자별로 여러번 발생 가능한 점을 고려하여 차수, 횟수, 당시 나이 등의 파생변수 생성
   - 면탈자 특성을 반영한 파생변수(1차 32종, 2차 16종)을 생성하여 분석성능 향상
5. 불균형 데이터 처리
   - 샘플링 기법
   - SMOTE 알고리즘

– 모델링

1. 탐색적 데이터 분석(EDA)

▶ 데이터 특성 확인
   - 병원별 병무용진단서 제출횟수 분석
   - 급증 질환 병역면탈의심자 분석
   - 손바닥 다한증 질환 상세분석
   - 틱관절 장애 질환 상세분석
   - 속임수 질환 면탈의심자 분석

▶ 데이터 간 연관성 확인
   - 4~6급 병역의무자와 병역면탈자간 연도별 병역처분사유 추세 분석
   - 중점관리질환(34종)의 연도별 발생 추세 분석
   - 일반입영자와 병역면탈자간 연기이력 및 특징분석 및 분석 착안사항 발굴
   - 일반입영자와 병역면탈자간 병역처분변경원 특징 분석 및 분석 착안사항 발굴
   - 사기 및 절도 이력 병역면탈자 분석
2. 분류 분석
   - 면탈자 특성(주요 질병, 진단 병원, 사기 이력 등)과 유사한 특성을 가진 의무자 분류

▶ 사용한 주요 분류 모델
   - Logistic Regression
   - Random Forest
   - XGBoost
   - QDA

▶ 모델 앙상블 기법 활용
   - Boosting
   - Bagging
   - Blending
3. 모델 선정
   - 예측모델의 성능평가지표인 Accuracy와 Recall 값을 비교하여 최종 Logistic Regression 모델을 확정 (Accuracy 97.5, Recall 86.5)
     * 실제 면탈자 인원(Positive)을 모델이 면탈자로 예측(True)한 비율인 재현율(Recall)을 평가지표로 채택
   - 해당 모델을 적용하여 추출된 병역면탈의심군에 대해 면탈여부 분류 및 예측 스코어를 포함한 리스트 도출

– 검증 및 고도화

- (검증) 추출된 병역면탈의심군(76명)에 대한 정밀분석을 통해 일부에 대한 추가 정보 수집 등 면탈여부 조사 중

• (고도화) 개발된 분석모델을 활용하여 새로운 면탈 의심군을 파악하기 위해서는 사용자 또는 모델 운영자가 데이터를 새로 확보하고 전처리하여 모델을 재실행해야 함. 이러한 불편함을 개선하기 위해 병무청에서는 '24년「데이터 통합 병역면탈 조기경보체계」구축 사업을 추진함.

- 결과 구현

▶ 병역면탈의심군 목록(List) 제공과 Tableau를 이용한 시각화로 사용자 편의성 제고

• 면탈의심 질병, 취업 및 연기횟수 등을 시각화 서비스로 제공

병역면탈 의심군 현황 대시보드 초기화면

판정검사 현황 대시보드 초기화면

정책활용/기대효과

병무청은「정신질환 관련 병역면탈 의심군 도출」모델을 활용하여 정신질환과 관련된 병역면탈의심자 76명을 식별하고 심층 분석을 진행하였다. 이 중 일부에 대해서는 수사 착수 여부를 검토하고 있다.

또한 병역면탈 이상징후 탐지 모델을 개발하여 면탈 의심자 발굴 소요시간을 기존 2개월에서 2주로 대폭 단축하였다. 아울러 정신질환뿐만 아니라 약 20여 개의 주요 속임수 가능 질환에 대한 면탈 의심군 추출 가능성도 확인하였다.

이전	수사착안 도출 (수사 경험)	의심자 추출 (1,885명 수작업)	면탈혐의자 수사	검찰송치	수사진행 (내사+입건)
	1개월 소요	1개월 소요			
			↓		
개선	(자동화)	수사 착안사항 도출 면탈 의심자 추출	면탈혐의자 수사	검찰송치	수사진행 (내사+입건)
	2주 소요				

이러한 성과를 바탕으로 병무청은 2024년「데이터통합 병역면탈 조기경보체계」구축 사업을 수행하여 분석 대상을 정신질환뿐만 아니라 22개 중점 질환 및 확인신체검사, 계속치료질환, BMI 등으로 확대하였다. 또한 탐색적 분석 결과는 병역면탈 기획수사 정책수립의 근거로 활용될 예정이다.

본 모델이 지속적으로 확장되어 과학적 병무행정 추진에 기여하길 기대한다.

PART 2-5
공공행정

저수지 수위 변화 예측 및 수문 조작 의사결정 지원 모델개발

한국농어촌공사

추진목적/배경

최근 기후 변화와 도시화가 가속화됨에 따라 수자원 관리의 중요성이 더욱 부각되고 있다. 특히 저수지는 농업, 산업, 생활용수 공급 등 다양한 분야에서 핵심적인 역할을 담당하고 있어 수위 변화의 정확한 예측과 효율적인 수문 조작이 필수적이다. 그러나 기존의 전통적 수위 예측 방법은 기상 조건 변화나 인위적 요인에 대한 민감도가 낮아 정확한 예측에 한계가 있다. 이러한 문제를 해결하기 위해 인공지능(AI) 기반의 수위 변화 예측 모델 개발과 이를 활용한 수문 조작 의사결정 지원 시스템의 필요성이 대두되고 있다.

본 프로젝트의 목적은 AI 기술을 활용하여 저수지 수위 변화를 보다 정확하게 예측하고, 이를 기반으로 수문 조작 의사결정을 지원하는 모델을 개발하는 것이다. 이를 통해 수자원 관리의 효율성을 높이고 예기치 않은 수위 상승이나 하강으로 인한 재해를 예방할 수 있을 것으로 기대된다. 또한 이러한 모델은 수자원의 지속 가능성을 확보하고, 물 부족 및 홍수와 같은 문제를 예방하는 데 기여할 것이다.

더불어 저수지 운영자와 정책 결정자에게 실질적인 의사결정 도구를 제공하여 수자원 관리의 체계성을 강화하고 환경 변화에 보다 유연하게 대응할 수 있는 기반을 마련하는 데 기여할 것이다. 이는 궁극적으로 지역 사회의 안전과 발전에 긍정적인 영향을 미칠 것으로 예상된다.

분석 사전 준비

– 활용 데이터

데이터명	형태	내용	출처	기준 년도	내·외부 데이터
저수지제원	CSV	장비번호, 표준코드, 유효저수량 등	한국농어촌공사	2023	내부
수위계측정보	CSV	장비번호, 관측일시, 수위	한국농어촌공사	2023	내부
저수지유역정보	CSV	장비번호, 면적, geometry	한국농어촌공사	2023	내부
기상관측소 강수량	CSV	장비번호, 일누적강수량	기상청	2023	외부
5km 격자 초단기실황	CSV	격자ID, 발표일시, 1시간강수량 등	기상청	2023	외부
5km 격자 초단기예보	CSV	격자ID, 예보일시, 1시간강수량 등	기상청	2023	외부
5km 격자 단기예보	CSV	격자ID, 예보일시, 1시간강수량 등	기상청	2023	외부

한국농어촌공사의 공공데이터포털에 공개된 수위계측정보, 公社 내부시스템(농촌용수종합정보시스템, RAWRIS)의 저수지제원 및 유역정보와 기상청 공공데이터포털의 관측 데이터를 사용하였다. 수집한 데이터에 대해 현황분석 및 전처리를 수행하였으며, 집중호우 시 수위 예측을 위한 치수(治水) 회귀분석과 기상청 데이터의 신뢰성 분석을 실시하였다. 분석 결과 12시간 누적강우량이 존재하면서 저수지 수위가 상승하는 경우를 학습한 예측모델이 가장 높은 신뢰도를 보였다. 이를 바탕으로 AI모델을 개발하여 公社 내부시스템(농촌용수종합정보시스템, RAWRIS)에 연계하였다.

분석과정

– 분석 환경

1. 분석 인프라 : 기관 내 PC 이용, 분석자원 인프라 지원
   - ·OS(Rocky Linux 9.2)
   - ·Memory(32GB)
   - ·Storage(295.8GB)
   - ·Graphics Driver(VMware SVGA II Adapter)
2. 분석 환경 : python 3.9

– 데이터 수집

1. 사용 데이터 출처 : 공공데이터 포털, 기관 자체 생성 데이터
2. 사용 데이터 형식 : csv

– 데이터 전처리

1. 기상청 격자 레이어를 생성하여 저수지 유역 폴리곤과 교차영역 분석
2. 저수지 유역과 기상청 격자가 중첩되는 면적 비율 산출

3. 격자 단위 기상 데이터를 면적 비율에 적용하여 저수지 유역 강우량 데이터 생성
4. 이상치 처리
   • 제정고 초과 값 결측 후 선형보간
   • Hampel Filter*
   * 시계열 데이터에서 이상치(outlier)를 검출하고 제거하는 필터링 기법으로, 데이터의 중앙값과 표준편차를 기준으로 이상치를 정의하여 이를 수정하거나 제외함
   • 강수 미발생 & 수위 상승 처리
5. 격자 데이터를 이용하여 유역면적 기상 데이터 생성
6. 집중 호우 발생 시 수위 예측 모델과 평상 시 수위 예측 모델 구분
7. 기상청 데이터 신뢰성 분석
   • ASOS 관측소와 격자 초단기 실황 비교
   • 초단기 예보 신뢰성 분석
8. 학습용 데이터 구축
   • 저수지 제원, 수위 계측 정보, 저수지 유역 정보, 기상 관측소 강수량, 5km 격자 초단기 실황, 5km 격자 초단기 예보, 5km 격자 단기 예보 등 최종 활용 데이터 결정

– 모델링

1. 치수 모델(호우 시) 개발 프로세스

▶ 사용 모델
   • AutoML (XGBoost, CatBoost, RandomForest, LightGBM 등)

▶ 개별 모델과 군집 모델 각각 개발 후 검증을 통해 선정
   • 개별 모델
     → 저수지별 강우 발생 데이터를 추출하여 학습 데이터를 생성한 후 학습
   • 군집 모델
     → 저수지 군집 후 군집별 학습 데이터를 생성하여 학습

2. 저수지 군집화

▶ 1차 군집
   • 저수지 유역면적, 만수면적, 유효저수량, 최대유역면적강수량, 평균유역면적강수량을 활용해 군집화(K-Means)
   • 0, 1 군집에 저수지가 집중되어 있으며 강수량에 따른 수위 변화를 반영하지 못함 → 2차 군집 시행

▶ 2차 군집
   • 6시간 누적 강수량에 따른 수위 변화율을 이용해 군집화 진행하여 7개의 군집 생성
   • 군집별 유효저수량, 만수면적, 유역면적 평균 확인하여 적절하여 군집된 것을 확인

3. 이수 모델 개발 프로세스

▶ 사용 모델
   • DLinear*
   * 시계열 데이터를 트렌드, 계절성, 잔차로 분해하여 각 요소를 독립적으로 예측하고, 이를 결합해 정확한 예측을 제공하는 모델로, 장기 예측에 강력함

▶ 수위 예측 모델과 저수율 예측 모델을 각각 개발하여 선택
   • 수위 예측 모델(2019년~2023년 데이터 활용)
   • 저수율 예측 모델(1991년~2023년 데이터 활용)

4. 평상 시 수위 예측 모델
   • DLinear, AutoARIMA*, LSTM을 활용해 이수 수위 예측 모델 개발
   * 시계열 데이터의 최적 ARIMA 모델을 자동으로 선택하고 학습하는 알고리즘으로, 데이터의 차분, 계절성, 자기상관을 고려하여 예측 성능을 최적화함
   • Hyper Parameter를 조정하여 16개 모델 개발

– 검증 및 고도화

1. 치수 모델 전체 저수지 정확도 확인
   - 2023년 저수지별 유역면적강수량 평균 초과 기간으로 1,663개 저수지 중 검증 제외 저수지 16개를 제외한 저수지로 검증
   - 전체 저수지의 87.7%가 강수 발생 상황에서 6시간 이후 예측까지 R2* 0.7 이상의 정확도를 가짐

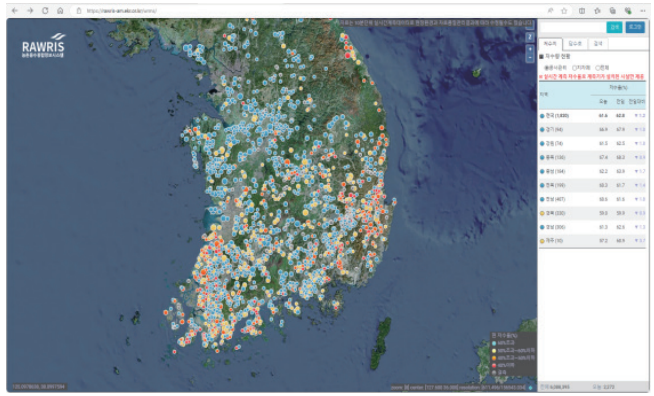
   * R2(결정계수) : 회귀 모델에서 모델이 설명하는 데이터 변동의 비율을 나타내는 지표로, 1에 가까울수록 모델이 데이터를 잘 설명함을 의미함

2. 치수 모델 대표 저수지 검증
   - 2023년 집중 호우 특별재난지역으로 선포된 지역의 수문이 있는 저수지 중 규모가 평균에 가까운 경천 저수지를 대표 저수지로 선정하여 검증함
   - 호우 발생 상황에서 기상 예보를 활용해 예측한 결과 실제 수위 대비 예측 수위의 정확도는 99.54%
   - 3시간 예측의 평균 오차는 0.07m, 12시간 예측의 평균 오차는 0.16m

3. 이수 모델 대표 저수지 검증
   - 2023년 저수지별 누적 강수량 0인 기간으로 50개 저수지로 검증
   - 전체 저수지의 78%가 강수 발생 상황에서 4일 이후 예측까지 R2 0.7 이상의 정확도를 가짐

– 결과 구현

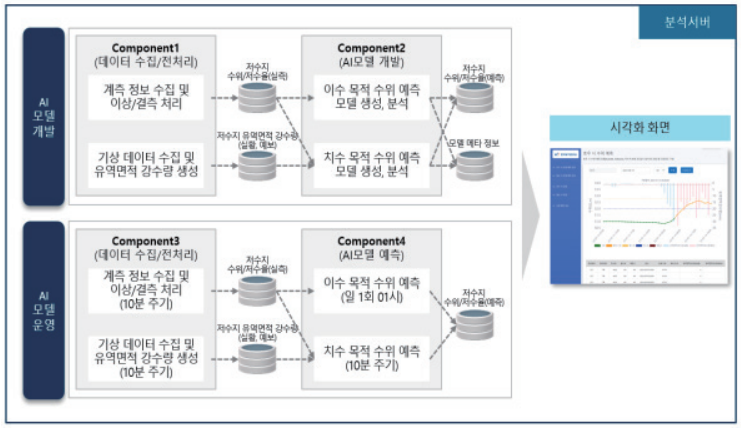
1. 시각화 서비스 개발
   - 치수 모델
     → 저수지 및 일시 조회 시 12시간 후 예측 수위까지 시각화
   - 이수 모델
     → 저수지 및 일자 조회 시 14일 후 예측 저수율까지 시각화



저수량 현황 시각화

2. 시스템 도입 및 고도화

▶ 저수지 수위 변화 AI 예측 모델 도입
   - 기상청 격자 데이터 자동 수집 및 적재
   - 저수지 유역 면적 기상데이터 생성
   - 저수지 수위 변화 예측 모델을 통해 호우 시/평상 시 수위 예측
   - AI 모델 운영
     → 운영 시 모델 강화 학습
     → 10분 주기 계측 정보 수집 및 이상치, 결측치 처리
     → 10분 주기 기상 데이터 수집 및 유역면적 강수량 생성
     → 일 1회 01시에 이수 모델 수위 예측
     → 10분 주기 치수 목적 수위 예측



▶ 운영시스템 반영

- 농촌용수종합정보시스템(RAWRIS) 연계 및 시범 운영
- 도입 이후 데이터 축적과 수위 예측 대상 확대 등에 기반한 모델 고도화 진행

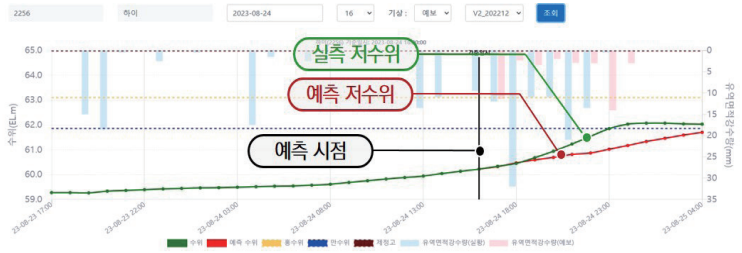
▶ 시스템 연계 확장

- 저수지 관리자 정보 전송 기능 고도화
- 홍수 위험 저수지 표출 및 피해 예상 지역 알림 시스템 구축

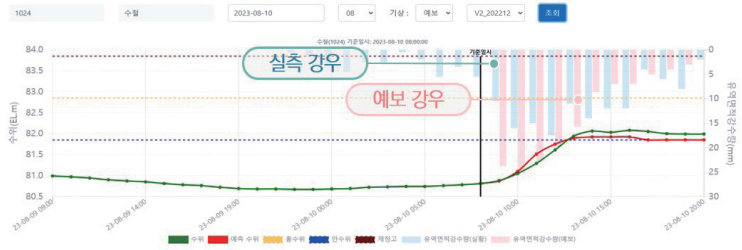
정책활용/기대효과

한국농어촌공사는 저수지 홍수 예·경보 시스템 구축을 통해 월류 위험이 높은 저수지를 대상으로 2시간 후의 예측 결과를 홍수 예·경보 상황전파 체계에 반영하여 골든타임을 확보했다. 또한 '23년과 '24년의 강우사례를 추가로 학습하여 예측 정확도를 향상시키고 모델을 고도화하였다. 아울러 농촌용수종합정보시스템(RAWRIS)에 탑재한 AI 기반 저수지 수위 예측 모델의 매뉴얼을 '24년 배포하였으며, 해당 홍수예측시스템과 재난안전종합상황실(DIMAS) 홍수 정보 표출 및 경보 전송 기능을 향후 고도화할 예정이다.

본 시스템을 활용해 저수지 관리자들은 수문 가동 시기 등 수문 조작에 관한 의사결정을 보다 효과적으로 내릴 수 있으며 재해를 사전에 감지하여 적절한 대응 계획을 수립할 수 있다. 또한 이번 모델이 홍수로 인한 국민의 불안을 해소하고 저수지 하류 지역 주민들의 안전을 확보하며, 재산 피해를 최소화하는 데 있어 기여할 것으로 기대된다.



고성군 하이저수지 수위 예측



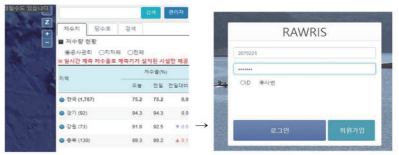
영덕군 회동저수지 수위 예측

활용 전	활용 후
· 시스템상 단순한 현재 저수지 수위 표시 · 실시간 계측 수위를 시설 담당자에게 발송하여 수위 상승에 따른 대처 시간 부족 · 발생한 강우에 따른 수위 예측 불가에 따른 위기상황대응을 위한 상황판단 지연	· 기상청 기상예보 자료 활용한 과거·현재·미래(2시간) 저수지 수위 관측 가능 · 예보된 강우로 인해 상승하는 저수지 수위 예측 및 조기경보를 통해 대처 시간 확보 · 재난상황시스템(DIMAS)와 연계하여 홍수발생 예측 시 주민대피 등 상황대응 위한 판단 가능

RAWRIS AI기반 저수지 홍수예측시스템 매뉴얼

1 **RAWRIS-AM 시스템 접속 :** https://rawris-am.ekr.or.kr/

2 **로그인 :** 우측상단 '관리자' 아이콘 클릭 → (사번으로 전환 → ID 사번, 비번 사번 입력)



3 **홍수저수지 조회 기능**

[경로: 왼쪽 중단 '홍수예측시스템' → 'AI홍수정보' → '홍수저수지']

- ㅇ 집중호우 시 홍수 위험 저수지를 직관적으로 확인할 수 있도록 저수위 예측 결과를 지도에 색상으로 표출
- ㅇ 예측 기준 시각, 예측 시간 등을 설정하여 예측 결과 조회 가능
- ㅇ 저수지 아이콘 클릭 시 계측정보, 그래프, 홍수예측 결과 조회 가능

홍수저수지 지도 표출	예측 기준 시각 등 설정 화면	저수지별 홍수 예측 결과조회

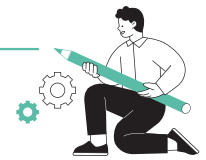
'24년 AI기반 저수지 홍수예측시스템 매뉴얼 배포

재난
안전

PART 3

재난안전

1. 전국상수도 운영데이터 통합모니터링 및 위기대응체계 구축
   한국수자원공사
2. 빅데이터 기반 시설물의 건설부터 유지관리까지 선제적 사고 예방
   국토안전관리원
3. 출동데이터를 활용한 골든타임 미확보 구역 특성 분석
   부산소방재난본부
4. 산재정보 분석을 통한 재해안전지수 개발
   근로복지공단



PART 3-1
재난안전

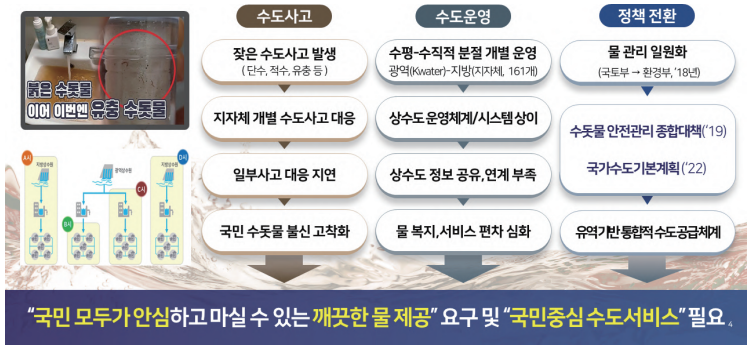
전국상수도 운영데이터 통합모니터링 및 위기대응체계 구축

한국수자원공사

추진목적/배경

국민 모두가 안심하고 마실 수 있는 깨끗한 물 공급을 위해 환경부는 2019년 11월 「수돗물 안전관리 종합대책」을 수립했다. 이에 따라 한국수자원공사는 광역-지방상수도 데이터통합 인프라를 기반으로 전국 수도시설(광역50, 지자체161)의 실시간 운영정보를 모니터링하고 있다. 또한 위기대응 컨트롤타워 역할을 하는 상황실을 구축하고, 유역수도 운영지원시스템을 운영하여 수도정보 기반의 기술지원을 수행하고 있다.

본 과제에서는 수돗물 안전관리 종합대책 및 국가수도기본계획에 따라, 통합데이터 기반의 상수도 관련 공동 위기대응체계 구축 및 고도화를 목표로 데이터를 분류하고 분석모델 및 알고리즘을 개발하였다. 이를 통해 중앙정부와 수도사업자 간의 데이터 기반 공동 위기대응체계를 구축하여, 국민이 안심하고 마실 수 있는 물 환경 인프라를 조성하고자 하였다.



분석 사전 준비

– 활용 데이터

데이터명	형태	내용	출처	기준 년도	내·외부 데이터
지자체 수도정보 (수량, 수질)	xlsx (db 추출)	수량(유량, 수위, 수압 등), 수질 등 실시간자료 4.6만개	161개 지자체	2023 (실시간)	내부
광역 수도정보 (수량, 수질)	xlsx (db 추출)	수량(유량, 수위, 수압 등), 수질 등 실시간자료 4만개	K-water	2023 (실시간)	내부
수도시설 제원정보	xlsx	시설명, 시설용량 등 시설관련 자료	환경부 (국가상수도정보시스템)	2023	외부
기상정보	xlsx	기상청 날씨정보 (강수, 바람 등)	기상청	2022	외부
뉴스정보	xlsx API	인터넷 수도사고 관련 뉴스정보	한국언론진흥재단 (Bigkinds)	2023	외부
도로 CCTV 정보	xlsx	7,000여개 도로 CCTV 정보 중 수도관로 인근 1,500 여개	도시교통정보센터	2023	외부
시군구 행정동 코드	xlsx	전국 시군구 행정동 코드	공공데이터포털	2023	외부

K-water 및 지자체와 환경부 간 연결된 유역수도망을 활용하여 실시간 수량·수질정보 4.6만개와 광역상수도 수도정보 4만개의 실시간 정보(1분 단위)를 엑셀 형태로 수집하였다. 또한 기상청 포털에서 강수량, 날씨, 풍량 등의 기상정보를 확보하였으며, 한국언론진흥재단 Bigkinds의 API를 활용하여 수도사고 관련 뉴스정보를 실시간 연계하였다. 아울러, 도로에 매설된 수도관로 인근 영상정보를 실시간으로 확인하기 위해 도시교통정보센터에서 제공하는 도로 CCTV 정보를 API를 통해 확보하였다.

분석과정

– 분석 환경

1. 분석 인프라 : 기관 내 PC 이용
2. 분석 환경 : python, anaconda 등

– 데이터 수집

1. 사용 데이터 출처 : 기관 보유 데이터, 지자체 연계 데이터, 공공데이터 등
2. 사용 데이터 형식 : xlsx

– 분석 과제

① 운영데이터 분석 및 경보설정을 통한 데이터 기반 감시
- 지자체 운영패턴(시설별, 시공간별) 분석 및 데이터 유형별 운영데이터 경보 분석
- 운영데이터 자동경보설정 및 경보기반 상시 사고모니터링 시스템 구축

② 영상기반 수도 관로사고 감시
- 이미지 기반 K-water형 관로 누수 사고 검출 알고리즘 개발
- 공공 개방데이터를 활용한 지능형 관망 CCTV 영상감시시스템 구축

③ 배수지 용수공급 관련 수요예측
- 상수도 위기대응의 핵심 시설인 배수지 용수 패턴분석 및 수요예측 모델 개발
- 시설물(배수지) 운영지원(영향분석/용수공급가능시간 등) 시스템 구축

④ 수도사고 관련뉴스 분석 및 공유체계 구축
- 상수도 핵심 키워드 중심 뉴스정보 입력데이터 분석 및 분류모델 개발
- 이벤트 기반의 실시간 뉴스정보 분석 및 공유 시스템 구축

– 데이터 전처리

① 운영데이터 분석 및 경보설정(시계열 데이터)

▶ 결측값 및 오류 점검

- null 형태로 수집된 결측 데이터 제거
- 숫자 형식으로 표현되지 않은 데이터 제외

▶ 이상치 점검

- 사분위 기법을 적용하여 ±1.5 IQR에서 벗어나는 이상치 제거
- 최대, 최소값으로 지정된 범위를 벗어나는 이상치 제거

② 영상기반 수도 관로사고 감시(영상 데이터)

▶ 관로 누수 데이터 수집 및 생성

- 물기둥, 누수, 물비침 등 합성 이미지 생성을 통한 관로사고 데이터셋 생성(앙상블 기반 알고리즘 개발)

▶ 데이터 강화

- 이미지 변조기술을 사용하고 학습데이터를 증강하여 모델 분류 정확도 강화(CCTV 정보(api) 기반 영상저장 후 관로사고 이미지 12000장 생성)

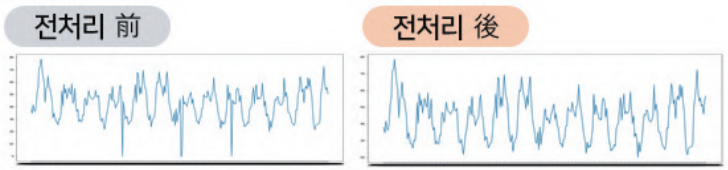
③ 배수지 용수공급 관련 수요예측

▶ 시계열 변수 생성 및 자료 구조 변경

- 지자체 배수지별 유입/유출유량, 수위, 기상데이터 등 8종 32천개 데이터

▶ 이상치 점검

- 사분위수 활용, Null, 0값 제거 등 이상치 제외



④ 수도사고 관련뉴스 분석 및 공유체계 구축

- 동일기사(유사도 100%), 중복제거(133건)
- 기사 URL 오류 방지 등을 위한 영어 및 특수문자 제거 (〈b〉 / 〈n〉 등)

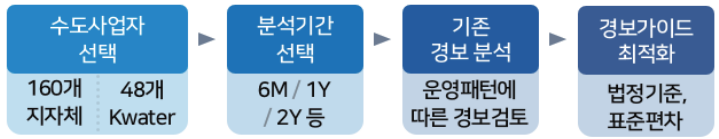
– 모델링

① 운영데이터 분석 및 경보설정을 통한 데이터 기반 감시

▶ 경보시스템 구현

- (경보분석) 통계기반 경보로직 설정(편차, IQR, 중앙값, 사용자 지정 등)

가이드에 따른 지자체 데이터 유형별(수량/수질)별 최적경보 분석



② 영상기반 수도 관로사고 감시

▶ 인공지능 기반 누수사고 영상분석 알고리즘 도입

- Auto DL을 통해 다종 누수사고 검출 AI 모델 학습
- 성능향상을 위한 앙상블* 기반 K-water형 알고리즘 모델 개발

* 여러 모델을 기준으로 학습 결과를 다시 재조합 또는 결합하여 최종 예측 성능을 극대화 하는 기술

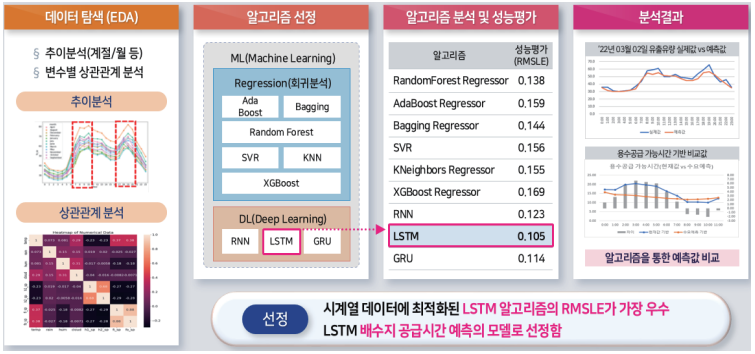
▶ 인공지능 기반 카메라 영상위치 분석

• 카메라 위치변경에 따른 관로위치 감지를 위한 LoFTR 알고리즘 기반 카메라 위치보정 알고리즘 도입

③ 배수지 용수공급 관련 수요예측

▶ 최적 알고리즘 선정

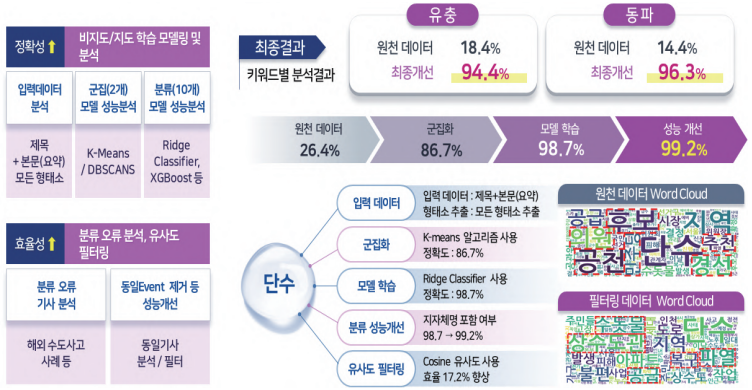
• 시계열 데이터에 최적화된 LSTM 알고리즘의 RMSLE가 가장 우수



④ 수도사고 관련뉴스 분석 및 공유체계 구축

▶ 정확성 효율성 향상을 위한 데이터 분석 수행

• 지도/비지도 학습 모델링 및 분석, 오류 분석 등 수행 후 정확도 최종 94~99% 달성(키워드 : 단수, 유충, 동파)



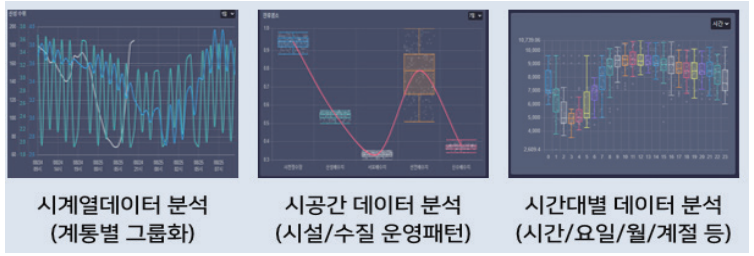
2. 상수도 이벤트 기반 실시간 뉴스정보 분석 및 공유시스템 구축 및 운영('23~)

- 결과 구현

① 운영데이터 분석 및 경보설정

• Python의 matplotlib 등을 활용한 데이터 변화 추이 시각화

• Java를 활용한 분석체계 구현

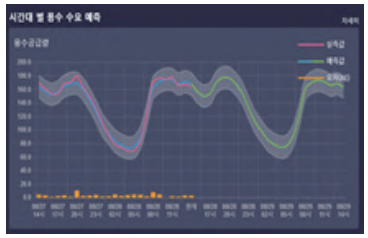


② 영상기반 수도 관로사고 감시

- Java를 활용한 웹서비스 구축
- CCTV 통합 모니터링을 위해 기존 관망 GIS와 연계하여 서비스함

③ 배수지 용수공급 관련 수요예측

- 최적 모델(RMSLE)을 활용한 배수지 운영환경별 수요 예측
- Java를 활용한 웹서비스 구축



④ 수도사고 관련뉴스 분석 및 공유체계 구축

- Java를 활용한 웹서비스 구축
- 실시간 사고뉴스 SMS 전송을 위한 C++ 기반 SMS 자동전송 프로그램 운영

빅보드 상황실 감시

뉴스 SMS 수신 → URL 연계로 뉴스 바로보기 가능

정책활용/기대효과

한국수자원공사는 물공급을 위한 단순한 데이터 모니터링의 한계를 넘어, 데이터 분석기법을 적용하여 실시간 수도사고 대응지원과 예방을 위한 기반을 마련하였다.

또한 국가 차원의 데이터 표준화 및 통합체계를 구축하고, 타 기관 및 정보와의 융합·제공이 용이한 시스템을 구성하였다. 이를 통해 단절되어 있던 광역-지방(지자체) 간 데이터 관리체계를 개선하고, 빅데이터 기반의 분석·활용체계 고도화를 위한 효율적인 데이터 취득체계를 갖추게 되었다.

아울러 국가 차원의 수도사고 대응을 위한 4개 알고리즘을 적용한 분석 서비스를 개발하고, 전국 지자체에 동일한 서비스를 제공할 수 있는 체계를 구축하였다. 해당 서비스는 유역수도지원센터 상황실에서 기술지원에 활용되고 있다.

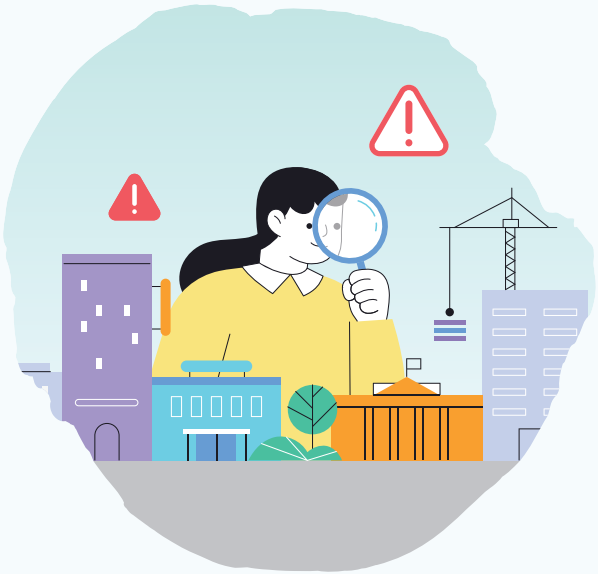
본 과제를 통해 수도사고 발생 시 데이터 기반의 기술지원 체계를 강화하고, 지역별 기술도입 여건과 재원 차이에 따른 물공급 서비스의 격차를 줄일 수 있게 되었다.

K-water는 앞으로도 빅데이터 기반 분석 서비스를 지속적으로 개발·적용하며, 국민이 신뢰할 수 있는 수도서비스를 제공하기 위한 정책을 추진할 것이다.

PART 3-2
재난안전

빅데이터 기반 시설물의 건설부터 유지관리까지 선제적 사고 예방

국토안전관리원



추진목적/배경

최근 시설물 사고 등에 의한 사회적 재난이 지속적으로 발생하고 있다. 최근 5년 간의 대표적인 사고로는 건설 시공(품질) 및 감리가 원인이였던 '2021년 광주 아파트 붕괴 사고', 건설 시공(품질) 문제로 발생한 '2023년 검단 아파트 무량판 붕괴 사고', 점검 및 유지관리 문제로 발생한 '2023년 성남시 정자교 붕괴 사고', '2024년 연희동 지반침하 사고' 등이 있다.

특히 건설공사 현장에서 사고가 발생할 경우 치사율이 상당히 높다. 고용노동부 자료에 따르면, 2024년 업종별 중대재해 사망자 수 중 64%가 건설업 종사자인 것으로 나타났다.

이러한 안전사고 저감을 위한 지속적인 정부 정책 이행을 위해 국토안전관리원은'시설물의 안전 및 유지관리에 관한 특별법'에 따라 시설물의 전 생애주기에 걸쳐 안전관리 및 안전점검을 수행하고 있다.

국토안전관리원은 빅데이터 기반의 선제적 유지관리를 통해 안전점검의 효율성을 높이고 사고 저감에 기여하고자 본 과제를 수행하였다. 주요 현안 과제에 대한 내·외부 평가 및 자문을 거쳐 ①AI 기반 건설공사 현장점검 우선순위 선정 서비스와 ②소규모 취약시설물 안전점검 대상 선정 지원 서비스를 우선 수행 과제로 선정하였다. 이후, 국토안전관리원 빅데이터 분석 프로세스에 따라 이를 추진하였다.

국토안전관리원빅데이터 분석과제 추진전략

기술적 측면	• 정형 · 비정형 **데이터 기반의 의사결정** 지원 • AI · 빅데이터를 통한 데이터 알고리즘 기반의 **공공서비스 강화**
관리적 측면	• 빅데이터 서비스 모델 개발 및 운영을 위한 안전관리 **주기적 요구사항 수렴** 수행 • 국토안전관리 관련 다양한 정보 연계를 통한 데이터 관리
법·제도적 측면	• 데이터 확보 및 빅데이터 기반 의사결정 업무 서비스 등에 필요한 법 · 제도 정비 추진 • 국토안전 관련 법령 개선사항 적용

분석 사전 준비

– 활용 데이터

데이터명	형태	내용	출처	기준 년도	내·외부 데이터
건설공사대장	CSV	공사번호, 공사이름, 공사종류 등	CSI	2023	내부
건설공사대장	CSV	공사ID, 공사명 등	KISCON	2023	외부
건설사고정보	CSV	사고명, 사고일시, 공사명 등	CSI	2023	내부
건설공사업체	CSV	공사번호, 시공업체 등	CSI	2023	내부
소규모 취약시설물정보	CSV	시설물번호, 시설물명 등	SFMS	2023	내부
건축물 제원	CSV	시설물번호, 건물번호 등	SFMS	2023	내부
보수보강 이력	CSV	시설물번호, 보수보강 등	SFMS	2023	내부
자체안전점검 실적	CSV	시설물번호, 점검순번, 점검일자 등	SFMS	2023	내부
현장조사	CSV	시설물번호, 손상코드, 손상원인 등	SFMS	2023	내부
안전점검평가 (상태평가)	CSV	점검일자, 안전등급 등	SFMS	2023	내부
안전조치	CSV	시설물번호, 조치방법, 조치예정일 등	SFMS	2023	내부

분석 과제 추진 과정에서 필요한 내부 데이터는 CSI와 SFMS 시스템 내 자료를 활용하였고, 외부 데이터인 건설공사대장(출처: KISCON)은 정보시스템 간 연계를 통해 수집하였다.

AI기반 건설공사 현장점검 우선순위 선정 서비스는 CSI 및 KISCON의 건설공사 대장, 건설사고 정보, 건설공사 업체 데이터 673,004건을 활용하였다.

소규모 취약시설물 안전점검 대상 선정 지원 서비스는 국토안전관리원 내부 시스템 SFMS의 소규모 취약시설물 정보, 건축물 제원, 보수·보강 이력, 자체 안전점검 실적, 현장 조사, 안전점검 평가(상태 평가), 안전조치 데이터 43,404건을 활용하였다.

분석과정

– 분석 환경

1. 분석 인프라 : 기관 내 분석자원 활용
2. 분석 환경 : Python, Jupyter Notebook

– 데이터 수집

1. 사용 데이터 출처 : 기관 자체 생성 데이터(CSI, SFMS), 외부 데이터 (KISCON)
2. 사용 데이터 형식 : csv

– 데이터 전처리

1. AI기반 건설공사 현장점검 우선순위 선정 서비스

▶ 결측값 및 이상치 처리

- 약 6,000개의 미입력, '없음'등 제거

▶ 범주형 변수 처리

- LeaveOneOutEncoder*, LabelEncoder** 활용

* LOO Encoder : 범주형 데이터 각 샘플의 범주를 제외한 나머지 샘플의 평균 타켓 값을 사용하여 해당 샘플을 인코딩하는 기법

** LabelEncoder : 각 고유한 범주를 고유한 숫자 값으로 매핑하는 기법

- 11개 범주형 변수 인코딩 실시

▶ 오버샘플링

- RandomOverSampler를 통한 약 19:1 비율의 데이터 불균형 해소
- 점검 대상 선정 목적의 공사 진행 구간별 데이터 증강 실시

2. 소규모 취약시설물 안전점검 대상 선정 지원 서비스

▶ 결측값 처리

- 보수조치 미입력 등 7,000개 row 결측치 제거

▶ 파생변수 생성

- 점검 연도, 준공 연도 활용한'경과연수'파생 변수 생성

▶ 오버샘플링

- LabelEncoder를 통한 7개 범주형 변수 인코딩

– 모델링

1. 탐색적 데이터 분석(EDA)

▶ AI기반 건설공사 현장점검 우선순위 선정 서비스

- 사고 발생 현황
  - → 건축공사에서 가장 높은 사고 건수 기록
  - → 민간 부문에서 사고 발생 빈도가 높음

·민간 부문에서 사고 발생 빈도가 높음

- 주요 사고 유형
  - → 넘어짐, 떨어짐, 자재 관련, 거푸집 관련 사고 비율이 높음

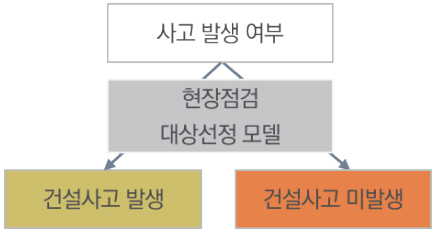
▶ 소규모 취약시설물 안전점검 대상 선정 지원 서비스

- 경과 연수와 등급의 상관관계
  - → 경과 연수가 증가할수록 등급이 낮아짐
- 위험 요인
  - → 보수조치가 없는 경우 C, D, E 등급 비율이 높음
  - → 2개 이상 구조형식 시설물의 경우 C, D, E 등급 비율이 높음

2. AI 기반 건설공사 현장점검 우선순위 선정 서비스

▶ 현장점검 대상선정 우선순위 모델

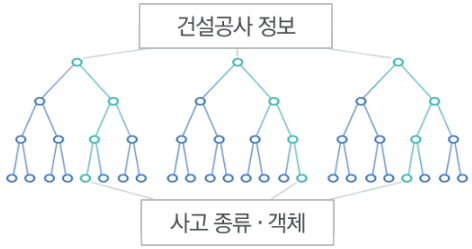
- 비교 모델: XGBoost, LightGBM 등
- 최종 선정: DNN 모델(변수 간 내재된 관계 도출에 우수)
- 성능 평가: AUC Score 0.9717



Evaluation	Value
Accuracy	0.9120
Precision	0.0958
Recall	0.9957
Loss	0.1889
AUC	0.9717

▶ 건설사고 종류 예측 모델

- 비교 모델: Random Forest, DNN, XGBoost, LightGBM 등
- 최종 선정: Random Forest
- 성능 평가: AUC Score 0.932



Evaluation	Value
Accuracy	0.930
Recall	0.921
AUC	0.932

3. 소규모 취약시설물 안전점검 대상 선정 지원 서비스

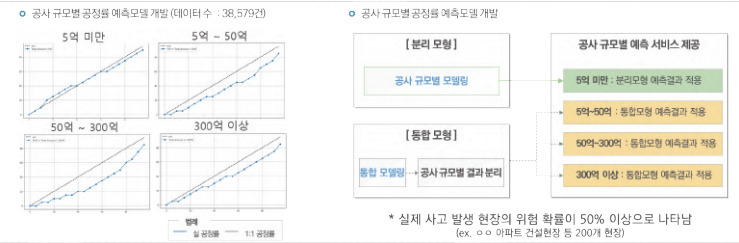
▶ 우선순위 선정 모델

- 비교 모델: Random Forest, SVM, XGBoost 등
- 최종 선정: XGBoost
- 성능 평가: AUC Score 0.950

– 검증 및 고도화

▶ 건설공사 현장점검 우선순위 선정 모델

- 현장별 예측 결과와 실제 현장 비교 검증
- 한계점 파악
  - → KISCON 시스템의 공정률 현행화 의무사항 부재
  - → 공사 규모별 예측 정확도 차이 발생
- 개선 조치
  - → 실제 건설현장 38,579개 실사 진행
  - → 공사 규모별 공정률 예측 모델 개발
  - → 데이터 신뢰성 보완
  - → 규모별 맞춤형 현장점검 대상 선정 모델 개발



공사 규모별 예측모델 개발

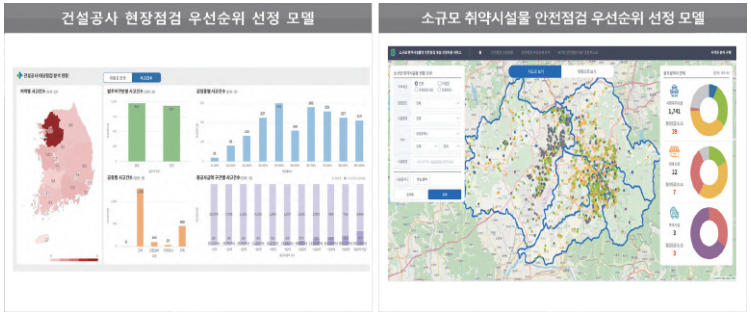
– 결과 구현

▶ 시각화 서비스 제공

- Python 활용 웹 기반 시각화 구현

▶ 소규모 취약시설물 안전점검 대상 선정 지원 서비스

- 안전관리 리포트 카드 제공, 지역별 비교 결과 제공



모델별 시각화 화면

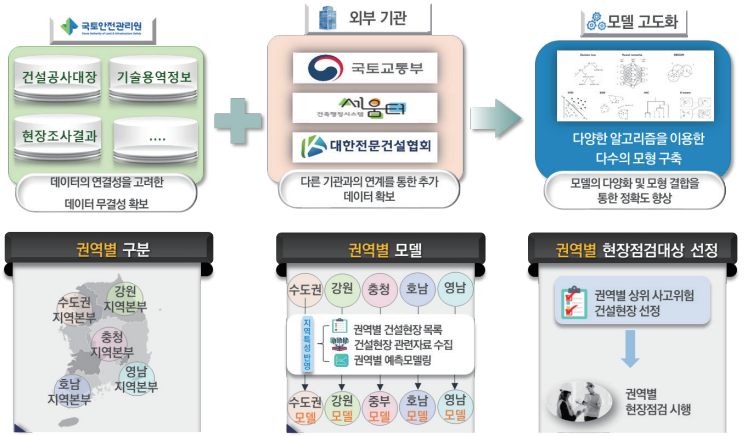
정책활용/기대효과

시설물 및 건설공사 사고 저감을 위해 국토안전관리원에서는 내·외부 데이터를 활용하여 빅데이터 기반의 서비스를 개발하고 활용하였다.

주요 성과는 다음과 같다. 첫째, AI 기반 건설공사 현장점검 우선순위 선정 서비스의 '건설공사 현장점검 우선순위 선정 모델'을 활용하여 국토안전관리원 주관의 현장 맞춤형 컨설팅을 지원하였다. 특히 2023년에는 건설공사 현장 맞춤형 사고 예방 컨설팅을 3,919건 실시하였고, 2024년에는 활용 범위를 확대하여 건설공사 작업자의 경각심을 고취하였다. 5만 개 이상의 현장에 대한 우선순위를 제공하여 건설공사 현장점검 대상 선정을 지원하였으며, 그 결과 전년 동기 대비 건설사고 사망자가 21%(2024년 8월 기준 2023년 141명에서 2024년 111명) 감소하였다.

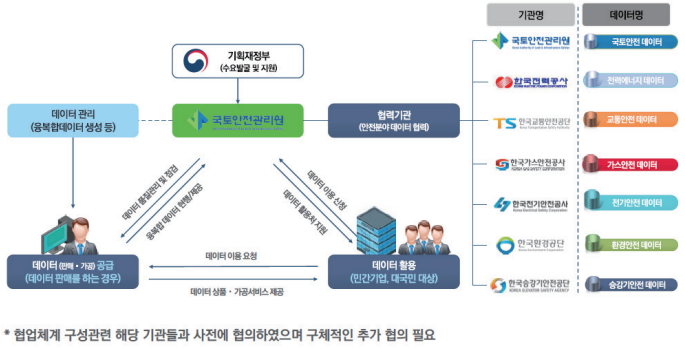
둘째, 소규모 취약시설물 안전점검 대상 선정 지원 서비스의 '소규모 취약시설물 안전점검 우선순위 선정 모델'을 활용하여 8만 개 이상의 현장에 대한 위험도를 예측하고 우선점검 필요 대상을 선정하였다. 2023년에는 선정 결과를 토대로 실제 86건의 현장점검을 실시하여 빅데이터 기반의 안전점검을 수행하였다. 결과적으로 2024년 소규모 취약시설에 대한 안전사고가 0건 발생하여, 선제적 대응을 통한 사망자 제로화라는 성과를 이루었다.

국토안전관리원은 향후 지리적·환경적 요소를 반영한 권역별 맞춤 모델을 개발하여 안전사고 저감을 위해 노력할 것이다. 아울러 우선순위 선정 모델을 시설물 점검과 지반 탐사 등에도 적용하여 위험 요인을 사전에 점검·보완할 계획이다.



권역별 맞춤 모델 개발 체계도

한편, 국토안전관리원은 기획재정부 공공기관 데이터 협업 과제 주관기관으로 선정되어 '대국민 대상 시설물 안전 통합정보 제공 서비스'를 구현하고 2024년 12월부터 제공하고 있다. 본 서비스는 7개 기관의 데이터를 융·복합하여 대국민을 대상으로 통합 안전정보를 제공한다. 이를 통해 안전문화를 확산하고 국민의 편의성을 높일 수 있을 것으로 기대된다.



* 협업체계 구성관련 해당 기관들과 사전에 협의하였으며 구체적인 추가 협의 필요

기획재정부 데이터 협업 과제 체계도

PART 3-3
재난안전

출동데이터를 활용한 골든타임 미확보 구역 특성 분석

부산소방재난본부

추진목적/배경

재난현장의 골든타임 확보를 위한 출동 취약 요인 개선 사업은 시민의 생명과 재산을 보호하고 사회적 비용을 절감하기 위해 추진되었다. 이 사업은 출동 데이터 및 공간 데이터를 분석하여 체계적이고 효과적인 대책 기반을 마련함으로써 재난 현장으로의 출동 소요 시간을 단축하는 데 중점을 두고자했으며 이러한 노력을 통해 재난 대응의 효율성을 극대화하고 골든타임 내 대응을 실현함으로써 시민 안전을 한층 강화하고자 하였다.

사업 추진 과정에서 골든타임 사각지대를 도출하고, 요인별 분석을 통해 도출된 결과를 기반으로 소방 정책과 지자체 협업 정책을 추진하여 기존의 구조적 문제를 보완하고, 안전 사각지대를 최소화 하기 위한 목적이다.

또한, 분석 결과를 지속적으로 모니터링하고 활용하기 위해 시각화 대시보드 생성을 통해 관련 업무 담당자들이 데이터를 손쉽게 공유하고 정책 추진에 활용할 수 있게 하였다.

분석 사전 준비

– 활용 데이터

데이터명	데이터 출처	자료형태	기준연도
화재 출동 정보	부산소방재난본부	CSV	2020년 01월 ~ 2022년 12월
구급 출동 정보	부산소방재난본부	CSV	2020년 01월 ~ 2022년 12월
구조 출동 정보	부산소방재난본부	CSV	2022년 01월 ~ 2023년 06월
소방차량 출동 GPS	부산소방재난본부	CSV	2023년 03월 ~ 2023년 06월
소방차량 진입·출입 곤란지역	공공데이터 – (자체수집)	CSV	2023년
소방용수 및 비상 소화장치함 설치	공공데이터 – (자체수집)	shp	2023년
실폭도로정보	공공데이터 – (자체수집)	JSON, XML	2023년
교차로 교통량정보	공공데이터 – (자체수집)	URL	2023년
교통소통정보	공공데이터 – (자체수집)	URL, XML	2023년 04월
교통돌발정보(행사 정보)	공공데이터 – (자체수집)	CSV, XML, JSON, TXT	2022년 01월 ~ 2023년 06월
보호구역정보	공공데이터 – (자체수집)	JSON	2023년
건물통합정보	공공데이터	shp	2023년

분석과정

– 분석 환경

1. 분석 인프라 : 기관 내 PC 사용, 시각화 시스템 내부서버
2. 분석 환경 : PostgreSQL, tableau

도입 대상 모듈

– 데이터 수집

사용 데이터 출처(기관 자체 생성 데이터 및 공공데이터 등 자체수집)

사용 데이터 형식(CSV, XML, SHP, JSON 등)

– 데이터 전처리 (변수 선정, 중복·이상·결측치 처리 등)

〈부산소방재난본부 데이터 변환 결과〉

▶ 국가 화재정보 정보

- 텍스트 기반 주소 데이터 위경도 좌표로 변환을 위한 주소형식 통일화
- 주소데이터 위경도 좌표로 변환
- 변환 불가 지점 및 부산시 이외 좌표 삭제 처리

국가 화재정보 정보				
단계	데이터수	제외 데이터	전처리 구분	전처리 내용
1	8821	0	파생변수 생성	주소 컬럼 생성
2		0	형식통일화	주소 형식 통일화
3		0	Geo Coding	위경도 좌표 변환
4		1270	행제거	위경도 좌표 변환 불가 지점 삭제
5				
전처리결과	8821	1270	7551	

국가 화재정보 데이터 전처리 결과

▶ 도로 정보

- 전국 단위 도로정보에서 부산시 도로 추출

도로정보				
단계	데이터수	제외 데이터	전처리 구분	전처리 내용
1	544457	527900	행제거	부산시 이외 정보 삭제
2				
3				
4				
5				
전처리결과	544457	527900	16557	

도로정보 데이터 전처리 결과

▶ 네비게이션 정보

- 도로-네비게이션 정보에서 부산시 이외 도로삭제 처리
- PGIS 경로탐색 알고리즘*에 사용하기 위한 시점, 종점, 비용 데이터 생성

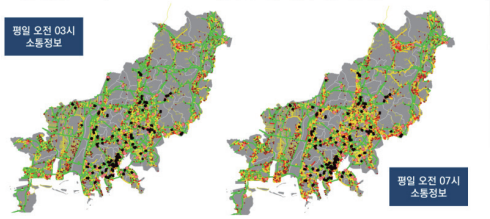
* PGIS 경로탐색 알고리즘 : 확률적 모델을 기반으로 하여, 불확실한 환경(교통량, 도로 상태, 기상 조건 등)에서 최적의 경로를 찾는 알고리즘

네비게이션 정보				
단계	데이터수	제외 데이터	전처리 구분	전처리 내용
1	99000	0	행제거	부산시 이외 정보 삭제
2	99000	0	컬럼 변경	테이블 구조 변경
3				
4				
5				
전처리결과	99000	0	99000	

네비게이션 데이터 전처리 결과

▶ 5분 단위 도로소통정보

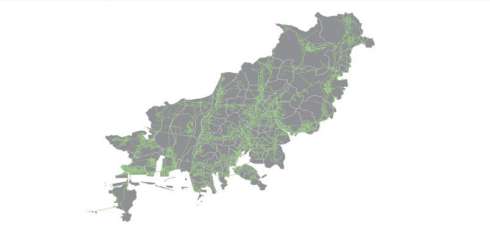
- 전국 5분단위 소통정보에서 부산시 이외 데이터 제거
- 링크별로 5분단위 소통정보에서 1시간단위 소통정보 생성
- 소통정보 이용하여 구간별 통행시간 생성

5분 단위 도로소통정보		
단계	전처리 구분	전처리 내용
1	행 제거	Link_id 체계 기반, 부산 지역 이외 데이터 제거
2	파생변수 생성	시간대별 평균 통행속도 생성 – 5분 단위 소통정보 이용
3	파생변수 생성	통행 시간정보 생성 – 통행속도와 링크 길이 이용
전처리결과		

5분 단위 도로소통정보 전처리 결과

▶ 표준노드링크

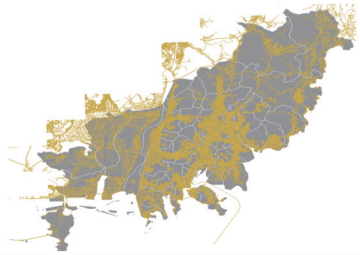
- 전국단위 표준노드링크에서 부산 이외 정보 삭제

표준노드링크		
단계	전처리 구분	전처리 내용
1	행 제거	Link_id 체계 기반, 부산 지역 이외 데이터 제거
전처리결과		

표준노드링크 데이터 전처리 결과

▶ 내비게이션 링크

내비게이션 링크

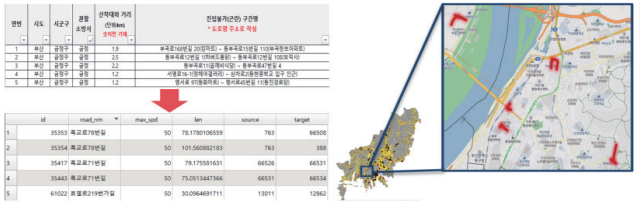
단계	전처리 구분	전처리 내용
1	파생변수 생성	pg_routing의 pg_createTopology() 사용하여 노드 생성
전처리결과		

내비게이션 링크 데이터 전처리 결과

▶ 소방차 진,출입 곤란지역

- 수기 데이터로 관리되고 있는 데이터 데이터베이스화 진행
- 텍스트 기반 주소 데이터 위경도 좌표로 변환을 위한 주소형식 통일화
- 주소 좌표변환
- 진입지점, 진출지점 구분하여 경로 탐색 알고리즘 적용하여 진출입 곤란 지역 경로 생성

소방차 진,출입 곤란지역

단계	전처리 구분	전처리 내용
1	파생변수 수정	비정형 주소 데이터를 특정 형식으로 변환
2	파생변수 생성	주소 형식 데이터 생성
3	파생변수 수정	주소 좌표 변환
전처리결과		

소방차 진,출입 곤란지역 데이터 전처리 결과

- 모델링/구현

사용한 분석 모델 : 로지스틱 회귀분석(Logistic Regression)

① 생존 분석(Survival Analysis) 모델

② 로지스틱 회귀분석(Logistic Regression)

③ 랜덤 포레스트(Random Forest)

④ 그래디언트 부스팅(Gradient Boosting)

▶ 활용 데이터 셋

- 출동 소요시간에 영향을 미치는 분석을 수행하기 위해, 부산소방재난본부 및 기타 협의를 통해 수집한 데이터를 기반으로 영향 요인 분석 데이터 셋을 구축함

① 사건발생 연/월/일 : 화재가 발생한 년/월/일을 분석 데이터로 구축

② 안전센터 거리 : 화재정보시스템의 안전센터 거리 데이터 활용

③ 날씨 상태 : 화재정보시스템의 날씨 데이터 활용

④ 불법주정차수 : 안전신문고 불법주정차 신고 데이터를 활용
불법주정차 신고 장소와 도로링크를 매칭하여 불법주정차 다발구간 생성

⑤ 도로 소통정보 : 표준노드링크의 소통정보를 활용하여 데이터 구축
소통정보는 국가 교통정보센터 기준을 따름

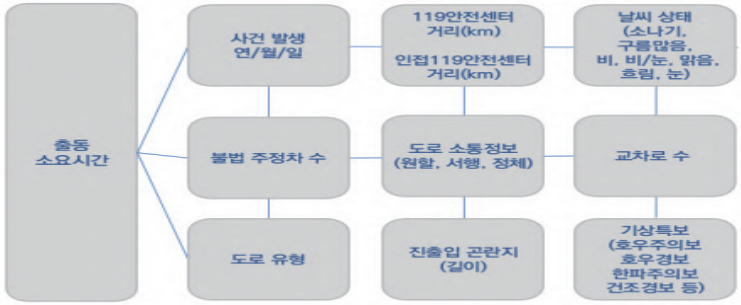
⑥ 교차로 수 : 표준노드링크의 노드정보를 활용하여 교차로 데이터 구축

⑦ 도로 유형 :표준노드링크와, 표준노드링크상에 없는 세부도로로 구분하여 구축

⑧ 진출입 곤란지역 : 소방재난본부에서 관리하는 진출입 곤란지역 활용

⑨ 기상특보 : 화재정보 시스템의 기상특보 데이터 활용

- 수집한 데이터셋의 모든 요인을 빅데이터분석 모델에 적용하여 분석 진행
- 각 모델별 결과를 도출하여 모델 성능 평가
- 골든타임 미확보 지역 특성분석에 관한 최적 모델 선정
- 선정된 모델의 결과를 활용하여 골든타임미확보 요인



분석 데이터 셋

▶ 생존분석 – Cox 비례 위험모형* (누적 데이터 전체)

* Cox 비례 위험모형 : 생존 분석에서 특정 사건의 발생 위험에 영향을 미치는 변수들을 분석하는 통계 모델로, 각 변수의 비례적 위험 비율을 추정하여 시간에 따른 위험을 평가함

- 2020~2023.06 전체 데이터 7551건을 활용하여 생존분석 모델로 분석.
- 생존분석 cox 비례위험모델로, 3년치 화재출동데이터를 기반으로 분석할 때 가장 유의미하게 나온 변수는 (p-value 0.05 이하) "119안전센터거리(km)"로 나타남
- 부산 소방재난본부의 출동 시간을 기준으로 생존 시간에 영향을 미치는 여러 요인을 분석하기 위해 사건이 발생한 시간과 장소를 기반으로 데이터를 연결(매칭)하려 했는데, 이 연결이 제대로 이루어지지 않는 문제로분석 모델에 영향을 미쳐서 결과의 정확도가 54%로 낮게 도출됨

생존분석 – 누적 데이터 전체

Feature	coef	exp (coef)	se (coef)	coef lower 95%	coef upper 95%	LPB4(p)
119안전센터거리	-0.048	0.953	0.021	-0.089	-0.007	5.608
정체	-0.004	0.996	0.005	-0.014	0.007	1.015
불법주정차 밀집	0.000	1.000	0.000	0.000	0.000	0.828
화재발생(월)	0.006	1.006	0.014	-0.021	0.033	0.584
원할	0.003	1.003	0.009	-0.014	0.020	0.462
서행	0.002	1.002	0.005	-0.009	0.012	0.371
화재발생(일)	0.001	1.001	0.005	-0.009	0.012	0.350
화재발생(년)	0.009	1.009	0.045	-0.079	0.097	0.236
화재발생(시)	-0.001	0.999	0.008	-0.017	0.014	0.198
모델 성능 평가 지수				Concordance = 0.54		
				Partial AIC = 5371.10		
				log-likelihood ratio test = 8.54 on 10 df		
				-LOG2(p) of ll-ratio test = 0.8		

▶ 생존분석 – Cox 비례 위험모형 (GPS 매칭 데이터)

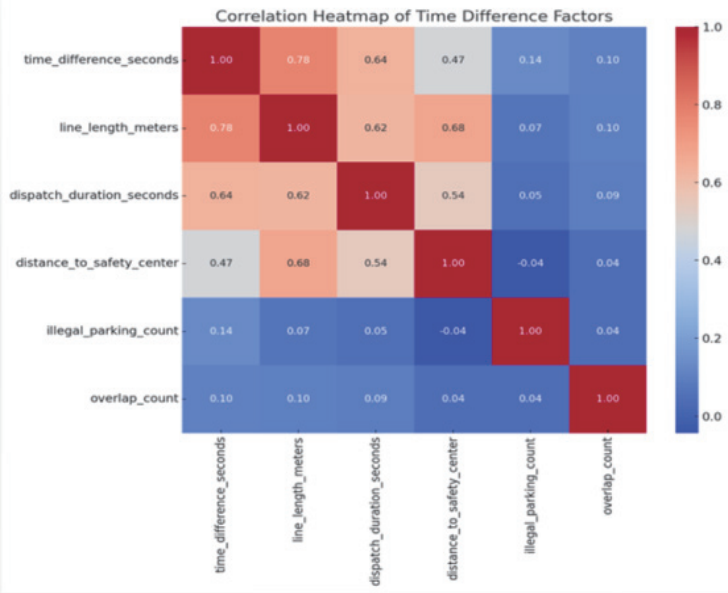
- GPS궤적 데이터와 매칭이 가능한 2023.03~06월 848건의 데이터를 활용하여 생존분석 모델로 분석.
- GPS 매칭 데이터로 생존분석 분석시, 분석 과정에서 의미 없다고 판단되는 데이터가 다수 제외되어 "불법주정차수", "진입곤란지역", "119안전센터" 3개 요인만 생존분석에 활용되어 결과가 도출되었음
- 해당 데이터에서도 유의미한 변수 (p-value 0.05 이하)로 "119안전센터거리"가 주요 요인으로 해석됨

생존분석 – Cox 비례 위험모형 3년치 화재출동데이터 기반 분석

Feature	coef	exp (coef)	se (coef)	coef lower 95%	coef upper 95%	LPB4(p)
불법주정차수	-0.07841	0.92459	0.08554	-0.24607	0.08925	0.78187
진입곤란지역	0.01448	1.01458	0.15339	-0.28616	0.31511	0.75114
119안전센터거리	-0.00100	0.99990	0.00005	-0.00019	-0.00001	0.99981
모델 성능 평가 지수				Concordance = 0.54		
				Partial AIC = 910.37		
				log-likelihood ratio test = 3.68 on 4 df		
				-LOG2(p) of ll-ratio test = 1.15		

▶ 데이터 상관분석

- 화재출동데이터와 GPS 매칭 데이터로 상관분석 수행 시에도, 안전센터와의 거리가 0.77로 양의 상관관계를 보이는 것으로 해석되었으며, 불법주정차수와 소방차 진입곤란지역의 변수가 낮은 상관계수를 보임 119안전센터 거리가 가장 상관성이 높다고 제시된 분석 결과로 나옴에 있어, 3년치 화재출동데이터를 기반으로 생존분석보다 단순하게 종속변수를 "골든타임 미확보(1)" / "골든타임 확보(0)"로 두고, 해당 종속변수에 영향을 미치는 요인 분석을 수행하고자 로지스틱 회귀분석을 수행함



상관분석 화재출동데이터 - GPS 매칭 데이터 분석

▶ 랜덤 포레스트

- 대량의 데이터를 모아서 기계학습(머신러닝)을 통해 분석을 하면서 특히, 사건별로 시간과 장소를 매칭하면서 불법 주정차가 얼마나 몰려 있는지와 소방차가 드나드는 경로를 살펴보았을 때 데 분석 결과에서 의미 있는 값이 나오지 않았고, 다만 총 이동 거리가 크게 영향을 준다는 점이 확인되어서 결국 이 분석 모델의 정확도는 31%로 낮게 나와 모델 성능이 낮았음

랜덤포레스트

no	Feature	Coefficient	no	Feature	Coefficient	no	Feature	Coefficient
1	링크길이합	0.334	11	화재발생(년)	0.023	20	강풍 주의보	0.000
2	119안전센터거리	0.162	12	비	0.004	21	강풍 경보	0.000
3	불법주정차 밀집	0.125	13	맑음	0.003	22	대설 주의보	0.000
4	정체	0.068	14	흐림	0.002	23	풍랑 주의보	0.000
5	119안전센터	0.054	15	구름많음	0.001	24	건조 주의보	0.000
6	서행	0.054	16	소나기	0.000	25	풍랑 경보	0.000
7	화재발생(일)	0.053	17	비/눈	0.000	26	태풍 주의보	0.000
8	화재발생(월)	0.046	18	눈	0.000	27	건조 경보	0.000
9	화재발생(시)	0.046	19	태풍 경보	0.000	28	호우 경보	0.000
10	원할	0.027	20	강풍 주의보	0.000	29	호우 주의보	0.000

모델 성능 평가 지수	MSE = 18405.747
	R_squared= 0.311417569

▶ Gradient Boosting Machines

- Gradient Boosting 모델의 모델 성능을 평가하는 지수는 "MSE" / "R-squared"로 평가할 수 있음. 그 중, MSE 값은 크기나 작다는 절대적인 수치가 아니며 해당 수치는 다른 비교모델의 MSE가 상대적으로 낮으면 모델이 좋다고 할 수 있음. 분석결과에서 도출된 요인인 "링크길이합" "119안전센터거리" "외에도 "불법주정차 밀집" 요인과 "도로소통정보 정체" 요인이 도출되어 도로 환경적 요인이 출동소요시간에 영향을 주는 요인으로 도출됨을 확인함. 하지만,

비교적 R-squared(1가까운 수치는 좋은 성능)이 낮게 나오게 되어, 추가 데이터 가공 후 재분석 해보고자 함.

- 부산 소방재난본부 출동 데이터특성을 분석한 결과 유의미한 파라메터는 추출이 되나, 이에 대한 모델 검증 시 정확도 26.67%로 실제 활용하기에는 적합하지 않은 모델로 판단함.

Gradient Boosting Machines

no	Feature	Coefficient	no	Feature	Coefficient	no	Feature	Coefficient
1	링크길이합	0.404	11	원할	0.007	21	비/눈	0.000
2	119안전센터거리	0.333	12	흐림	0.001	22	태풍 주의보	0.000
3	불법주정차 밀집	0.130	13	비	0.000	23	대설 주의보	0.000
4	정체	0.046	14	건조 주의보	0.000	24	건조 경보	0.000
5	화재발생(년)	0.020	15	소나기	0.000	25	구름많음	0.000
6	서행	0.015	16	강풍 주의보	0.000	26	호우 경보	0.000
7	화재발생(월)	0.014	17	강풍 경보	0.000	27	호우 주의보	0.000
8	119안전센터	0.013	18	태풍 경보	0.000	28	눈	0.000
9	화재발생(시)	0.009	19	풍랑 주의보	0.000	29	맑음	0.000
10	화재발생(일)	0.008	20	풍랑 경보	0.000	30	한파 주의보	0.000
모델 성능 평가 지수		MSE = 19600.41661						
		R_squared= 0.266723405415815						

▶ 로지스틱 회귀분석

- 로지스틱 회귀분석 결과, 모델 성능을 평가하는 지수인 Accuracy (정확도)가 약 91.2%로 성능은 가장 좋음.다만, 각 요인들의 영향도를 보다 정확하게 분석하기 위해 데이터 셋 재구축 후 해당 모델을 재분석 해보고자 함. 요인 중, "119안전센터거리"와 "링크길이합 (최적경로 거리 합)" 두 개의 독립변수가 다중공선성의 문제가 있어, 이후 재분석시 1개 변수만 활용하고자 함.

로지스틱 회귀분석

no	Feature	Coefficient	no	Feature	Coefficient
1	119안전센터거리	0.401	16	풍랑 주의보	0.000
2	비	0.011	17	태풍 경보	0.000
3	119안전센터	0.011	18	강풍 경보	0.000
4	정체	0.008	19	강풍 주의보	0.000
5	서행	0.007	20	불법주정차 밀집도	0.000
6	구름많음	0.003	21	소나기	0.000
7	링크길이합	0.000	22	눈	0.000
8	호우 주의보	0.000	23	비/눈	0.000
9	호우 경보	0.000	24	화재발생(시)	-0.002
10	한파 주의보	0.000	25	화재발생(일)	-0.002
11	건조 경보	0.000	26	화재발생(년)	-0.002
12	대설 주의보	0.000	27	흐림	-0.005
13	태풍 주의보	0.000	28	맑음	-0.007
14	풍랑 경보	0.000	29	화재발생(월)	-0.011
15	건조 주의보	0.000	30	원할	-0.014
모델 성능 평가 지수		Accuracy = 0.913			

다) 분석 모델 선정

▶ 로지스틱 회귀분석 모델 선정

- 로지스틱 회귀분석 모델을 사용했더니, 91.2%라는 높은 정확도로 다른 분석 모델들은 결과 정확도나 "R-squared(결정계수-모델이 주어진 데이터를 얼마나 잘 설명하는지)" 값이 낮게 나왔지만, 로지스틱 회귀는 안정적인 결과를 보여줬으며 변수들 간의 관계를 해석하기 쉬운 장점이 있어서, 다른 모델에 비해 분석 결과를 이해하고 설명하기가 훨씬 편리하고 정확도도 높고 결과를 해석하기도 쉬운 믿을 만한 모델임
- 다중공선성 문제로 제외된 변수를 최적화하여 모델을 재평가함으로써, 더욱 정확하고 해석 가능한 결과를 얻을 수 있을 것으로 기대됨
- 최종적으로 로지스틱 회귀분석 모델을 이용하여 상세 데이터 분석에 활용하기로 함

〈화재 출동 골든타임 미확보 요인 분석〉

▶ 1차 분석 독립변수 설정

- 사건발생 연/월/일, 119안전센터 거리, 날씨 상태, 불법주정차 수, 진출입곤란지, 도로소통정보, 기상특보 데이터 중 연관성이 거의 없거나 데이터가 부족하여 의미있는 분석이 불가능한 데이터를 제거 후 1차 독립변수 선정함.

독립변수 설정

▶ 1차 로지스틱 회귀분석 결과

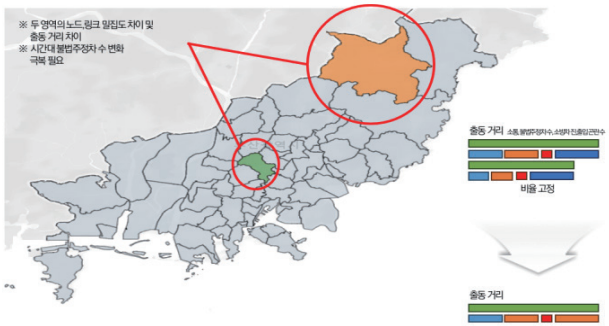
로지스틱 회귀분석 결과

	Feature	Coefficient
1	119안전센터-사고지점 거리	0.401394309
2	비	0.011152081
3	119 안전센터	0.010650447
4	정체	0.008138753
5	서행	0.006589163
6	구름 많음	0.003064542
7	링크길이 합	0.000322242
8	불법주정차 밀집	0.000310361
9	소나기	0.000306419
10	눈	0.000258679
11	화재발생(시)	-0.001691436
12	화재발생(일)	-0.001994675
13	화재발생(년)	-0.002112219
모델 성능 평가 매트릭스		
accuracy	0.912	

▶ 분석 내용

- 부산 소방재난본부 출동데이터 기반으로 최적경로 계산 후 먼저 최적경로 거리를 100% 기준으로 잡고 이 기준으로 소방차 이동효율을 분석하고 추가적으로 시간대별 불법주정차 건수를 반영
- 정확도는 91.2%로 가장 성능이 좋으나, 요인의 영향도 확인 어려움
- 119 안전센터 거리와 링크 길이의 합 두개의 독립변수가 다중공선성의 문제가 있어, 이를 1개의 변수로 활용할 필요성 확인

▶ 1차 독립변수 설정 문제



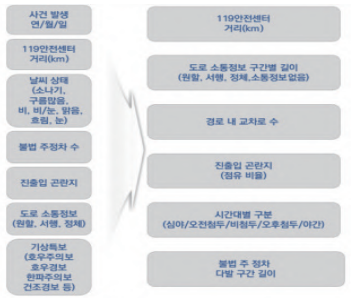
로지스틱 회귀분석 내용

▶ 분석 내용

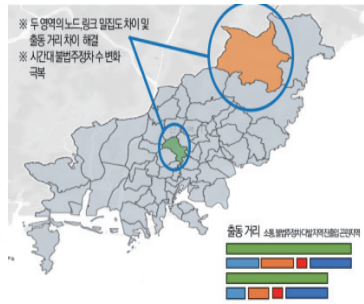
- 초기 데이터는 119 안전센터 별 골든타임 미확보 (7분 이상)을 초래하는 위험지수를 산정하기 위하여 7분을 기준으로 비율(100%) 형태로 구성
- 도로소통정보를 원활, 서행, 정체구간 길이를 데이터로 추출하고 분석하다 보니 왜곡을 발생시킬 수 있는 요인들이 확인되었음
- 작은 도로에서는 소통상태가 항상 원활하다고 나올 가능성이 커서 실제상황보다 소통이 원활하다는 잘못된 결과가 나올 수 있음
- 최적경로와 진출입 곤란지역 매칭에서도 결과가 많거나 적게 나오는 것 자체가 왜곡의 원인이 될수 있어서 주의가 필요
- 또한, 시/공간 분석을 위하여 실제 출동 발생 시간대를 기준으로 출동 경로에 해당하는 불법주정차 수, 소방차 진,출입 곤란지역을 매칭하였으나 해당 시간대의 출동정보와 매칭되지 않는 정보가 많아 결과적으로 유의미한 값이 산출되지 않음

▶ (n)차 독립변수 문제해결 가공 값 설정

- 119안전센터 거리의 연관성이 가장 높아 다른 요인의 연관성 확인을 위해 안전센터 별로 그룹핑 하여 분석 해보았으나 그룹핑시 결과를 뽑기위한 데이터 수량 부족으로 의미있는 결과를 얻기가 어려움이 있음
- 도로 소통정보 원할/서행/정체/소통정보없음으로 세분화
- 도로 소통정보 출동경로 상 원할/서행/정체/소통정보없음의 거리의 비율로 값 변경
- 진출입 곤란지역의 매칭건수에서 전체 경로상 진출입곤란지역의 거리의 비율로 변경
- 불법주정차 신고데이터와 출동정보의 시공간 매칭에서 불법주정차 다발구역(누적 3년간 50건 이상의 불법 주정차 신고된 지점)을 지정. 불법주정차 다발구역 통과 비율 값으로 변경



(n)차 독립변수 문제해결 가공 값 설정



로지스틱 회귀분석 결과

▶ 로지스틱 회귀분석 결과

로지스틱 회귀분석 결과

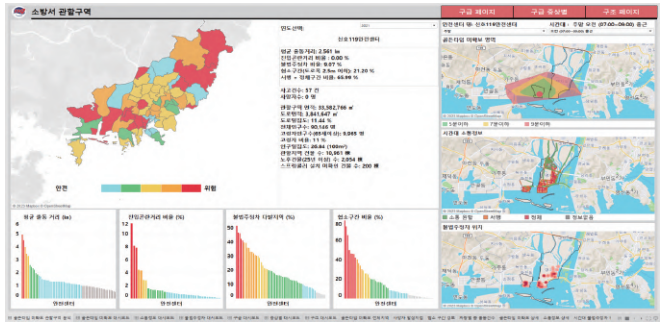
	Feature	Coefficient
1	119안전센터-사고지점 거리	0.645446232
2	서행 구간 길이	0.328603269
3	최적경로 거리	0.302998962
4	정체 구간 거리	0.292163387
5	소통정보 없음 구간 거리 (30km 최속)	0.148528889
6	불법주정차 다발구간 거리	0.13135424
7	09~17시 출동	0.072530986
8	비	0.059679081
9	07~09시 출동	0.057535867
10	소방차 진입곤란지역	0.014475863
11	17~19시 출동	-0.0121547351
12	22~07시 출동	-0.02284297
13	19~22시 출동	-0.055503351
14	원활 구간 길이	-0.164572114
모델 성능 평가 매트릭스		
accuracy	0.912	

▶ 분석 내용

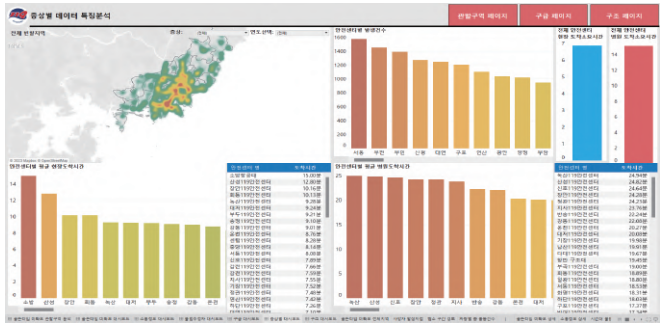
- 부산 소방재난본부 출동 데이터의 약 83%가 7분 내 도착
- 데이터 분포가 7분내 출동에 많아서 7분이상 걸린 출동 데이터 양이 적어 정확도가 낮아지거나 결과에 차이가 생길수 있어 분석이 어려웠음
- 최종 로지스틱 회귀분석을 통한 총 정확도 90.3%의 정확도를 보임.
- 분석된 요인별 Coefficient를 odds ratio로 변환하여 골든타임 안전지수에 활용

- 시각화 :

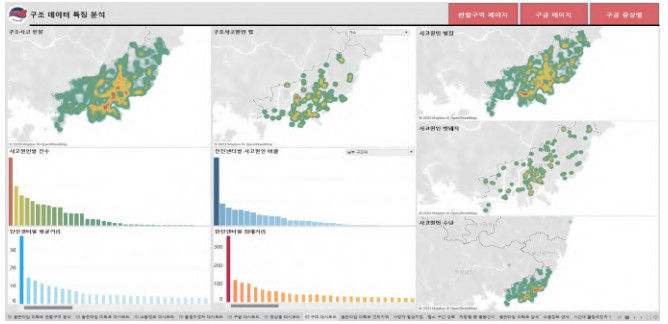
시각화에 활용한 도구 기입 - 태블로



화재 출동 유형 화면



심정지, 중증외상, 심뇌혈관 질환 화면



구조 출동 유형 화면

정책활용/기대효과

부산 소방재난본부는 소방차 진입 곤란 지역 개선을 위해 협소도로, 급경사도로 등 소방차의 진출입이 어려운 지역에서 발생하는 출동 제약 문제를 해소하고자 노력하였다. 이를 위해 지자체와 협의를 통해 개선 방안을 마련하였으며, 그 결과 부산진구 내 3개 이면도로의 진입 곤란 문제가 해소되었다.

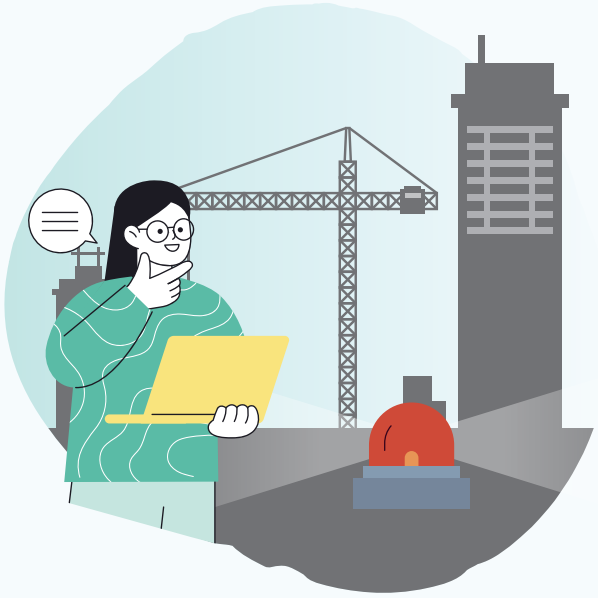
또한, 골든타임 사각지대 해소를 위해 빅데이터 분석을 활용하여 소방차 진출입이 어려운 지역 중 인명 피해가 자주 발생하는 지역을 우선적으로 고려하였다. 이를 기반으로 비상소화장치 4개소를 신규로 설치하였으며, 부산진구 1개소, 사하구 1개소, 기장군 2개소에 설치함으로써 사각지대 문제를 해결하고 긴급 상황 대응력을 강화하고자 하였다.

본부는 화재 등 출동 정보를 분석하여 출동 시 골든타임을 저해하는 요소를 확인하고 이를 제거할 수 있는 방안을 마련하기 위한 분석을 처음으로 진행하였다. 이를 통해 특정 지역에서의 미확보된 곳의 제한점이나 장애물을 파악하고, 출동로 확보를 위한 개선 방향을 모색할 수 있는 중요한 과제를 수행하였다. 이러한 노력을 바탕으로 향후 출동로 확보와 관련한 보다 구체적인 개선 방안을 수립할 수 있는 기반을 마련할 수 있었고 거시적 및 미시적인 효과를 들여다 볼 수 있는 좋은 경험이 되었다.

산재정보 분석을 통한 재해안전지수 개발

데이터의 힘으로 산재보상과 산재예방의 선순환 구조를 만들어 가다

근로복지공단



추진목적/배경

우리 사회는 산업재해 예방과 감소를 위해 다양한 방법으로 근로자의 안전한 일터 문화를 조성하려는 노력을 지속하고 있다. 그러나 산업·노동환경의 급격한 변화로 새로운 업종과 근로형태가 등장하면서 예상치 못한 위험이 발생하고, 이로 인한 국내 산업재해 발생 건수는 꾸준히 증가하고 있다.

* 최근 3년간 산재보험 산재신청 33.0% 급증
('20년)147,512건 → ('23년)196,206건

근로복지공단은 지난 30년간 산재보상 업무를 직접 수행하며 축적된 막대한 산재보상 처리 데이터를 보유하고 있다. 과거 산재보상 데이터가 산업재해 발생 후 진행된 보상 결과의'기록'에 그쳤다면, 현재는 데이터를 통해 지급 패턴과 재해발생 형태를 구분하고 위험 사업장의 형태와 환경적 요인의 특성을 파악함으로써 그 의미가 크게 확장되었다.

공단은 축적된 데이터를 분석하여 유사 및 인접 산업에 내재된 산재 위험을 정보로 전환하고, 이를 구조적·직관적으로 쉽게 이해할 수 있도록 시각화 및 지수화하여 지자체와 사업장에 공유하기 위해 본 모델을 개발하였다.

본 모델을 통해 산업재해 예방의 필요성과 경각심을 갖는 사회적 분위기가 조성되고 안전과 예방이 강화되어 모든 근로자들이 일터에서 안심하고 일할 수 있는 사회가 실현되길 기대한다.

분석 사전 준비

– 활용 데이터

데이터명	형태	내용	출처	기준년도	내·외부 데이터
최초요양신청서 처리 현황(2017~2022)	CSV	• 사업장 주소 : 지오코딩을 통해 좌표값 형성 • 지수산정 : 연령대, 직종, 외국인 등을 통해 지수 산정 • 직종별 데이터 정리 : 직종코드를 활용해 직종데이터 정리	근로복지공단	2017~2022	내부
재활서비스 이용 현황 (2017~2022)	CSV	• 취업자 기본정보 : 이름, 나이, 주민등록번호 • 취업정보 : 취업기관명, 업종, 주소, 취업일자 등 • 중증도 지수 : 사업장의 산재 위험도를 나타냄	근로복지공단	2017~2022	내부
적용사업장 (2017년 이전 성립사업장)	CSV	• 사업장 주소 : 사업장소재지를 활용하여 좌표값을 형성한 이 후 산재미발생 지역에 대한 값으로 활용	근로복지공단	2017년 이전	내부
적용사업장 (2017년 이후 성립사업장)	CSV	• 사업장 주소 : 사업장소재지를 활용하여 좌표값을 형성한 이 후 산재미발생 지역에 대한 값으로 활용	근로복지공단	2017년 이후	내부
기상데이터(ASW)	CSV	• 일별 날씨정보를 산재데이터와 매핑하여 분석	기상청	2017~2022	외부
외국인근로자 근무현황	CSV	• 시군구(법정동) 업종별 외국인근로자 수	EIS 고용행정 통계	2022	외부
지역별고용조사 데이터	CSV	• 시군구(행정동) 업종별 근로자수, 성별, 연령	마이크로 데이터	2022	외부
시군구/성/연령별 취업자(근무지기준)	CSV	• 시군구(행정동) 업종별 근로자수, 성별, 연령	통계청	2022	외부
전국 그리드 (1KM)	SHP	• 전국 그리드 1KM 단위	SGIS	-	외부
시군구 행정경계 (행안부)	SHP	• 법정동 행정경계, 시군구명, 시군구코드	공간정보 포털	2022	외부
시군구 행정경계 (통계청)	SHP	• 행정동 행정경계, 시군구명, 시군구코드	SGIS	2021	외부

공단 내부에서 협조를 받아 데이터를 수집하고, 외부데이터는 공공데이터를 활용하였다.

분석과정

– 분석 환경

1. 분석 인프라 : 기관 내 PC 이용
2. 분석 환경 : Python, QGIS 등

– 데이터 수집

1. 사용 데이터 출처 : 기관 자체 생성 데이터 및 공공데이터
2. 사용 데이터 형식 : CSV, SHP 등

* PoC 단계에서의 데이터 확보 방안 및 분석 환경으로, 실제 구축 시 이와 다를 수 있음

– 데이터 전처리

1. 데이터 정제 및 융합

▶ 내부데이터 전처리

- 적용 사업장 데이터
  - → 관리번호/산재업종 null 제거
  - → 데이터 타입 변경
  - → 특근자 이진분류(해당있음, 해당없음)
  - → 고용상시인원 null값 0처리
  - → 주요 컬럼 정의(관리번호, 사업장 소재지, 산재업종코드, 산재업종명 등)
- 최초요양신청 데이터
  - → 관리번호 null 제거
  - → 데이터 타입 변경
  - → 외국인 여부 null값 내국인으로 변경
  - → 고용형태명 null값 제거 후 정규직, 비정규직 형식 통일
  - → 주요 컬럼 정의(원부번호, 관리번호, 재해당시연령, 외국인여부 등)

• 재활서비스 현황 데이터

→ 중증도 지수/고용형태명 null 제거 및 형식 통일

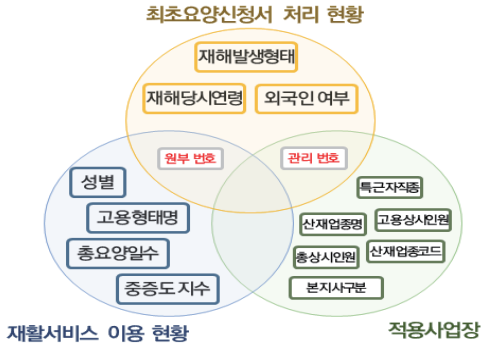
→ 업종코드 null 제거

→ 외국인 여부 null값 내국인으로 변경

→ 원부번호 데이터 타입 변경

→ 주요 컬럼 정의(연도, 원부번호, 생년월일, 재해일자, 성별, 중증도지수 등)

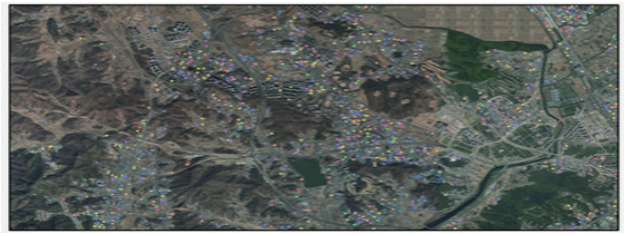
▶ 원부번호와 관리번호를 Key값으로 사용하여 데이터 융합 및 필요 데이터 추출



산재 데이터 주요 속성

2. 적용사업장 지오코딩

사업장 정보('17~'22년)는 지오코딩 과정을 거쳐 공간정보화하여 사용



사업장 지오코딩 결과

– 모델링

1. 산재 유형별 안전지수 개발 프로세스

▶ 분석 모델 및 방법론 개요

• 산재 유형마다 중증도, 빈도, α지수(재해취약자지수) 지수를 산출

• 근로복지공단에서 사업장 단위의 데이터를 적용하면 해당되는 지수가 산출될 수 있는 프로세스 제공

▶ 안전지수

• 산재 이력을 기반으로 산재유형별 중증도, 빈도 지수 모델의 학습데이터 구축

• 머신러닝 기반의 분류 모델을 적용하여 중증도 빈도 지수 산출

• 안전 지수는 중증도 지수와 빈도 지수를 기반으로 산출

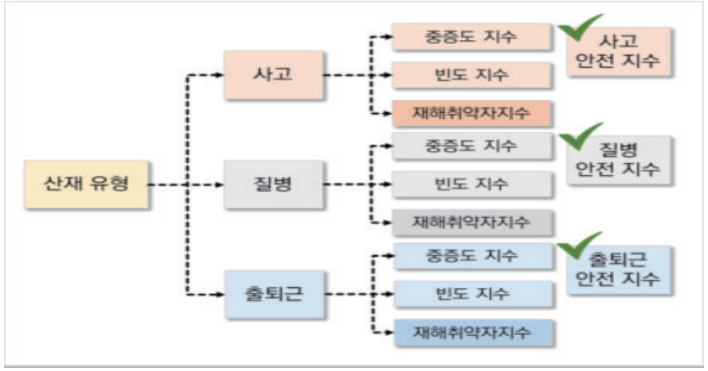
• 적용 범위는 사업장 단위로 적용

▶ α지수(재해취약자 지수)

• 사업장마다 연령대, 성별, 고용형태, 외국인 여부를 파악하기 힘들기 때문에 외부데이터에서 제공되는 시군구형태의 데이터를 적용

• 각 변수 마다 산재와의 상관관계 및 비율을 통해 재해취약자 지수 산출

• 적용 범위는 시군구 단위로 적용



산재 유형별 안전지수 개발

2. 중증도 지수

▶ 학습데이터

중증도 지수 학습데이터 구성

구분	활용 데이터	속성정보
산재 유형 분류	재활서비스 이용 현황	재해발생형태
독립변수	적용사업장	특근자직종, 본지사구분, 산재업종코드, 총상시인원, 고용상시인원, 고용노무제공자수
	지역별고용조사 데이터	시군구 단위 성별, 연령대, 업종별 근로자 수
	지역별 체류자격별 등록외국인 데이터	시군구 단위 외국인 근로자수
종속변수	최초요양신청서 처리 현황	중증도 지수

- 극도로 갈수록 중증도의 위험도가 커지고, 중증도 단계마다 산재 발생 건수의 비율이 불균형함
  → 3단계(A, B, C)로 재분류하여 종속변수의 각 단계 분포 보완

중증도 종속변수 재분류

종속변수	중증도 단계	산재 건수	건수	비율
A	무장해	275,912	275,912	72.86%
B	경미	23,528	89,036	23.51%
	경도	53,284		
	중등도	12,224		
C	고도	11,188	13,724	3.62%
	극도	2,536		

3. 빈도 지수

▶ 학습데이터

빈도 지수 학습데이터 구성

구분	활용 데이터	속성정보
산재 유형 분류	재활서비스 이용 현황	재해발생형태
독립변수	적용사업장	특근자직종, 본지사구분, 산재업종코드, 총상시인원, 고용상시인원, 고용노무제공자수
	지역별고용조사 데이터	시군구 단위 성별, 연령대, 업종별 근로자 수
	지역별 체류자격별 등록외국인 데이터	시군구 단위 외국인 근로자수
종속변수	재활서비스 이용 현황 및 적용사업장	사업장별 사고 발생 건수

- 산재 유형마다 사업장에서 중복으로 발생하는 산재 수의 비율이 다르고, 산재가 발생하지 않는 사업장이 대다수를 차지하여 데이터 불균형이 심함
  → 산재가 발생하지 않는 데이터를 비율에서 제외하여 종속변수의 각 단계 분포 보완

빈도 종속변수 재분류

종속변수 (건수)	사고			질병			출퇴근		
	건수	비율 (무사고 포함)	비율 (무사고 제외)	건수	비율 (무사고 포함)	비율 (무사고 제외)	건수	비율 (무사고 포함)	비율 (무사고 제외)
0	3,838,542	94.39	-	3,838,542	99.47	-	3,838,542	99.54	-
1	170,516	4.19	74.73	17,299	0.45	85.26	15,695	0.41	87.84
2이상 4미만	44,071	1.08	19.31	2,222	0.06	10.95	1,705	0.04	9.54
4이상	13,594	0.33	5.96	768	0.02	3.79	468	0.01	2.62

4. 재해취약자 지수

▶ 지수 개발 개요

- 내부 데이터를 활용하여 5가지 요인에 따른 산재 발생 비율 분석(특근자직종, 본지사구분, 외국인 여부, 재해당시연령, 성별)
- 각 비율은 통계 데이터를 기반으로 상대값으로 적용
- 사업장에는 존재하지 않는 근로자의 정보는 시군구 단위의 통계 데이터를 활용하여 시군구 단위의 재해취약자 지수 산출
- 산재 유형, 산재 대분류 업종코드마다 각 요인의 산재비율 도출

▶ 활용 데이터

재해취약자 지수 활용 데이터

구분	활용 데이터	속성정보
내부 데이터	적용사업장	특근자직종, 본지사구분, 산재업종코드
	최초요양신청서 처리 현황	외국인 여부, 재해당시연령
	재활서비스 이용 현황	재해발생형태, 성별
외부 데이터	시군구_성_연령별_취업자_근무지기준	시군구 단위 성별, 연령대, 근로자 수
	지역별 체류자격별 등록외국인 데이터	시군구 단위 외국인 근로자 수

– 검증 및 고도화

1. 사고 안전지수 모델 성능 결과

▶ 사고 중증도 지수

- 비교 모델: Random Forest, LightGBM, XGBoost, Catboost
- 최종 선정: LightGBM
- 파라미터 튜닝 결과 정확도 2.38% 향상 → 정확도 87.09%

▶ 사고 빈도 지수

- 비교 모델: Random Forest, LightGBM, XGBoost, Catboost
- 최종 선정: Catboost
- 파라미터 튜닝 결과 정확도 1.34% 향상 → 정확도 91.12%

2. 질병 안전지수 모델 성능 결과

▶ 질병 중증도 지수

- 비교 모델: Random Forest, LightGBM, XGBoost, Catboost
- 최종 선정: XGBoost
- 파라미터 튜닝 결과 정확도 8.38% 향상 → 정확도 76.32%

▶ 질병 빈도 지수

- 비교 모델: Random Forest, LightGBM, XGBoost, Catboost
- 최종 선정: Catboost
- 파라미터 튜닝 결과 정확도 0.53% 향상 → 정확도 92.68%

3. 출퇴근 안전지수 모델 성능 결과

▶ 출퇴근 중증도 지수

- 비교 모델: Random Forest, LightGBM, XGBoost, Catboost
- 최종 선정: LightGBM
- 파라미터 튜닝 결과 정확도 2.56% 향상 → 정확도 72.93%

▶ 출퇴근 빈도 지수

- 비교 모델: Random Forest, LightGBM, XGBoost, Catboost
- 최종 선정: Random Forest
- 파라미터 튜닝 결과 정확도 6.5% 향상 → 정확도 99.10%

4. 안전지수 산정 프로세스

▶ 이력 정보를 통한 예측 결과 보정

- 산재가 발생하지 않은 사업자의 예측결과 보정을 위해 최소값 0.5로 적용
- 기존 산재 이력이 큰 사업장의 경우 1.5 적용

·각 지수는 위험도 5단계 분류를 위해 0 ~ 5로 정규화 적용

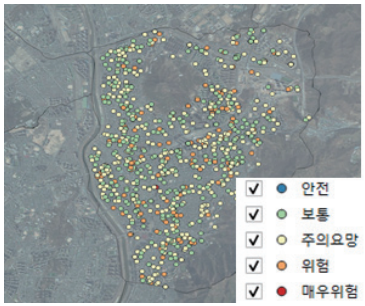
안전지수 산정 프로세스

– 결과 구현

1. QGIS 활용 주제별 시각화

▶ 사업장 단위

- 산재 유형별 중증도, 빈도, 안전지수 시각화
- 위험단계 필터링을 통해 위험 사업장 파악 가능

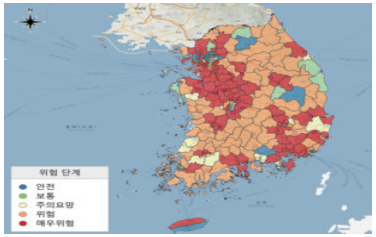


사업장 단위 안전지수 시각화

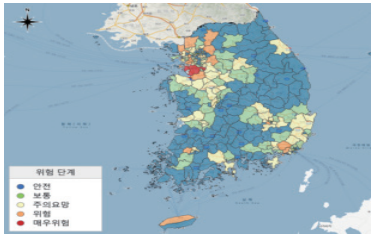
사업장 단위 안전지수 시각화(필터 적용)

▶ 시군구 단위

- 산재 유형별(사고, 질병, 출퇴근) 지수(중증도, 빈도, 안전)를 기준값 (상댓값, 절댓값) 기준으로 시각화



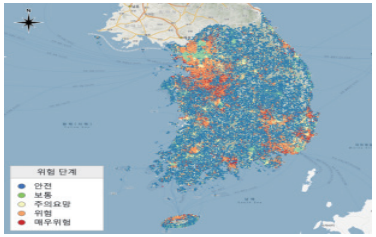
시군구 단위 안전지수 시각화(상댓값)



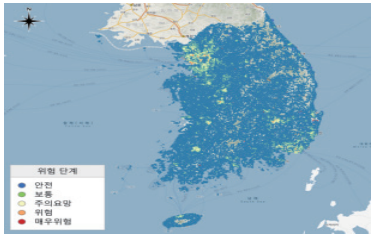
시군구 단위 안전지수 시각화(절댓값)

▶ 그리드 단위

- 산재 유형별(사고, 질병, 출퇴근) 지수(중증도, 빈도, 안전)를 기준값 (상댓값, 절댓값) 기준으로 시각화

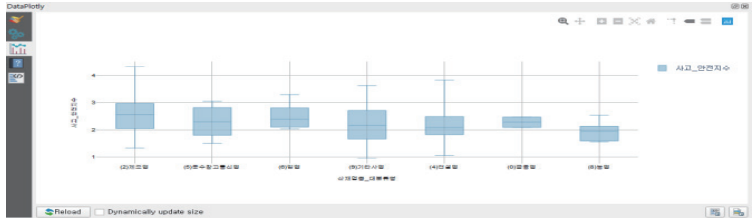


그리드 단위 안전지수 시각화(상댓값)



그리드 단위 안전지수 시각화(절댓값)

2. 대상 지역 업종별 안전지수 현황 시각화



업종별 안전지수 현황 시각화

정책활용/기대효과

개발된 모델은 사업장, 근로 업종, 산업재해 발생 건수 및 사고, 질병 유형 등을 기반으로 지역의 산업재해 위험정도를 알 수 있도록 '재해안전지수'를 산출하여 제시한다. 근로복지공단은 이 지수를 지리정보와 연계하여 시각화된 모형을 개발하였으며, 지수의 변화 비교를 통해 직관적으로 산재 위험 정도의 변화를 알 수 있도록 시계열 비교 결과를 함께 제공한다.

또한 분석 모델의 활용성을 높이기 위해 개발모델의 데이터 현행화 과정으로 '23년 데이터를 업데이트하였으며, 개발모델의 시범 활용을 위해 '24년 대형 인명사고가 발생했던 경기도 화성 지역과 대형 공업 시설이 많아 산재 빈도가 높은 울산 지역을 대상으로 모델을 활용한 분석 보고서를 작성하였다.

공단과 협업하여 분석을 진행한 두 지자체(화성시청, 울산시청)를 직접 방문하고, 지자체의 산업 예방 안전 활동에 활용될 수 있도록 보고서의 분석 결과를 설명과 함께 제공하였다. 그 결과 제공된 정보의 활용에 대한 긍정적인 반응을 얻을 수 있었다.

아울러 정보 제공 활동에 그치지 않고 개발된 데이터의 활용 확대 방안을 모색하기 위해 안전보건공단, 울산연구원과 업무 협의를 진행하였으며 지속적인 실무자 협의를 통해 데이터 교류와 활용 방안에 대해 함께 모색하기로 하였다.

* 안전관리 협의 : 화성시청(9.19.), 울산시청(11.26), HD현대중공업(10.25)
* 데이터 활용 협의 : 안전보건공단(10.10), 울산연구원(11.26)

본 과정에서 공단은 해당 모델이 제공하는 정보의 유용성과 필요성을 확인할 수 있었다. 다만 지자체 등과의 지속 가능한 교류를 위해서는 본 PoC 모델의 고도화를 통해 실시간 정보가 제공되어야 하며, 동시에 이용자가 쉽고 빠르게 정보를 제공받을 수 있도록 접근성을 개선할 필요가 있다. 이를 위해서는 모델이 제공하는 정보가 사용자 입장에서 실효적인 정보로 재구성되고 사용 및 전달될 수 있도록 유관 기관 및 활용 기관과의 협력과 개발이 우선 진행되어야 한다.

근로복지공단은 개발된 재해안전지수가 양질의 정보로 계속 활용될 수 있도록 고도화 방안을 지속적으로 모색하고자 한다.

산업
경제

PART 4

산업경제

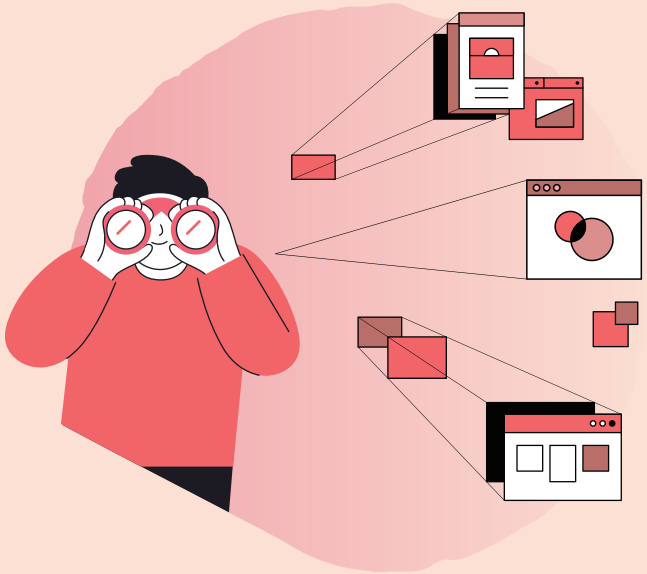
1. 중소벤처기업 전용 빅데이터 플랫폼
   중소벤처기업진흥공단
2. 상권변화 요인을 활용한 상권 부실징후 예측
   소상공인시장진흥공단
3. 수산물 공급데이터를 활용한 수산종자 수급예측
   한국수산자원공단

PART 4-1
산업경제

비즈패스파인더(BizPathFinder)

인공지능과 빅데이터 분석으로 열어가는 중소벤처기업 성공 비즈니스 가이드

중소벤처기업진흥공단



추진목적/배경

빅데이터 기술의 발전과 생성형 AI의 출현 등 급격한 기술 진보는 데이터경제 시대를 열었으며, 이에 따라 데이터를 활용한 기업경영은 이제 거스를 수 없는 시대적 흐름이 되었다. 특히 중소기업은 한 번의 의사결정이 기업의 성공과 실패를 좌우할 수 있기 때문에 자원배분과 미래전략 수립 측면에서 대기업보다 더 데이터 기반의 객관적이고 합리적인 의사결정이 요구된다.

하지만 국내 중소기업의 디지털 성숙도는 40.7점에 불과하며. 데이터 분석시스템을 보유한 기업은 전체의 30%, 전문인력을 보유한 기업은 26%에 불과해 여전히 데이터 활용에 어려움을 겪고 있다.

〈 중소기업 디지털 성숙도 조사 주요내용 〉

구 분	비율	구 분	비율
고객 데이터 분석시스템 보유기업	30.0%	디지털화 전략 준비기업	35.7%
전문인력 보유기업	26.0%	관련 교육 계획 · 실시기업	14.3%

* 2022년 중소기업 디지털 성숙도 조사 (중소기업중앙회)

또한 중소기업 간 역량 격차가 심각한 상황으로, 중기업은 소기업에 비해 신기술 도입수준이 2.8배, 활용수준이 2.3배 높아 기업 규모별로 데이터 활용 역량에 큰 격차가 존재한다.

〈 기업규모별 디지털 격차 〉

구 분	대기업	중기업	소기업
SW신기술 도입수준 (10점만점)	9.03	3.09	1.09
SW신기술 활용분야수 (개수)	13.60	4.33	1.88

* 국내 기업의 디지털 전환 촉진을 위한 주요 요인 도출 및 실증 연구 (소프트웨어정책연구소)

이러한 문제를 해결하기 위해 중소벤처기업진흥공단은 중소기업의 데이터 분석 및 활용 인프라를 구축하고, 기업 간 데이터 활용 격차를 해소하고자 중소벤처기업 전용 빅데이터플랫폼인 비즈패스파인더(BizPathFinder)의 개발을 추진했다.

비즈패스파인더(BizPathFinder)는 빅데이터와 인공지능을 활용하여 중소기업에게 「비즈니스의 길을 찾아준다.」는 의미를 담고 있으며, ❶기업포지셔닝 분석, ❷미래 성장경로 예측, ❸적합 정책사업 추천, ❹수출 품목 제안 서비스 등을 통해 중소기업의 데이터 기반 경영을 지원한다.

분석 사전 준비

– 활용 데이터

데이터명	형태	내용	출처	기준 년도	내·외부 데이터
고객정보	DBMS	중진공 정책사업 신청정보 등	중진공	1978~	내부
주요통계	API	국내총생산, 경제성장률, 고용·실업률, 기업·종사자수, 매출액 등	통계청	2017~2024	외부
경기현황	API	소비자/생산자 물가, 통화량, 기준금리, 성장·수익·생산성 지표 등	한국은행	2017~2024	외부
수출	CSV	월별·국가별·품목별 수출실적 등	무역통계진흥원	2017~2024	외부
고용	CSV	월별 고용현황 등	고용정보원	2017~2024	외부
지식재산정보	FTP	특허, 실용신안, 디자인, 상표정보 등	특허정보원	2017~2024	외부
기업정보	FTP	기업개요, 재무정보, 신용정보 등	나이스평가정보, 한국평가데이터	2024	외부

비즈패스파인더(BizPathFinder)는 중소벤처기업진흥공단 내부 및 외부에서 확보한 58만 개 기업, 208종, 1억 개 이상의 데이터를 활용하여 빅데이터 분석서비스를 제공한다.

내부데이터는 중진공의 42개 현장조직을 통해 축적된 기업정보를 바탕으로 하며, 수출, 고용, 특허 등의 외부 데이터는 업무 협약, 공공데이터 활용, 구매, 스크래핑 등을 통해 수집하였다.

특히 기획재정부의 데이터협업 과제를 통해 협업기관의 데이터를 실시간으로 비즈패스파인더와 연계하였으며, 한국특허정보원과의 업무 협약을 통해 중진공 지원기업 전체에 대한 특허데이터를 확보하였다.

분석과정

– 분석 환경

1. 분석 인프라 : 기관 자체 빅데이터 플랫폼 서버장비 이용
2. 분석 환경 : python, Visual Studio Code, PyCharm, Jupyter Notebook 등

– 데이터 수집

1. 사용 데이터 출처 : 기관 자체 생성데이터 및 외부 연계 데이터
2. 사용 데이터 형식 : DBMS, CSV 등

– 데이터 전처리

1. 결측값 및 오류 점검
   - 재무데이터 내 일부 NULL 값은 0으로 대체
   - 수출 품목제안 분석에 불필요한 업종코드 제외
   - 중진공 고객정보의 기업 설립일자와 법인 설립일자가 모두 NULL인 기업 제외
2. 이상치 점검
   - 기업 설립일자 이상치 제거(예: 설립년도 1000년)
   - 중진공 정책사업 시작년도(1978년) 이전의 지원사업 데이터 제거
3. 파생변수 생성
   - 정책사업 추천 시, 자금 지원요건에 따른 결과값 필터링을 위한 파생변수 설정 → 업력 7년 미만(1), 업력 7년 이상(0)
     (예: 추천사업으로 창업자금이 나왔지만 기업 업력이 10년인 경우, 추천 결과에서 제외)
   - 대표자의 나이에 따라 청년 여부를 Y(청년 해당), N(청년 미해당)으로 구분

– 시각화

- Anaconda, TensorFlow, PyTorch, NumPy, Scikit-learn, Pandas, SciPy 등 Python 패키지를 활용하여 데이터 통계분석 및 시각화 처리

– 모델링

〈성장경로 예측 분석방법 요약〉

① 성장 변수 선별	② 성장 특성별 그룹화	③ 그룹 內 위치 측정	④ 미래성장경로 추정
AI분석(머신러닝) (K-means, Randomforest)	AI분석(머신러닝) (K-means)	AI분석(머신러닝) (주성분분석)	재무비율 분석
• **고성장기업 데이터 학습** • **성장성**(시가총액증감)에 영향을 미치는 **주요변수(재무비율) 분석**	• 선별된 변수를 활용, **중소기업 데이터 학습** • 5년간의 성장 특성을 반영하여 기업 분류	• 그룹 내 **기업** 간 **재무성과** 차이(변별력)를 **학습** • 복합 재무성과 지표를 단일 지수화·서열화	• 소속그룹의 재무특성 활용 재무 추정 • 기업의 미래 2개년재무 예측

▶ 분석 프로세스

① 기업성장을 기업가치(시가총액) 증가로 정의하고, 고성장기업 1,573개사의 성장정보 학습

② Randomforest(랜덤포레스트)를 활용하여 성장성(시가총액증감)에 영향을 미치는 주요변수(37개 재무비율) 선별

③ 선별된 주요 변수를 활용하여, K-Means 군집화 기법을 통해 중소기업을 성장 특성별로 5개 그룹으로 분류

④ PCA(주성분분석)를 통해 매출성장, 부채비율 등 다양한 재무성과를 단일 지수화하고, 각 그룹 내 기업들의 성과를 서열화

⑤ 분석 대상기업과 유사한 성장특성(5년간의 재무변화)을 가진 그룹을 판별하고, 해당 그룹의 성장률 변화를 활용하여 향후 2개년의 미래 재무 추정

– 검증 및 고도화

- Train data(훈련용 데이터), Test data(테스트 데이터)를 7:3 비율로 나누어 모델 성능 검증
- 최종적으로 과거 재무데이터를 입력하여 예측한 현재의 재무비율과 실제값을 비교하여 모델 설명력 검증

– 결과 구현

▶ 비즈패스파인더 주요 서비스

① 성장경로 예측

- 서비스 이용 기업과 유사한 특징을 가진 기업군을 분석하여 기업의 미래 재무를 추정하고 미래 성장경로를 예측하는 서비스
- 자원조달, 투자결정 등 경영계획 수립에 활용 가능

② 적합 정책사업 추천

- 3개년 연평균 20% 이상 성장하는 고성장 기업의 빅데이터를 분석하여, 고성장 기업의 과거 경영환경과 유사한 중소기업에게 당시 고성장을 촉진했던 정책사업을 추천하고, 중소기업의 고성장을 유도하는 서비스

③ 기업 포지셔닝 분석

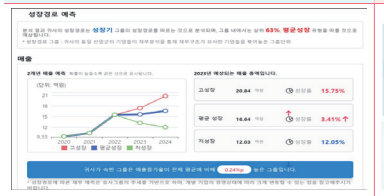
- 동일지역, 동일업종의 기업들과 자사의 수준을 비교하고, 기업의 재무비율을 분석하여 업계 내 포지션과 재무현황을 제공하는 서비스

④ 수출 품목 제안

- 기업의 생산품목과 제품정보를 인공지능 언어모델을 활용하여 업종, HS코드, 수출통계 등과 비교 분석하고, 유망 수출품목을 제안하며, 해당 품목의 3개년 수출실적과 주요 수출국 현황 정보를 제공하는 서비스

【 비즈패스파인더 분석서비스 주요 화면】

① 성장경로 예측 서비스



② 적합 정책사업 추천 서비스

③ 기업포지셔닝 분석

④ 수출품목 제안

정책활용/기대효과

2024년 한 해 동안 중소기업은 비즈패스파인더(BizPathFinder)에 약 165만 회 접속하여 빅데이터 분석서비스를 기업경영에 활용하였다. 또한 중소벤처기업진흥공단의 주요 7개 정책사업을 활용한 기업 중 약 54.3%는 적합 정책사업 추천서비스를 이용하여 해당 사업을 신청한 것으로 나타났다.

〈정책사업 지원기업 중 분석서비스 이용기업 현황〉

(단위 : 개사, %)

구분	기업진단	연수사업	정책자금	제조바우처	내일채움	수출	합계
지원기업	2,858	2,986	15,289	2,081	10,693	5,021	38,928
서비스 이용기업	2,459	1,850	8,388	1,143	5,240	2,072	21,152
활용율	86.0%	62.0%	54.9%	54.9%	49.0%	41.3%	54.3%

비즈패스파인더(BizPathFinder)의 빅데이터 분석서비스는 자본, 인력, 기술, 시간 등 자원이 부족한 중소기업이 데이터를 활용한 객관적이고 효율적인 의사결정을 내릴 수 있도록 지원한다. 이를 통해 중소기업은 스스로 인프라를 구축하여 데이터를 분석하는 데 드는 시간과 비용을 절감하고 업무 생산성을 향상시킬 수 있다.

또한, 다양한 통계 분석 서비스를 제공하여 대내외 경제 현황을 파악할 수 있도록 한다. 기업 포지셔닝 분석과 성장 경로 예측 서비스를 통해 자사의 업계 내 위치를 비교하고, 이를 바탕으로 기업 성장 전략을 수립할 수 있게 하여 지속 가능한 성장을 촉진한다.

더불어 성장 단계에 맞는 정책을 제시함으로써 기업이 유동성 확보 등 필요한 자원을 적시에 확보하고, 지원 정책을 보다 효과적으로 활용할 수 있게 한다.

중소벤처기업진흥공단은 주요 서비스를 지속적으로 고도화하고 새로운 서비스를 개발할 예정이다.

2025년에는 기존의 수출 품목 제안 서비스를 고도화하여 '수출 전략 국가 및 품목 추천 서비스'로 개선할 계획이다. 기존에는 관세청의 국내 수출입 통계 데이터를 활용했으나, 앞으로는 UN에서 제공하는 전 세계 국가별, 품목별 수출입 데이터를 추가 분석하여 생산 제품의 수출 경쟁력과 전 세계 국가별 무역 환경을 분석하고, 품목별 및 국가별 수출 포지셔닝 전략 수립을 지원할 것이다.

신규 서비스로는 「미래위험 예측서비스」를 개발할 예정이다. 이 서비스는 부실 기업의 데이터를 분석하고 주요 부실 패턴을 파악한 후 유사 중소기업에게 부실 징후를 사전에 알려준다. 이를 통해 중소기업은 경영 위기를 사전에 대비할 수 있다.

PART 4-2
산업경제

상권변화 요인을 활용한 상권 부실징후 예측

부실징후 예측 모델 : 빅데이터 기반 상권 모니터링 체계 마련

소상공인시장진흥공단

추진목적/배경

코로나, 경기침체, 인구감소 등으로 인해 소상공인 산업 분야의 위험도가 점점 높아지고 있다. 2023년 국세통계 기준 사업자 폐업률은 19.3%로 전산업 평균 9.9%에 비해 상당히 높은 수준이다.

현재 상권의 잠재력 변화에 따른 상권 적응 수준, 젠트리피케이션, 상권발달 요인에 대한 연구는 현저히 부족한 실정이다. 기존의 상권분석은 주로 사업자(가맹점) 기준의 데이터에 의존하여 분석 및 제공되었기 때문에 잠재적 수요자 및 경쟁점 관점에서의 추가적인 분석이 필요한 상황이다.

일부 민간 금융회사에서도 관련 정보 서비스를 제공하고 있지만, 이는 단순한 시계열 데이터 제공에 불과하다. 또한 본 기관의 상권정보 시스템에서 서비스하는 데이터 모델(창업 기상도, 입지 평가, 매출 예측 등) 역시 다양한 업종의 창업으로 인한 경쟁점 증가에 따른 위험정보를 제공하고 있지 않다. 따라서 상권 부실 징후 예측 모델의 개발이 필요한 상황이었다.

이러한 문제를 해결하기 위해 소상공인시장진흥공단은 상권의 변화요인을 정의하고 각 요인에 따른 상권 부실 징후 예측 모델을 개발하여 지역상권의 잠재력을 시계열적으로 측정하고자 하였다.

분석 사전 준비

– 활용 데이터

데이터명	형태	내용	출처	기준년도	내·외부 데이터
매출(카드사)	xlsx	매출_기초단위구, 매출_행정동	카드사	2020~2023	외부
소득·소비	xlsx	월별 추정 소득, 월별 추정 소비	카드사	2022	외부
유동인구	xlsx	월별 유동인구 수 (성별, 연령별, 시간별)	통신사	2018~2023	외부
주거인구	xlsx	월별 주거인구 수 (건물 및 행정동)	행안부	~2023	외부
직장인구	xlsx	월별 지역별(행정동) 단위 직장인구 수	카드사	~2022	외부
업소	xlsx	생활밀접업종 사업체	국세청, 통계청,소진공	~2022	내·외부
배달 건 수	xlsx	행정동별, 업종별, 월별, 성별, 연령별, 시간별, 요일별 배달 건 수	배달플랫폼	~2023	외부
배달 매출액	xlsx	행정동별, 업종별, 월별, 성별, 연령별, 시간별, 요일별 배달 매출액	배달플랫폼	~2023	외부
부동산	xlsx	권역별, 규모별, 층별 임대시세	한국부동산원	~2023	외부
공동주택	xlsx	전국 공동주택 단지, 동, 호, 정보	국토교통부	2023	외부
시설종합	xlsx	시설종합(주소, 좌표), 시설학교(주소, 좌표)	data.go.kr	2022	외부
교통	xlsx	교통버스 및 지하철 (위치, 노선, 승하차 인원 등)	data.go.kr	2022	외부
상권	xlsx	발달상권, 르네상스, 전통시장 경계영역	소진공	2022	내부

각 공공·민간 기관과 데이터 교류를 위한 협약을 체결하고 업무 협의를 진행하였으며, 민간 데이터 구매 및 공공데이터 포털 활용 등 다양한 경로를 통해 필요한 데이터를 확보하였다.

각 데이터는 행정동, 기초단위구, 좌표, 건물 등이 서로 다른 단위로 구성되어 있어 이를 공단에서 지정한 상권별 데이터로 재가공하여 활용할 수 있도록 정제하였다. 또한 분석용 데이터는 공단 내부에서 개인정보 보안 검토를 거친 후 CSV 파일 형태로 변환하였다.

분석과정

– 분석 환경

1. 분석 인프라 : 기관 내 PC 이용
2. 분석 환경 : Python, Tableau (대시보드)

– 데이터 수집

사용 데이터 출처 : 민간, 공공데이터, 기관 자체 보유데이터

사용 데이터 형식 : csv

– 데이터 전처리

1. 결측값 처리 및 데이터 전처리
   - 건물 단위, 셀 단위 등의 데이터를 상권, 행정동, 자치구 단위 데이터로 그룹화
   - 업종 코드가 없는 데이터 제외
   - 분기/반기별 업데이트 데이터의 경우, 이전 최신 기준 값으로 보간
   - 시계열 데이터 누락 값의 경우, 선형 보간법을 사용하여 보간
   - 장소 기준 누락 값의 경우, 평균값으로 대체
2. 파생변수 생성
   - 과거 데이터 기반, 증감률 변수 생성

- 상권별 매출액, 가게 수를 기반으로 가맹점 당 매출액 변수 생성
- 자치구, 타상권 매출액과 비교하여, 상대적 매출액 변수 생성
- 연령대별, 성별 등을 활용하여 다양한 연령대 기준 인구 데이터 변수 생성
- 상권 면적 기반으로, 유동/주거/직장인구, 점포 밀도 변수 생성
- 업종 개/폐업 날짜를 기준으로 업종 다양성, 개폐업점포비율, 순개폐업점포수, 생존율 변수 생성

– 모델링

1. 상권 유형 구분
   - 상권별 유동/주거/직장인구, 주요 업종, 매출액, 상권 인프라 환경 등의 특성을 기반으로 상권을 4가지 유형으로 구분(대학가 상권, 주거지 상권, 오피스 상권, 상업지구 상권)
2. 매출 영향변수 도출
   - RandomForestRegressor 모델을 활용하여, 월별·상권·업종별 매출에 영향을 미치는 변수 파악
3. 상권부실징후 지표 생성
   - 상권부실징후를 측정하기 위해 크게 5가지 카테고리(성장성, 안정성, 영업력, 지역 지표, 임대)를 기준으로 세부 평가 항목들 생성
   - 상권 유형별 특징을 고려하여 세부 평가 항목 조정 및 전문가 자문을 통해 가중치 조정
   - 최종 산출된 점수를 기준으로 상권 부실 징후를 5단계로 구분(양호, 관심, 주의, 경계, 위험)
4. 상권 부실징후 예측모델 생성
   - 시계열 예측 모델을 활용하여 향후 3개월 상권 부실 징후 예측 모델 설계
   - 4가지 모델 생성 후, 평가지표를 활용한 예측모델 성능 비교를 통해 상권유형별 부실 징후 예측 모델 생성

▶ 활용 모델

- ARIMA(Autoregressive Integrated Moving Average)
- LSTM(Long Short-Term Memory)
- RNN*(Recurrent Neural Network)
- GRU**(Gated Recurrent Unit)

* RNN : 시퀀스 데이터를 처리하는 신경망 구조로, 이전의 출력을 다음 입력에 반영하여 시간적 순서나 문맥을 학습할 수 있는 특징을 가짐

** GRU : RNN의 변형으로, 장기 의존성을 학습하기 위해 업데이트 게이트와 리셋 게이트를 사용하는 순환 신경망 모델

▶ 활용 평가지표 : MAPE, MAE, RSE, Accuracy

▶ 최종 활용 예측모델

- 대학가 상권 : GRU 사용
- 주거지 상권 : ARIMA 사용
- 오피스 상권 : ARIMA 사용
- 상업지구 상권 : ARIMA 사용

– 검증 및 고도화

- 새로운 데이터 업데이트 이후, 예측한 부실징후 지표와 실제 값으로 산출한 부실징후 지표 비교를 통해 예측모델 정확도 재검증 진행
- 서울시 254개 상권에 대해 92.5%의 높은 정확도 기록

	평가지표			
상권유형	Accuracy	MAPE	MAE	RSE
전체 (254개 상권)	92.5%	8.8%	0.16	20.7%
대학가 (14개 상권)	90.5%	9.9%	0.19	24.7%
주거지 (51개 상권)	92.8%	8.1%	0.16	19.4%
오피스 (111개 상권)	95.5%	8.4%	0.15	19.1%
상업지구 (78개 상권)	88.5%	9.7%	0.18	23%

– 결과 구현

Tableau를 활용하여 데이터 변화 추이를 시각화하여 표현하고, 상권별, 시/군/구, 행정동별 부실징후 및 예측결과 제공

부실징후 예측모델 시각화

정책활용/기대효과

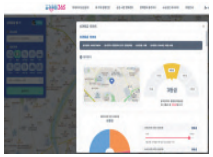
소상공인시장진흥공단은 부실징후 예측을 위한 지수 모델링, 매출 영향 요인 분석 기법, 상권 유형화 기법을 「소상공인365 상권 진단 서비스」에 반영하였으며 이를 공단의 정책자금 심사 시 상권 평가 자료로 활용할 예정이다.

본 사업을 통해 개발된 상권 부실 징후 예측 모델로 상권 유형별 향후 3개월의 부실 징후 점수 및 지표를 예측할 수 있게 되었다.

이용자는 소상공인 빅데이터플랫폼을 통해 상권 진단을 위한 5종 지수(집객력, 구매력, 성장성, 안정성, 활성화)에 대한 평가 결과를 5개 등급으로 제공받을 수 있다. 또한 입지 평가 보고서를 통해 선택한 지역과 업종에 대한 매출 영향 요인 및 매출 분위 정보도 함께 확인할 수 있다.

향후 공단은 부실징후 예측 지수 모델링 등 서비스에 활용되는 모델링의 신뢰도를 검증하고, 이를 기반으로 모델을 고도화하여 상권 분석 보고서 서비스를 제공할 계획이다.

빅데이터플랫폼 서비스 반영 현황

상권 정보 시스템 (기존)			
	상권평가보고서 **'상권등급'**	입지평가보고서 **'종합입지등급'**	경쟁분석보고서 **'평가종합지표'**
빅데이터 플랫폼 (반영)			
	상권등급리포트 **'상권등급'**	입지평가보고서 **'입지평가등급'**	경쟁분석보고서 **'평가종합지표'**

PART 4-3
산업경제

수산물 공급데이터를 활용한 수산종자 수급예측

한국수산자원공단

추진목적/배경

한국인의 1인당 연간 수산물 소비량은 63.5kg으로 이는 전 세계 1위이다. 삼면이 바다로 둘러싸인 우리나라는 다양한 수산물을 생산하며, 양식 기술이 발달해 있어 어업생산량보다 양식생산량이 약 2배 더 많다.

그러나 기후변화 등 해양환경의 변화와 수산물 수요 변화로 인해 수산물 공급시장이 불안정해지고 있다. 지난해 고수온 현상으로 양식장 굴의 25% 이상이 집단 폐사했으며, 김의 경우 한류 열풍으로 인기가 치솟아 1조 원의 수출 기록을 달성했지만 생산 환경이 악화되면서 종자의 수급이 불안정해지고 시장 가격이 상승하는 문제가 발생했다.

한국수산자원공단은 불안정한 수산물 수급에 대응하기 위해 수산종자 생산업 실태조사 데이터와 수산물 빅데이터를 기반으로 수산종자의 수요 예측 모델을 개발하였다.

특히 넙치, 전복, 김을 주요 양식 수산물로 선정하고 양식생산량, 수온, 국제 환율 등 외생 변수를 반영한 빅데이터 모델링을 적용하여 5개년 예측치를 산출하였다. 이를 통해 수산종자의 안정적 수급을 도모하고자 하였다.

분석 사전 준비

– 활용 데이터

데이터명	형태	내용	출처	기준년도	구분
수산종자생산업 실태조사	xlsx.	• 인력 수급 현황 • 어류, 패류, 해조류, 기타 종자 생산, 판매 사육 현황	한국수산 자원공단	2020~2021	내부
해양수질자동 측정망 센서자료	xlsx.	• 연도별, 월별, 정점 코드별 수온, 염분 등	해양환경 공단	2003~2022	외부
소비자심리지수	xlsx.	• 연도별, 월별 소비심리지수	KOSIS (통계청)	2008~2023	외부
기후통계분석	csv.	• 평균기온, 일평균 강수량 등	기상청	2012~2023	외부
국제유가 도입현황	xlsx.	• 최근 12개월 월평균 원유가 추이, 수출입 동향	KOSIS (통계청)	2003~2023	외부
관광통계 (주요관광지점 입장객)	xlsx.	• 연도별, 월별, 지역별 내/외국인 관광객 수	TDSS (관광개발정보시스템)	2004~2023	외부
수산양식물 수급 현황	xlsx.	• 연도별 양식 생산량 합계(광어, 전복, 김)	한국해양 수산개발원	품종별 상이	외부
수산종자 수급 현황	xlsx.	• 연도별 종자 생산량 합계 (광어, 전복, 김)	한국해양 수산개발원	2015~2023	외부
기간별 평균 환율	xlsx.	• 연도별, 월별, 기준환율 등	우리은행	2012~2023	외부
방류실적데이터	xlsx.	• 연도별, 지자체별, 품종별 매입방류 양 및 계약 단가	한국수산 자원공단	2017~2022	내부

분석 데이터는 수산공단에서 조사한 수산종자생산업 실태조사 결과를 주로 활용하였으며 그 외 외생 변수는 통계청 국가통계포털, 기상청, 해양수산개발원 등에서 정기적으로 조사·공개하는 데이터를 통해 확보하였다.

분석과정

– 분석 환경

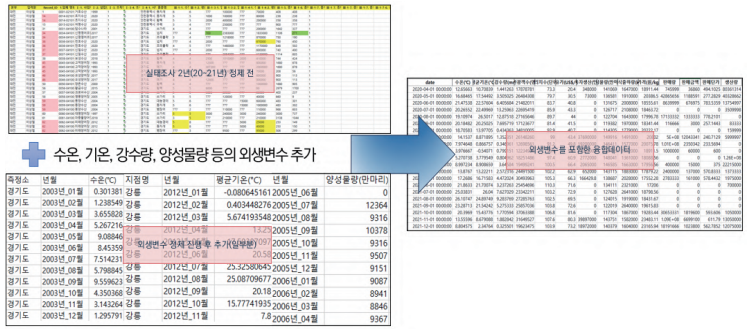
1. 분석 인프라 : (주)선도소프트 및 기관 내 PC 이용
2. 분석 환경 : python, Tableau, MLOPS* 등
   * 머신러닝(Machine Learning)과 운영(Operations)을 합친 용어로 프로덕션 환경에서 머신러닝 모델이 지속적이고 안정적으로 배포되도록 유지, 관리, 모니터링 해주는 것

– 데이터 수집

1. 사용 데이터 출처 : 기관 자체 생성 데이터 및 외부기관 공개 데이터
2. 사용 데이터 형식 : csv, xlsx

– 데이터 전처리

1. 종자별로 분리된 실태조사 데이터를 정제 후 정형화하여 품종 컬럼을 추가한 하나의 데이터로 융합
2. 실태조사 데이터에서 예측 분석을 진행할 데이터를 정제하여 수온, 기온, 환율 등의 외생변수와 융합



실태조사/외생변수 데이터 정제 전/후 및 융합 데이터

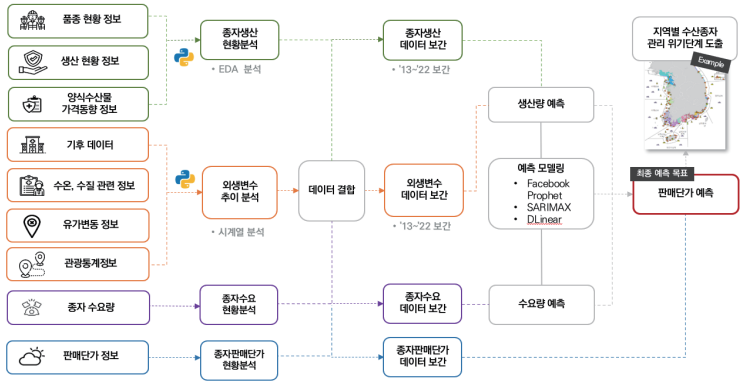
3. 해당 융합데이터에 대해 과거 추정 및 데이터 보간을 진행하고 예측 결과 및 위험단계 데이터 도출

– 모델링

▶ 사용한 분석 모델

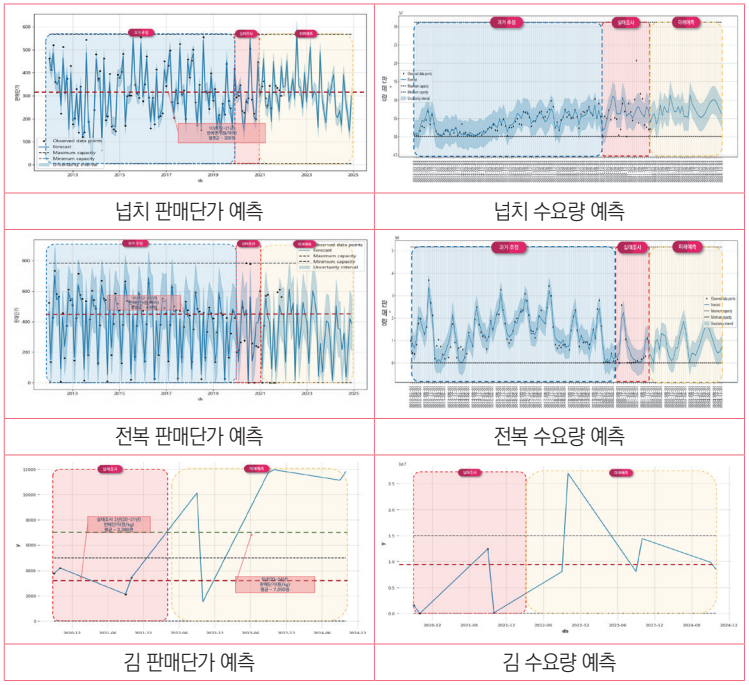
- Facebook Prophet
  - → 유연한 모델 구조로 직관적인 파라미터 조절이 가능
  - → Trend, Seasonality, Holiday effects 등을 조절하는 매개변수를 제공하여 주기적인 특성과 이벤트, 계절성을 고려한 예측 수행
  - → 데이터의 이상치를 식별하고 잠재적 특징들을 반영하여 예측의 안정성을 높임
  - → 큰 규모의 데이터셋에 대한 처리를 지원하며, 병렬 처리 및 GPU 가속을 사용해 효율적으로 작동
- SARIMAX
  - → ARIMA 모델의 확장 버전
  - → 계절성을 처리하는 SARIMA 모델에 외생변수를 처리하는 기능 추가
- DLinear
  - → 간단한 선형 모델
  - → 장기 시계열 데이터에서 시간 순차성 정보를 보존하면서 추세와 주기성에 대한 특징을 추출
  - → 단변량 예측 분석만 가능하여 외생변수 추가가 어렵다는 단점이 있음

▶ 분석 프로세스



빅데이터 분석 프로세스 도식화

- 데이터 추이 분석 진행 후 필요 시 데이터를 보간하여 모델링
- 최종 예측 목표는 종자 판매단가이며 이에 대한 지역별 수산종자 위기단계를 도출
- 실태조사 2년간('20년~'21년) 데이터와 10년간('12년~'21년)의 과거 외생변수 데이터를 Prophet 모델로 분석하여 과거 8년치('12년~'19년) 추정



넙치 판매단가 예측 | 넙치 수요량 예측

전복 판매단가 예측 | 전복 수요량 예측

김 판매단가 예측 | 김 수요량 예측

- 검증 및 고도화

▶ 성능 검증

- 분석모델을 통해 예측된 값과 전수조사 방식으로 이루어진 과거 실태조사 값 간 비교 등을 통해 이루어짐
- MAE(Mean Absolute Error), MSE(Mean Square Error), RMSE(Root Mean Square Error)를 성능 비교 지표로 사용
  → 세 지표 모두 값이 낮을수록 모델의 성능이 높음을 나타냄

- 생산량 예측 분석에 대한 세 모델 비교 지표

평가기준	Prophet(다변량)	SARIMAX(다변량)	Dlinear(단변량)
MAE	0.0943	0.1552	0.1551
MSE	0.0131	0.0365	0.0329
RMSE	0.1145	0.191	0.1815

- 판매량 예측 분석에 대한 세 모델 비교 지표

평가기준	Prophet(다변량)	SARIMAX(다변량)	Dlinear(단변량)
MAE	0.0869	0.1882	0.1707
MSE	0.0102	0.0505	0.0431
RMSE	0.1008	0.2248	0.2076

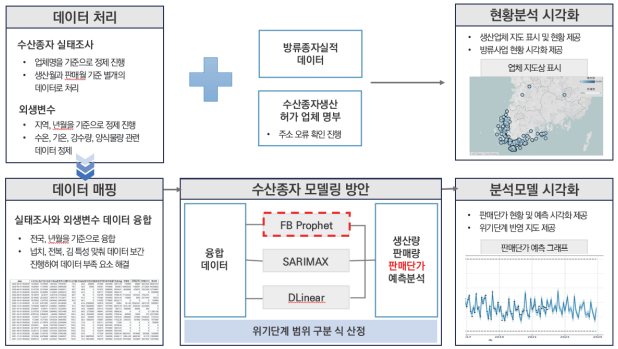
- 판매단가 예측 분석에 대한 세 모델 비교 지표

평가기준	Prophet(다변량)	SARIMAX(다변량)	Dlinear(단변량)
MAE	0.1018	0.1276	0.1511
MSE	0.0164	0.0248	0.0323
RMSE	0.1282	0.1573	0.1797

→ 생산량, 판매량, 판매단가 예측 분석 모두 fb Prophet 모델이 가장 적절하다고 평가됨

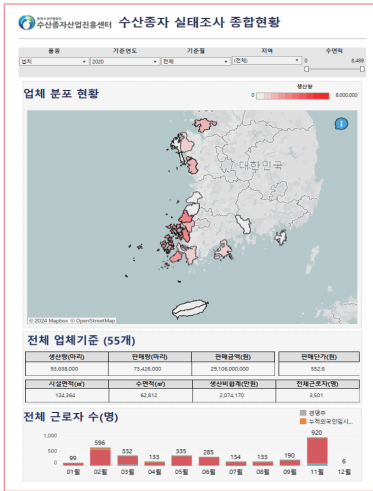
- 결과 구현

1. 시각화 프로세스

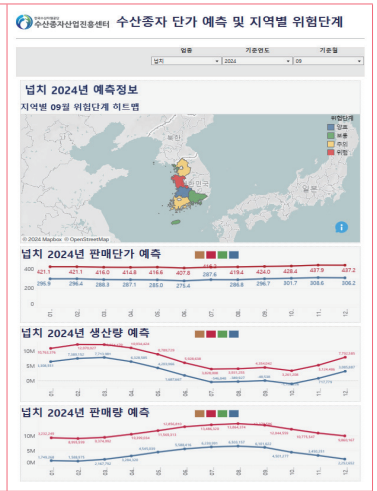


수산물 공급 및 수요 빅데이터 분석 시각화 프로세스 도식화

2. 시각화 결과



업체별 현황 지도



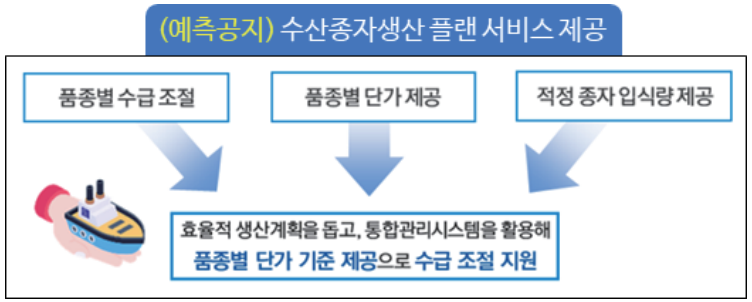
수산종자 단가 예측 및 지역별 위험단계

정책활용/기대효과

본 수산종자 수급 예측 모델은 시범활용 과정에서 국가 통계로 매년 시행되는 '수산종자 생산업 실태조사'의 정확도를 높이는데 활용되었다. 한국수산자원공단은 모델을 통해 도출된 예측치와 실태 조사 값 간 차이가 큰 경우 심층 조사를 실시하여 통계의 정확도를 높였다. 그 결과, '수산종자 생산업 실태조사'는 통계치 공표 2년 만에 통계청 주관 국가통계 품질진단에서 최고등급을 달성하는 성과를 거두었다.

또한 모델개발 과정에서 수산분야 최초로 Facebook Prophet 모델을 활용하여 김, 전복, 넙치의 수요 예측 및 변동치를 분석하였다. 본 모델은 다른 품종의 예측에도 활용할 수 있어, 기후변화와 수요 변동으로 공급이 불안정한 양식수산물의 종자 수급 관리 도구로 자리매김할 것으로 기대된다.

나아가 향후 판매가격과 수급 상황 등에 대한 예측을 바탕으로 체계적인 종자 생산이 가능해지고, 수산물 가격 안정성을 확보하는 등 현안 해결에도 기여할 것으로 기대된다.



보건
의료

PART 5

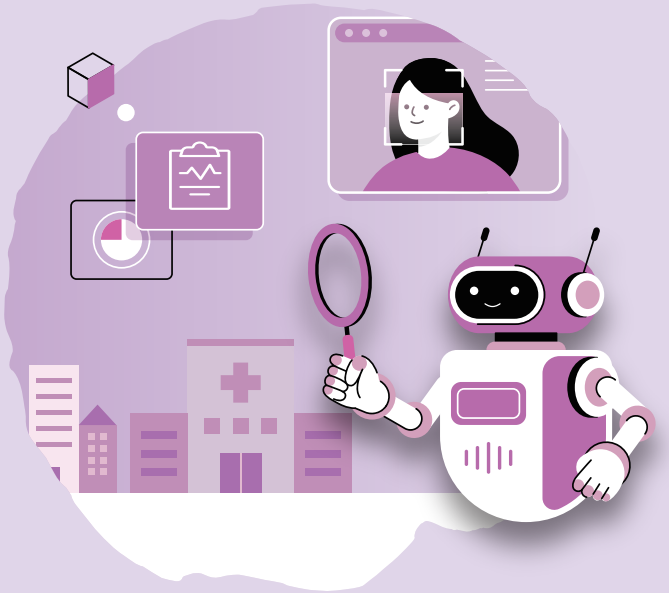
보건의료

PART 5-1
보건의료

상병·요양데이터를 활용한 산재의료 의사결정 지원 모델개발

AI 의학자문 모델 : 산재요양 처리의 신속, 정확성을 불러온 혁신!

근로복지공단



추진목적/배경

최근 업무상 재해 인정기준이 완화되면서 산재요양 신청 건수가 매년 지속적으로 증가하고 있다. 이로 인해 연관 업무 및 산재심사 업무처리 전반에 부담이 가중되고 있다.

현재의 업무 처리 절차는 산재 신청 건이 접수되면 산재 요양기간의 적절성을 판단하는 과정에서 의사의 자문을 활용하고 있으나 약 80%는 판정 결과가 명확한 단순자문에 해당한다. 공단 소속 상근의사와 비상근 위촉 자문 의사가 직접 진료계획서의 적정성 여부를 검토하는데, 이로 인해 신청부터 결과 확인까지 약 5.5일가량 소요되어 민원처리 기간 단축이 필요했다.

이를 해결하기 위해 최근 5년간 산재관련 자문 결과 데이터를 토대로 AI기반의 과학적 학습을 통해 단순 의학자문을 대체하고 업무 효율성을 높일 수 있는 모델 개발을 추진하였다.

분석 사전 준비

– 활용 데이터

데이터명	형태	내용	출처	기준 년도	내·외부 데이터
급여원부정보	CSV	원부번호, 연령, 성별 등	근로복지공단	2018 ~ 2022	내부
재해정보	CSV	재해일자, 상병종류코드 등	근로복지공단	2018 ~ 2022	내부
상병코드별 정보	CSV	상병구분, 상병코드 등	근로복지공단	2018 ~ 2022	내부
정렬요양기간	CSV	요양시작·종결일자, 요양구분(최초요양/재요양) 등	근로복지공단	2018 ~ 2022	내부
의학적소견_주치의 진료계획서	CSV	주치의소견횟수, 수술여부 등	근로복지공단	2018 ~ 2022	내부

분석데이터에 개인정보가 포함되어 있음을 고려하여 근로복지공단 내부에서 협의를 진행하였다. 그 결과 개인을 식별할 수 있는 원부번호(ID)를 비식별화하고, 연령 및 성별을 범주화하여 가명 처리된 데이터를 분석에 활용하도록 협의하였다. 분석용 데이터는 가명처리 이후 공단 내 정보화본부와 법무팀의 승인 절차를 거쳐 CSV파일 형태로 반출하였다.

최종적으로 가명처리한 급여원부정보, 재해정보, 상병코드별 정보, 정렬요양기간, 의학적 소견(주치의 진료계획서) 데이터 등을 활용하여 학습용 데이터셋을 구축하고, 이를 AI 모델 개발에 활용할 수 있게 되었다.

분석과정

– 분석 환경

1. 분석 인프라 : 기관 내 PC 이용
2. 분석 환경 : python, MLOPS* 등

* 머신 러닝(Machine Learning)과 운영(Operations)을 합친 용어로 프로덕션 환경에서 머신 러닝(ML) 모델이 지속적이고 안정적으로 배포되도록 유지, 관리, 모니터링 해주는 것

– 데이터 수집

1. 사용 데이터 출처 : 기관 자체 생성 데이터
2. 사용 데이터 형식 : csv

– 데이터 전처리

1. 결측값 및 오류 점검
   - 상병종류코드 등의 결측값을'9999'로 대체
   - 날짜 형식으로 표현되지 않은 데이터 제외
   - 상병코드값에 공백이 있는 경우 공백을 제거하고 활용하며, 대문자 알파벳으로 시작하여 숫자 2~5글자까지 이루어진 값(S02620)만 허용하고 해당하지 않는 경우 제거
2. 이상치 점검
   - 특정 샘플식별자(원부번호)의 최초 요양시작일이 재해일자보다 이전인 경우 해당 데이터 제외 (오류 데이터/이상치로 판단)
   - 요양일수는 요양종료일 – 요양시작일로 재계산하며, 요양일수가 음수인 데이터는 제외
3. 파생변수 생성
   - 교통사고 여부가 요양일수 예측에 주요한 변수인지 확인을 위해 교통사고자 유형코드 값이 있는 경우 1(Y), 없는 경우 0(N)으로 변수

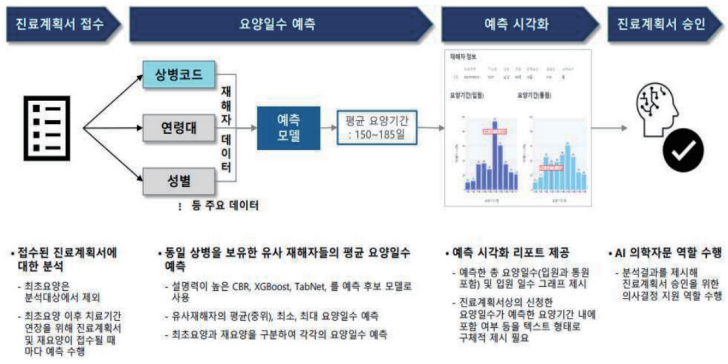
생성. 단, 한 재해자가 교통사고자유형코드와 NA(결측값)을 동시에 가지고 있는 경우 교통사고여부 1(Y)인 것으로 판단

- 병원종류코드가 07인 경우 1(한방병원), 0(한방병원X)으로 변수 생성

– 모델링

1. 사용한 분석 모델
   - 상병·요양 데이터를 활용한 산재의료 의사결정 지원모델은 재해자의 상병코드, 연령대, 성별 등의 정보를 활용하여 치료종결 이전의 요양기간(요양일수)를 예측하는 AI regression 모델 활용
   - 활용 변수, 모델 구조, 요양기간 및 상병코드 처리 방식, 앙상블 적용 등에 따라 통계, 기계학습, 앙상블, 딥러닝 모델과 같은 다양한 모델을 개발하여 성능평가 후 최종 모델을 결정
   - 딥러닝 모델로는 전자의료기록(EMR) 형태의 복합 데이터에 대한 예측에 주로 사용되는 Multimodal-Net을 벤치마킹함

2. 예측 모델 프로세스



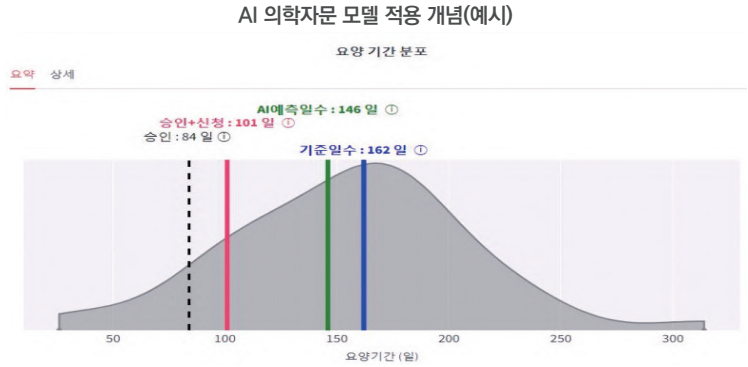
- 진료계획서에 포함된 재해자 정보, 요양기간, 자문의사 소견 등 다양한 데이터를 활용하여 요양기간 예측 모델 개발
- 분석 및 예측결과를 진료계획서 승인을 위한 의사결정에 활용할 수 있도록 유사재해자의 요양기간 분포 그래프와 예측한 총 요양일수를 시각화 서비스로 제공

– 검증 및 고도화

- 모델 성능평가 지표인 RMSE를 사용한 모델별 성능 비교에서 딥러닝 모델(RMSE 58.66)이 가장 높은 성능을 보여 해당 모델로 성능 고도화를 진행함
- AI모델 고도화를 위해 요양일수 이상치 처리 및 스케일링 방식 변경, 유사재해자의 요양일수 중위값 변수 추가, 진료계획서 단위로의 학습용 데이터 단위 변경, 파생변수 생성 및 추가 변수 조정으로 성능 고도화를 진행해 기존성능(RMSE 58.66) 대비 58% 개선(RMSE 24.36)

– 결과 구현

1. 시각화 페이지
   - Python 등을 활용하여 구현
   - 재해 신청자와 유사한 특성을 가진 재해자에 대한 분석 정보 및 AI 예측결과 확인 가능



→ 재해자의 요양일수(승인+신청 101일)가 과거 유사재해자의 기준일수(중앙값 162일)와 AI예측일수(146일) 이내에 해당하여 의사자문 생략가능

2. AI 시스템 배포 및 테스트

▶ AI 요양기간 예측 프로그램 개발

- 해당 프로그램은 11가지 컴포넌트로 구성됨
- 지속적으로 AI모델을 개발하고 관리 및 서비스할 수 있도록 MLOps 구현을 위한 프로그램을 패키징하여 배포
- 데이터 수집 → 데이터 가공 → AI모델 개발 → AI모델 저장 → AI모델 서빙 → 시각화 까지의 전반적인 프로세스를 구현할 수 있도록 프로그램을 구성

▶ 시스템 배포(요양기간 예측모델 및 프로그램 배포)

- 근로복지공단의 주요 시스템 중 AI 모델 개발용으로 활용되는 [지능형 시스템]에 배포
- 요양기간 예측모델을 포함하여, 데이터 수집 및 가공 프로그램, 관리/모니터링툴, 시각화 서비스 등을 패키징하여 배포

정책활용/기대효과

요양기간 예측 모델이 성공적으로 개발됨에 따라 근로복지공단은 AI 모델 결과의 현장 활용을 추진하였다.

먼저 「산업재해보상보험 의학자문 운영 지침」을 개정('24.5.8.)하였다.

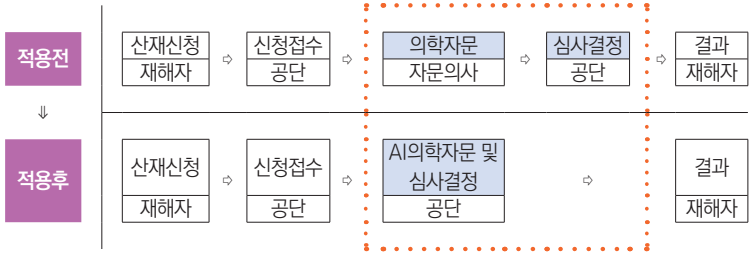
요양일수 산정 시 재해자가 신청한 요양일수가 과거 유사재해자의 요양일수 중앙값과 AI 예측 일수 이내인 진료계획서는 의학자문을 생략하고 AI 모델 결과를 활용하도록 하였다. 그 후 AI 모델을 공단 내부에서 활발하게 활용할 수 있도록 권역별 요양재활 현장 간담회를 개최하여 AI 모델 활용방안 및 지침 개정사항 등을 안내하였다.

이러한 노력을 통해 근로복지공단은 예산 절감과 업무 시간 단축이라는 두 가지 성과를 얻을 수 있었다.

첫째, 약 5개월에 걸쳐 전체 진료계획서 101,242건 중 11,484건(11.7%)을 AI 의학자문 모델로 처리하였고, 11,848건 중 2,291건에 대한 자문을 생략하여 비용을 절감하였다.

둘째, AI 의학자문 모델 활용으로 업무처리 시간이 약 5.5일에서 3.6일로 1.9일(34.5%) 단축되었다.

AI의학자문 실시 전·후 프로세스



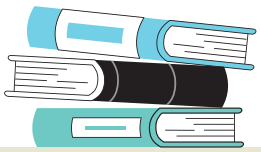
향후 본 모델이 재요양 판단, 장해판정 등 의학자문이 필요한 분야로 확장되어 공단 내 데이터 기반의 객관적 행정이 이루어지기를 기대한다.

2024 공공부문 데이터 분석·활용 우수사례집

발 행 처 | 행정안전부 공공지능데이터분석과, 한국지능정보사회진흥원

엮 은 이 | 행정안전부 공공지능데이터분석과
과장 조아라
팀장 홍성수
주무관 김나은
한국지능정보사회진흥원 지능데이터기반행정팀
팀장 강경훈
책임 김승현

발 행 일 | 2025년 3월 10일



행정안전부 NIA 한국지능정보사회진흥원