데이터 분석 최종결과보고서

Ⅰ. 참가자 정보

제 목	충남・세종・대전 지역 음주운전 사고 분석 및 예측	
팀 명	폴리스노트	
성 명	연세대학교 이창대 외 2명(김세중, 김정수)	
연락처	휴대폰	
	E-mail	

Ⅱ. 개요

○ 분석/시각화 목적

- 충남・세종・대전 지역 음주운전 사고 예측
- 예측기반 음주운전 영향요인 분석
- 영향요인과 관련한 15개 군집 제시
- 음주운전 관련 정책 제안(단속 거점/기간 제안)

○ 배경 및 필요성

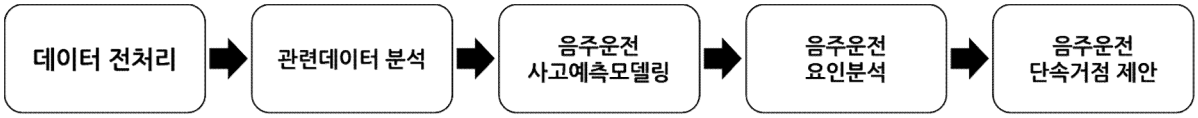
- 충청도 지방의 음주운전 사고의 심각성
  - ■ 주어진 데이터를 사고 유형별로 분류해 본 결과 충청도 지역의 교통사고 중 음주운전 비율이 높게 나옴
  - ■ 교통사고 중 음주운전으로 주제를 세분화한 뒤 전국적인 비율과 비교를 한 결과 비수도권 지역 중에서는 충남의 인구수 대비 음주운전 사고비율이 가장 높은 결과가 도출됨
  - ■ 인구당 음주운전 사고로 인한 부상자 비율은 제주도를 제외하고 충남과 충북이 각각 전국적인 비율 중 가장 높은 것으로 확인됨.
- 기존 지표 교통안전지수의 한계
  - ■ 전국 기초자치단체를 대상으로 교통사고 심각도별 사고건수와 사상자수를 기초로 인구와 도로연장을 고려하여 지자체별 교통안전 수준을 평가한 지수인 교통안전지수는 빅데이터를 활용하여 교통사고와 관련된 요인들을 종합적으로 측정하고 추후의 교통환경을 개선시키기 위한 목적으로 도로교통공단으로부터 매년 발표되고 있지만 해당 자료에는 음주운전 관련 요인이 포함되어 있지 않음.
  - ■ 뿐만 아니라 현재 널리 이용되는 교통사고 관련 지표에는 음주운전을 조명할 수 있는 데이터가 존재하지 않는 실정임.

⇒ 이에 음주운전 사고비율이 높은 충남 지역에서 이를 줄이기 위한 방안으로 교통사고 데이터셋을 이용해 앞으로 발생할 사고를 예측하는 모델을 구축하고자 함

⇒ 더불어 모델에서의 변수중요도를 파악해 세부적인 요인분석을 진행하며 군집을 제시하고, 그에 맞는 정책적 해결방안을 제안하는 것을 계획.

○ 분석/시각화 프로세스

프로젝트 로드맵



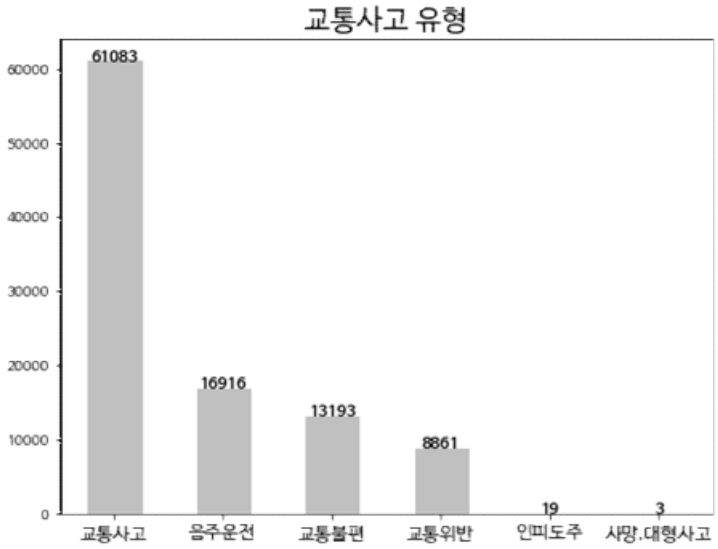
- EDA(Exploratory data analysis)
  - ■ 주어진 데이터를 분석하여 의미있는 인사이트 탐색
  - ■ 그 외 논문, 신문기사, 통계자료 등 음주운전과 연관있는 다른 데이터들을 수집하여 의미있는 인사이트 탐색
  - ■ 최종 모델에 활용될 변수 선택
  - ■ 최종 독립변수 (시간 & 공간):
    - ▶ 지역구(중분류), 거주인구, 지역별 주점개수, 지역별 도로환경지수, 지역별 차량대수
    - ▶ 사건발생 시간대, 휴일여부, 요일, 발생월
- 모델구축
  - ■ 충남, 대전, 세종시 등에서 발생한 교통사고가 음주운전 사고인지 예측하는 모델
  - ■ EDA를 통해 선정된 변수들에 대해 3가지 모델(XGBoost, Randomforest, SVM)을 통해 가장 성능이 좋은 모델을 채택
  - ■ 최종모델: XGboost
  - ■ 성능: 정확도 86%
- 요인분석 및 거점 클러스터링
  - ■ 모델에 Shap Values와 Feature Importance를 통해 음주운전 사고의 중요요인 분석
  - ■ 요인분석에 따라 중요하다고 판단된 변수들을 선별해 클러스터링 진행
  - ■ 진행한 클러스터링을 기반으로 음주운전 단속 거점 제시 및 시사점 도출

Ⅲ. 분석/시각화 결과 상세내용

○ 분석/시각화 결과 상세내용

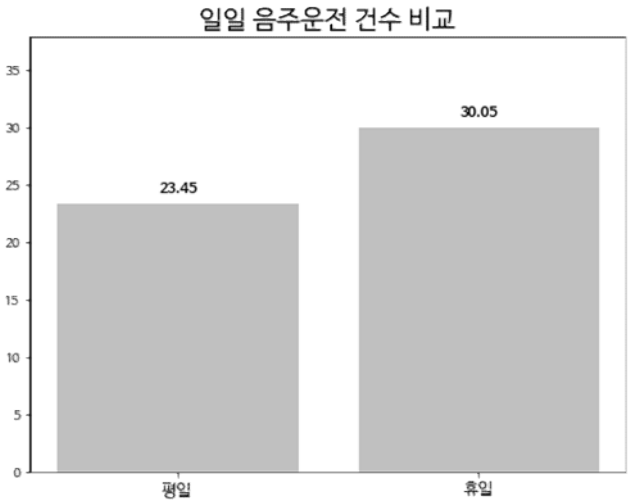
1. 데이터 수집

충남・세종・대전 지역의 2020, 2021년 교통사고 데이터 추출
동일한 사고를 좌표 및 시간을 바탕으로 중복된 값을 제거했고, 결측값을 제거함.
주소 데이터에 오류 및 결측값이 존재하여, Kakao API를 활용해 주어진 X,Y 좌표에서 주소를 끌어냄
⇒ 2020년 12월에서 2022년 2월까지의 교통사고 데이터 총 10만여 건의 데이터 확보

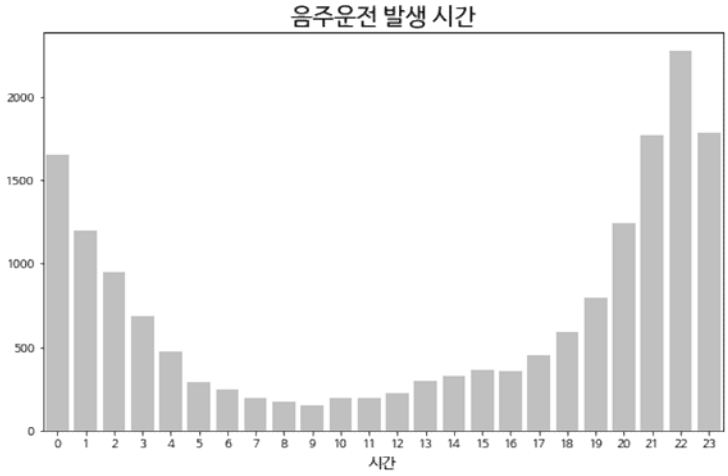


- 교통사고 유형 중 음주운전이 2등으로 28%를 차지. 음주운전이 교통사고의 상당 부분을 차지하므로, 본 연구를 통해 음주운전에 미치는 요인을 분석하고, 음주운전을 예측할 수 있는 모델을 개발하고자 함

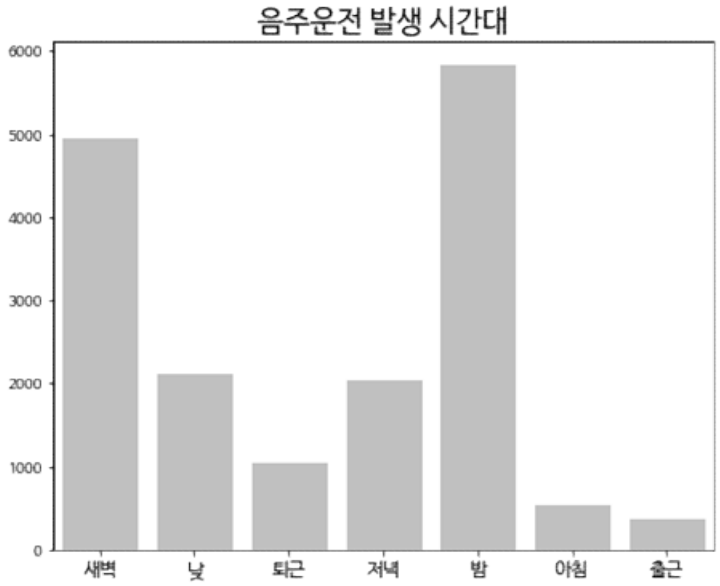
2. 탐색적 분석(EDA) 및 변수탐색



- 충남,세종,대전 지역의 평일, 휴일 일일 음주운전 건수 비교
- 휴일에 더 많은 음주운전이 발생

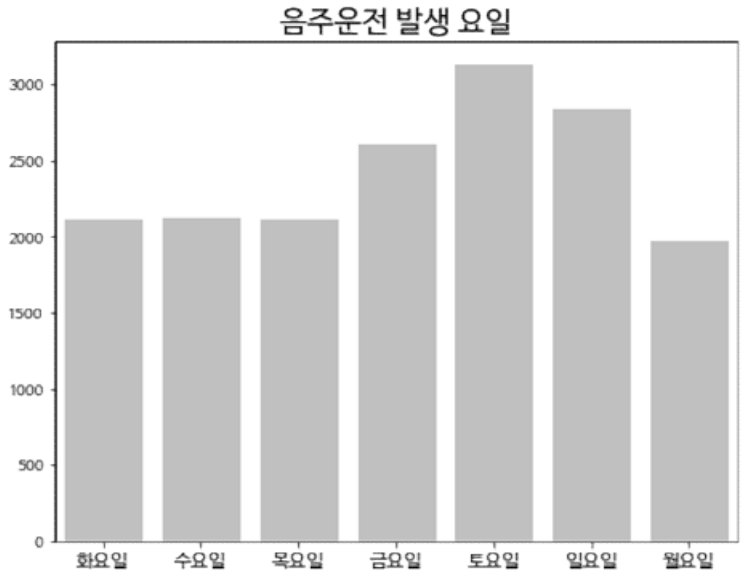


- 시간대에 주기성이 존재
- 낮에 감소했다가 밤과 새벽에 음주운전 사고가 증가하는 것을 알 수 있음

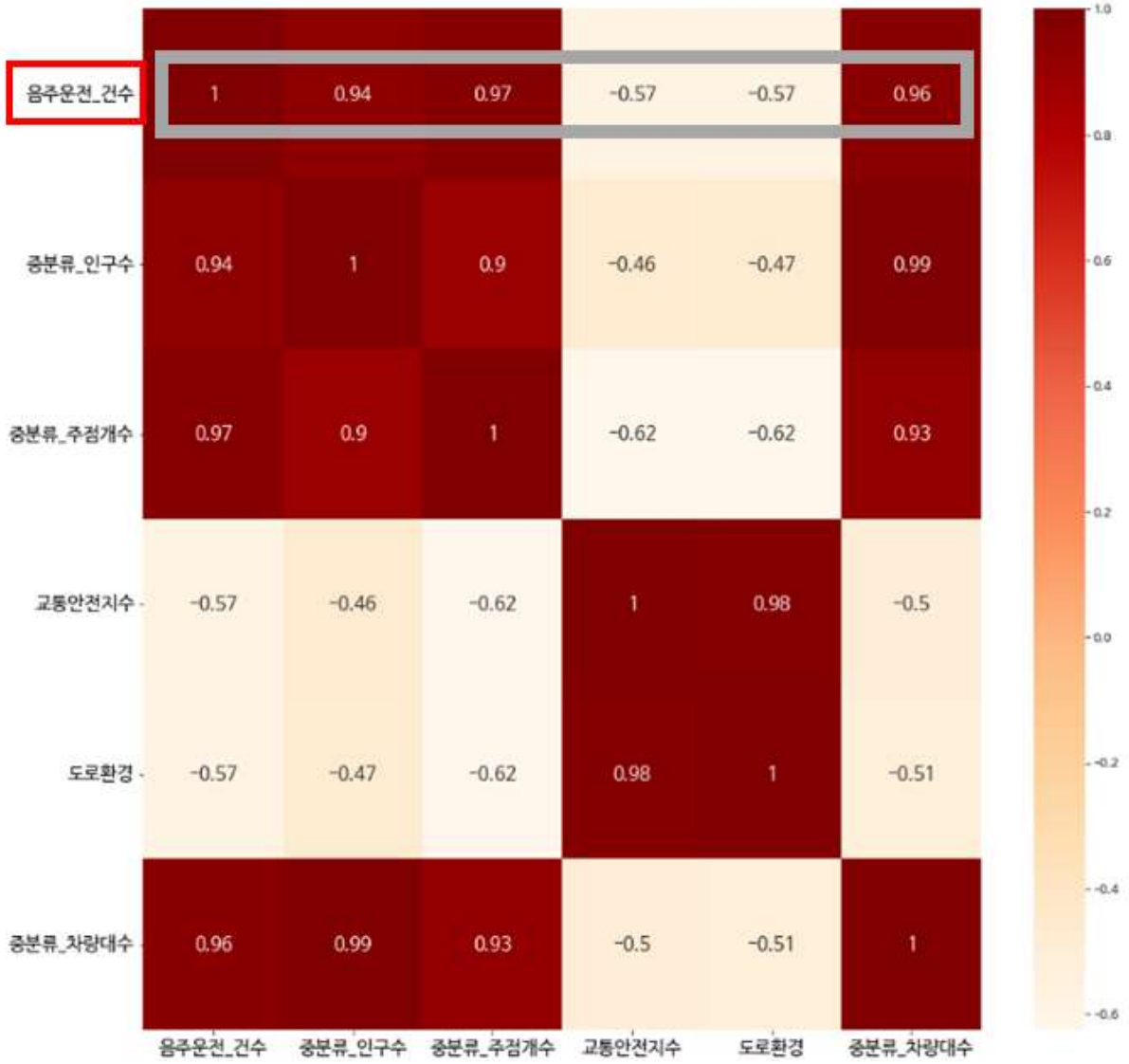


새벽 : 0시 ~ 5시
아침 : 5시 ~ 7시
출근 : 7시 ~ 9시
낮 : 9시 ~ 17시
퇴근 : 17시 ~ 19시
저녁 : 19시 ~ 21시
밤 : 21시 ~ 24시

- 밤과 새벽 시간의 음주운전 발생 빈도가 높음

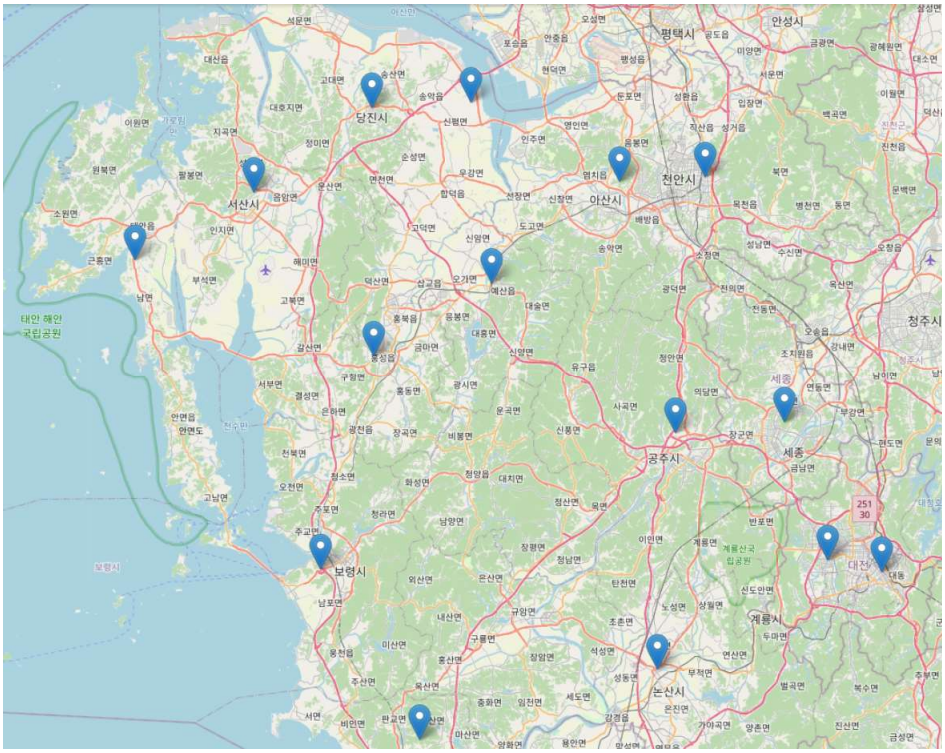


- 음주운전 발생 요일은 금요일에 증가해 월요일에 감소하는 추세를 보임

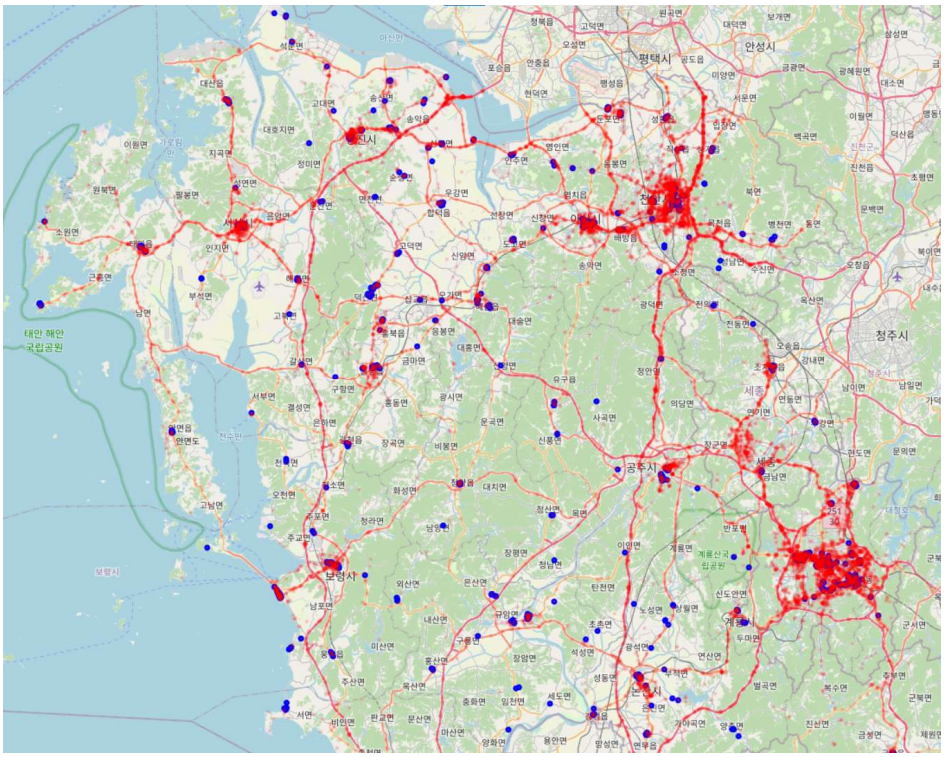


- 주어진 데이터에서 지역별 음주운전 발생건수를 바탕으로 다른 독립변수들과의 상관관계를 조회함
- 교통안전지수, 도로환경지수와 음주운전 건수는 높은 음의 상관관계를 보임
- 인구수, 주점개수, 차량대수와 음주운전 건수는 매우 높은 양의 상관관계를 보임.

3. 지도 시각화



- K-means Clustering을 활용해 X, Y 좌표 데이터에 대해 15개 군집 생성
- 각 군집의 중점을 위 지도에 마커로 표시



- 빨간색 점은 음주운전사고 발생위치, 파란색 점은 주점위치를 의미
- 주점이 몰려 있는 곳에서 주로 사고가 발생했음을 확인할 수 있음
- 위 지도의 군집 중점과 아래 지도의 빨간색 군집(음주운전)이 일치함

3. 모델 구축

- 경쟁모델
  - ■ XGBoost: 트리계열의 부스팅모델로 regularization term을 추가하여 모델의 성능향상을 유도함
  - ■ Random Forest: 트리계열의 배깅모델로 기존의 배깅이 갖고 있던 샘플 간의 종속성 문제를 일부 해결
  - ■ Support Vector Machine (SVM): 커널 기반의 모델로 커널 선택에 따라 분류데이터에 비선형적인 경계선을 주는 것이 가능
    (SVM의 경우 데이터 정규화가 필수적이기 때문에 연속변수를 정규화시킨 데이터를 이용하여 적합)

- 최종모델 XGBoost 선택
  - ■ 요인분석에는 구동방식을 정확히 알 수 없는 블랙박스 모델보다 설명력이 높은 모델이 유리함
  - ■ XGBoost는 트리계열의 모델로 기본적으로는 블랙박스 모델로 알려져 있으나 분기하는 시점에서 얻는 gain의 증가량이나 분기 당시의 weight 등을 기반으로 한 변수중요도를 제공
  - ■ 그 외에도 Shap Values라는 또 다른 변수중요도 지표가 사용 가능함
  - ■ Random Forest나, SVM보다 성능 면에서 우위를 보임

- ⇒ 최종 패러미터 선택: {'objective': 'binary:logistic', 'colsample_bytree': 1.0, 'gamma': 0, 'learning_rate': 0.7, 'max_depth': 9, 'n_estimators': 1000, 'subsample': 0.5}

○ 결과 해석 및 시사점

1. 요인분석

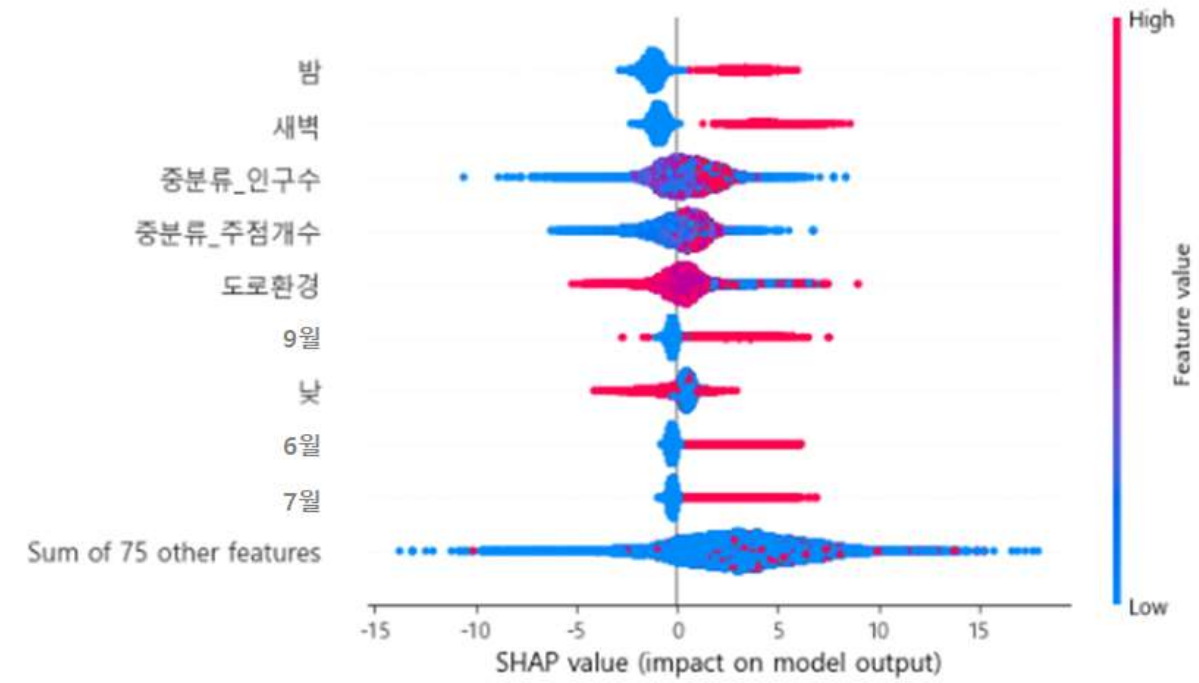
- Gain Feature Importance 결과
  - ■ 지역:
    - □ 연동면과 전동면은 비록 음주운전 사고 수 자체는 적지만 인구당 사고 수는 각각 4위, 6위를 기록
    - □ 연동면과 전동면의 교통안전지수는 81.25로 교통안전지수 상위 10위안에 들지만 실제로는 유의해야 하는 지역임을 알 수 있음
  - ■ 월:
    - □ 11월은 음주운전사고가 제일 적은 월, 12월은 4번째로 적은 월
    - □ 모델 입장에서 음주운전사고가 적게 일어나는 주요 시기를 11, 12월로 판단했음을 알 수 있음
  - ■ 시간대:
    - □ 아침과 출근은 음주운전사고가 가장 적은 시간대
    - □ 새벽은 음주운전사고가 많은 시간대이나 가장 많은 시간대는 아님
    - □ 하지만 모델 입장에서 중요한 시간대로 새벽을 다른 음주운전사고가 많은 시간대보다 가중치를 주었음

지역	변수중요도
연동면	3.71
전동면	3.61
소담동	3.38

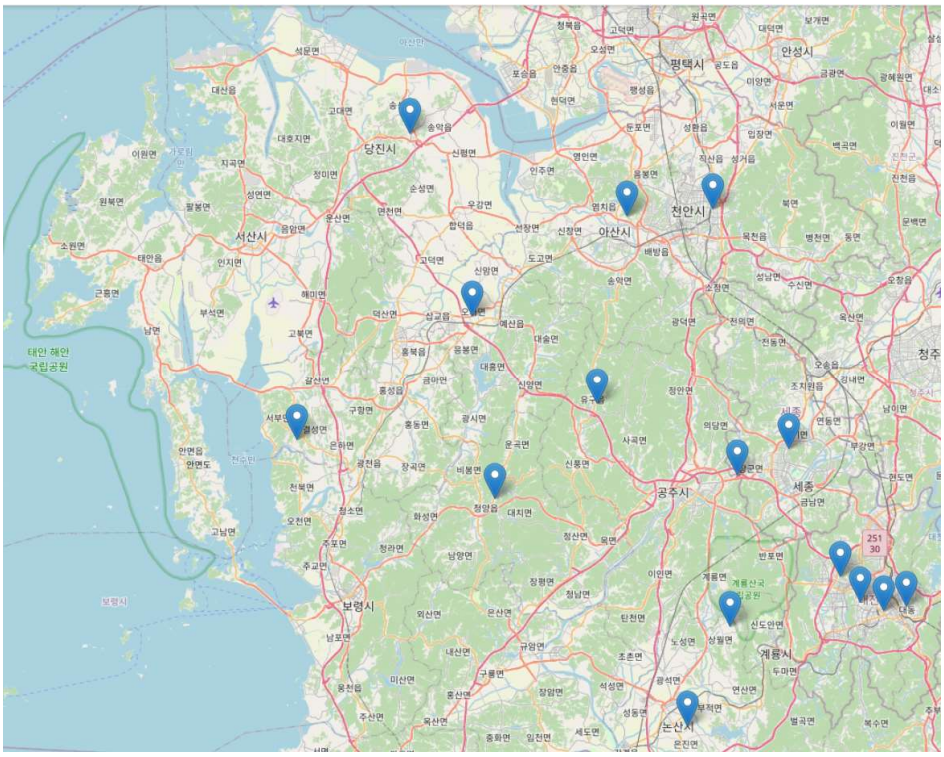
월	변수중요도
12	2.5
11	2.1
8	1.60

시간대	변수중요도
아침	1.96
출근	1.957
새벽	1.85

- Shap Values 결과
  - ■ 밤, 새벽, 낮
    - □ 밤과 새벽일수록 음주운전사고 발생과 상관성 높음
    - □ 즉, 밤, 새벽에 음주운전 단속을 시행하는 기존 시조는 옳다 할 수 있으나 이건 기존 관례로 인한 당연한 결과일 수 있음
    - □ 오히려 낮시간대 음주운전사고 발생에 기여한 부분이 있으므로, 낮시간대 중 특정 시간대에 단속을 하는 것도 나아보임
  - ■ 도로환경
    - □ 도로환경이 좋을수록 음주운전사고 발생이 적음
    - □ 즉, 음주운전사고를 막기 위해서는 전반적인 도로환경 개선을 검토해볼 수 있음
  - ■ 9월, 6월, 7월
    - □ 음주운전이 자주 발생했던 월로, 이 시기에 단속을 강화해야 함을 알 수 있음



- 추출된 변수 기반 클러스터링



■ 좌표 정보와 Shap Values로부터 추출된 중요한 정보 중 지역별 인구수, 지역별 주점개수, 도로환경을 바탕으로 K-means Clustering을 활용해 15개 군집 형성

■ 좌표 정보만으로 형성한 클러스터와 양태가 다르게 나타남

■ 각 군집은 비슷한 인구수, 주점개수, 도로환경을 가진 지역으로 묶임

거점번호	x좌표	y좌표	대분류	중분류	소분류	지역별 인구수	지역별 주점개수	도로환경
1	127.019161	36.806665	충청남도	아산시	염치읍	333602	281	72.59
2	127.150623	36.815879	충청남도	천안시	동남구	657821	420	70.30
3	126.689868	36.90737	충청남도	당진시	송악읍	167955	160	48.15
4	127.37387	36.336881	대전광역시	서구	괴정동	470374	122	75.88
5	127.18792	36.491794	충청남도	공주시	동현동	29862	4	84.42
6	126.819358	36.463361	충청남도	청양군	청양읍	99673	91	81.21
7	127.409308	36.325735	대전광역시	중구	용두동	227108	132	78.88
8	127.343923	36.368324	대전광역시	유성구	궁동	356093	136	78.67
9	126.518596	36.535555	충청남도	홍성군	결성면	61769	58	75.94
10	127.265003	36.523566	세종특별자치시	도담동		8929	3	83.95
11	127.112145	36.185641	충청남도	논산시	은진면	112842	82	72.62
12	126.97457	36.579285	충청남도	공주시	유구읍	174477	98	78.86
13	126.785092	36.685569	충청남도	예산군	오가면	77329	72	64.08
14	127.176894	36.3078	충청남도	논산시	상월면	45853	27	81.31
15	127.444529	36.332239	대전광역시	동구	대동	219751	33	85.86

■ 지도에 표시되어 있는 15개 군집의 중점의 정보는 위와 같음
■ 15개 군집 간 **서로 다른** 지역, 인구수, 주점개수, 도로환경을 가짐
■ 각 군집 내에서는 군집의 중점과 비슷한 정보 분포를 이룸
■ 가령, 15번 군집은 그 중점인 대전광역시 동구 대동에 인접하며, 비슷한 인구수, 주점개수, 도로환경점수를 가짐.

2. 시사점

- 교통안전지수가 높은 세종시 연동면과 연기군 전동면의 음주운전 사고와의 상관성이 높게 나타나므로 **음주운전을 포함시킬 수 있는 교통사고 관련 지수가 필요함**
- 위에서 제시한 군집별 중심을 참조해 **군집에 맞는 음주운전 단속, 정책** 등을 시행할 수 있음. 단순히 행정구역별로 음주운전과 관련된 정책을 시행하는 것보다, 데이터 기반으로 유사성을 지닌 구역별로 음주운전 정책을 진행했을 때 그 성과가 더 분명하리라 판단됨

○ **기대효과**

1. 분석 결과의 활용 가능한 분야
- 효과적인 음주운전 단속에 관한 계획 수립
- 차후 추가될 관련 데이터셋의 모델 학습 방법론 제공

2. 기대효과
- 음주운전 단속과 관련하여 시간적, 공간적 요인분석을 통해 단속 자원에 대한 효율적인 활용
- 사고 발생 빈도가 잦은 지역과 시기에 음주운전 예방정책의 집중을 통해 충남 지역의 음주운전 비율의 전반적인 감소를 도모
- 교통 관련 빅데이터가 추가되고 있는 시점에서 해당 데이터에 맞는 특성 추출 및 학습 알고리즘 선택에 대해 핵심 인사이트 제공

Ⅳ. 기타

○ **건의 사항**

- 교통사교별 사상자 정보(부상자, 사망자) 필요
- 좀 더 긴 기간의 교통사고 관련 데이터 필요

○ 활용 데이터 및 참고 문헌 출처 (필수)

- **공공데이터포털**

공휴일 데이터
https://www.data.go.kr/tcs/dss/selectApiDataDetailView.do?publicDataPk=15012690
행정안전부 단란주점, 유흥주점 영업 데이터
https://www.data.go.kr/dataset/15006701/fileData.do

- **국가통계포털(kosis)**

행정구역별 인구수 데이터
https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1B040A3
자동차 등록대수
https://kosis.kr/statHtml/statHtml.do?orgId=110&tblId=DT_11001N_2013_A033&vw_cd=MT_ZTITLE&list_id=110_11001_006_08&seqNo=&lang_mode=ko&language=kor&obj_var_id=&itm_id=&conn_path=MT_ZTITLE

- **교통사고분석시스템 TAAS**

2022년판 전국 기초자치단체별 교통안전지수(도로교통공단)
2022년판 교통사고 통계분석(도로교통공단)

- **충남, 대전, 세종 지자체 사이트 (인구수 데이터)**

충남
http://www.chungnam.go.kr:8100/cnnet/stats/cnHumanStats.do?mnu_cd=CNNMENU02122
대전
https://www.daejeon.go.kr/sta/index.do
세종
https://www.sejong.go.kr/stat/bbs/list.do?key=1912123631625