대전시 교통사고 위험지역 100개소

(TEAM. 호반우)

01 서론

/ 분석배경
/ 분석주제

02 전처리

/ 데이터분류
/ 데이터통합
/ 모델사용 전처리

03 모델링

/ 인공지능 모델
/ 통계학적 분석

04 분석결과

/ 모델 비교
/ 시각화
/ 분석의의

분석방향성

교통사고 요인 및 비율

*자료 : AASHTO, 2010

원인	도로 및 환경				인적		차량
유형	순수요인	복합요인			순수요인	복합요인	순수요인
		인적	차량	인적·차량		차량	
비율	3%	27%	1%	3%	57%	6%	3%
	34%				63%		3%

AASHTO(미국 주도로 및 교통 행정관 협회)는 교통사고 발생원인을 도로 및 환경요인, 인적요인, 차량요인으로 구분
통제 가능한 요소를 찾기 위해, 도로 및 환경요인 중점 분석

인적요인

운전자 개인성향, 운전성숙도,운전성향, 음주 및 약물 복용 여부 등을 의미(본 분석에서는 인적요인은 통제되었다고 전제)

분석주제

1. 교통사고 위험소 100개 제시
2. 위험지역은 반경50m
3. 연령대별 사고유형과 교통안전 시설물과의 인과관계 고려

분석방향성

여러가지 모델을 활용한 다층적 분석

교통사고 경중이 반영된 가중치를 통한 심도있는 분석

정책적으로 통제가능한 요소를 고려하여
효율적인 교통사고예방 정책 수립에 기여

가중치 반영

ECLO(Equivalent Casualty Loss Only,인명피해 심각도) 개념

: 기상정보와 연계한 교통안전예보지수 개발 및 활용방안에서 활용한 인명피해 심각도 지수
: 교통사고의 경중을 고려할 수 있어 인공지능 모델 가중치로 활용

$$ECLO = (1 * MI) + (3 * MO) + (5 * SE) + (10 * FA)$$

MI는 부상신고자수,MO는 경상자수, SE는 중상자수, FA는 사망자수

PSI(Potential for Safety improvement,잠재력 교통안전개선지수) 개념

: AASHTO에서 제안한 잠재적 안전도 개선 지수

$$PSI_i = Y_i - \hat{Y}_i \quad , \quad \hat{Y}_i = \mathrm{F(VKT)}$$

F는 안전성능함수, VKT는 노출빈도

PSI_i는 종속변수 잔차와 동일

PSI 가 높을수록 안전개선효과가 높음.

$PSI_i > 0$ (예측 사고건수 > 실제 사고건수)
:동일 노출빈도 대비 잠재적 안전개선효과가 높은 지역

ECLO가중치

: 단순 ECLO를 곱하기보다 Min-Max Scailing을 통한 정규화 값 가중

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

PSI가중치

: 실제 위험도가 높은 지역을 탐색하기 위해 식을 다시 세움

$$\text{위험도 점수} = PSI_i * \frac{acci\_cnt}{\max(acci\_cnt)}$$

*PSI가 동일할 경우 실제 사고건수를 가중치


서론 01
/ 분석배경
/ 분석주제
02 전처리
/ 데이터분류
/ 데이터통합
/ 데이터정리
모델링 03
/ 인공지능 모델
/ 통계학적 분석
04 분석결과
/ 모델 비교
/ 시각화
/ 분석의의

데이터 분류

31개의 방대한 데이터가 있어 분류할 필요성 존재. 아래와 같이 4개 카테고리로 범주화

사고 데이터	• 1.대전광역시_교통사고내역(2017~2019) • 2.대전광역시_교통사고격자(2017~2019)
기타 데이터	• 16.대전광역시_기상데이터(2017~2019) • 29.코드정의서
로컬 데이터	• 11.대전광역시_동별_인구현황 • 12.대전광역시_인구정보 • 13.대전광역시_인구정보(고령) • 14.대전광역시_인구정보(생산가능) • 15.대전광역시_인구정보(유소년) • 23.대전광역시_도로명주소(건물) • 24.대전광역시_건물연면적_격자 • 25.대전광역시_법정경계(시군구) • 26.대전광역시_법정경계(읍면동) • 27.대전광역시_행정경계(읍면동) • 28.대전광역시_연속지적도
도로 및 도로시설 데이터	• 3.대전광역시_신호등(보행등) • 4.대전광역시_신호등(차량등) • 5.대전광역시_안전지대 • 6.대전광역시_횡단보도 • 7.대전광역시_도로속도표시 • 8.대전광역시_정차금지지대 • 9.대전광역시_교통안전표지 • 10.대전광역시_교통CCTV • 17.대전광역시_교통링크(2018) • 18.대전광역시_교통노드(2018) • 19.대전광역시_상세도로망(2018) • 20.대전광역시_평일_일별_시간대별_추정교통량(2018) • 21.대전광역시_평일_일별_혼잡빈도강도(2018) • 22.대전광역시_평일_일별_혼잡시간강도(2018) • 31.대전시_중앙분리대

도로 및 도로시설 데이터

Geometry 데이터 (도로시설 및 인구데이터 통합)

3.대전광역시_신호등(보행등)
4.대전광역시_신호등(차량등)
5.대전광역시_안전지대
6.대전광역시_횡단보도
7.대전광역시_도로속도표시
8.대전광역시_정차금지지대
9.대전광역시_교통안전표지
10.대전광역시_교통CCTV
12.대전광역시_인구정보
13.대전광역시_인구정보(고령)
14.대전광역시_인구정보(생산가능)
15.대전광역시_인구정보(유소년)
18.대전광역시_교통노드(2018)
24.대전광역시_건물연면적_격자
30.대전광역시_차량등록현황_격자
31.대전시_중앙분리대

Gid	acci_cnt	…	CCTV	…	TOTAL_p	…	BUILDING_size	geometry
다바 866110	0		0		13		146.5900	MULTIPOLYGON (((127.35…

road_count함수를 활용하여 격자 내 도로시설 카운트
인구자료에 gid가 있으므로 사고격자 데이터와 결합.

Link_id 데이터

19.대전광역시_상세도로망(2018)
20.대전광역시_평일_일별_시간대별_추정교통량(2018)
21.대전광역시_평일_일별_혼잡빈도강도(2018)
22.대전광역시_평일_일별_혼잡시간강도(2018

상세도로망_LinkID기준 결합

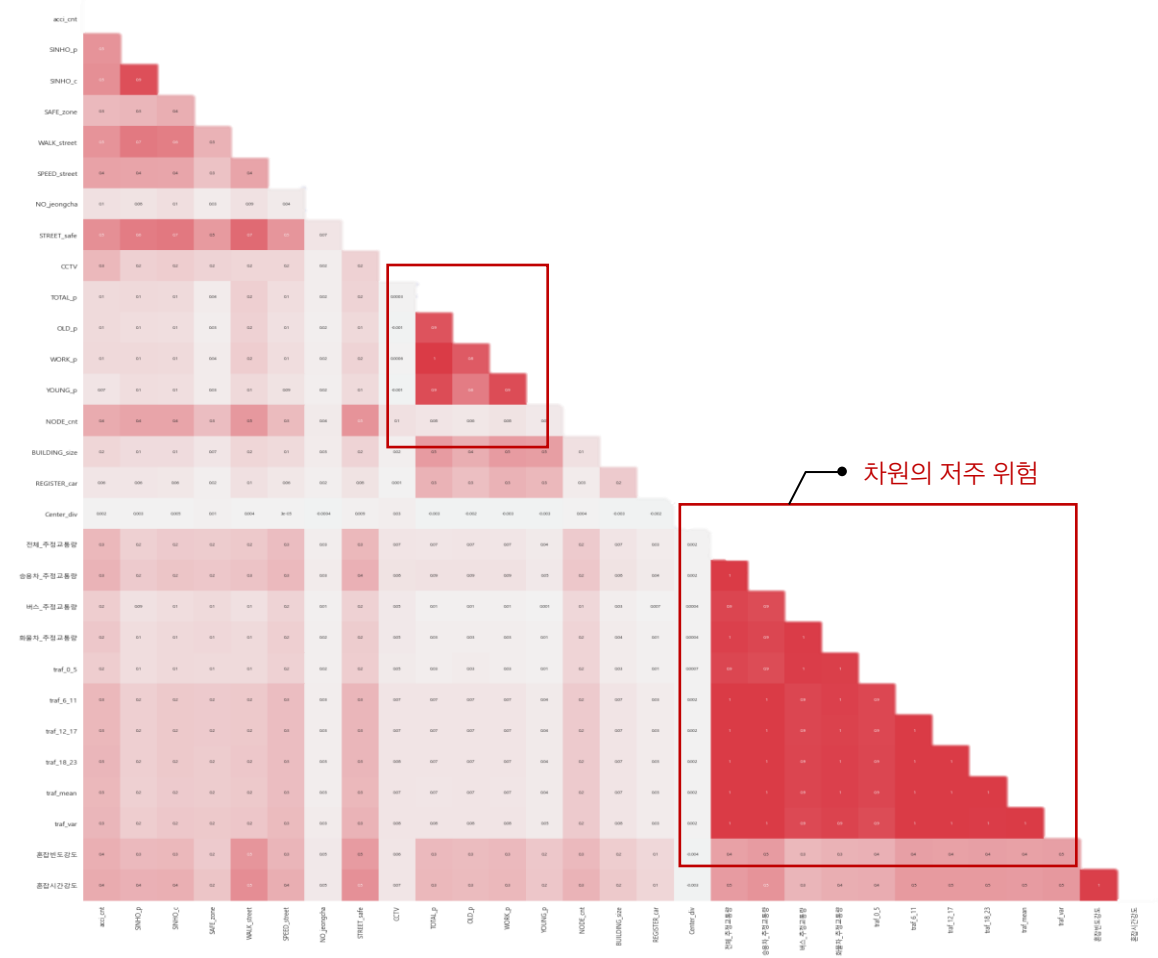
*교통량의 시간대별 데이터 평균,분산 데이터 생성

독립변수 생성 (count기준 : 30개)

보행자 신호등	차량용 신호등	안전지대	횡단보도	전체_추정교통량	Traf_0_5
정차금지지대	속도제한표지	교통안전표시	CCTV	승용차_추정교통량	Traf_6_11
총인구	고령인구	생산가능인구	유소년 인구	버스추정_교통량	Traf_12_17
교통노드	건물연면적	등록차량	중앙분리대	화물차추정_교통량	Traf_18_23
				혼잡빈도강도	Traf_var
				혼잡시간강도	Traf_mean

상관관계 분석

Correlation Plot



차원의 저주와 다중공선성

차원의 저주

변수 개수(차원)이 늘어
모델성능이 저하되는 현상

다중공선성

회귀분석에서 독립변수들 간에
강한 상관관계가 나타나는 문제

VIF 지수

*R은 결정계수

$$VIF = \frac{1}{1-R^2}$$

다중 회귀 모델에서
독립 변수간 상관 관계를 측정

10이상이면 다중공선성 존재.

전체_추정교통량	Traf_0_5	Traf_var
승용차_추정교통량	Traf_6_11	Traf_mean
버스추정_교통량	Traf_12_17	혼잡시간강도
화물차추정_교통량	Traf_18_23	

총인구	고령인구
생산가능인구	유소년 인구

높은 VIF값 → 대체변수로 전환 혹은 구성변수 제거 필요

02 전처리_데이터통합

최종 전처리

- 총인구
- 생산가능인구
- 고령인구
- 유소년 인구

→ 대표데이터인 총인구 데이터 선택 → **총인구**

- Traf_0_5
- Traf_6_11
- Traf_12_17
- Traf_18_23
- Traf_var
- Traf_mean
- 혼잡시간강도

- 승용차_추정교통량
- 전체_추정교통량
- 버스추정_교통량
- 화물차추정_교통량

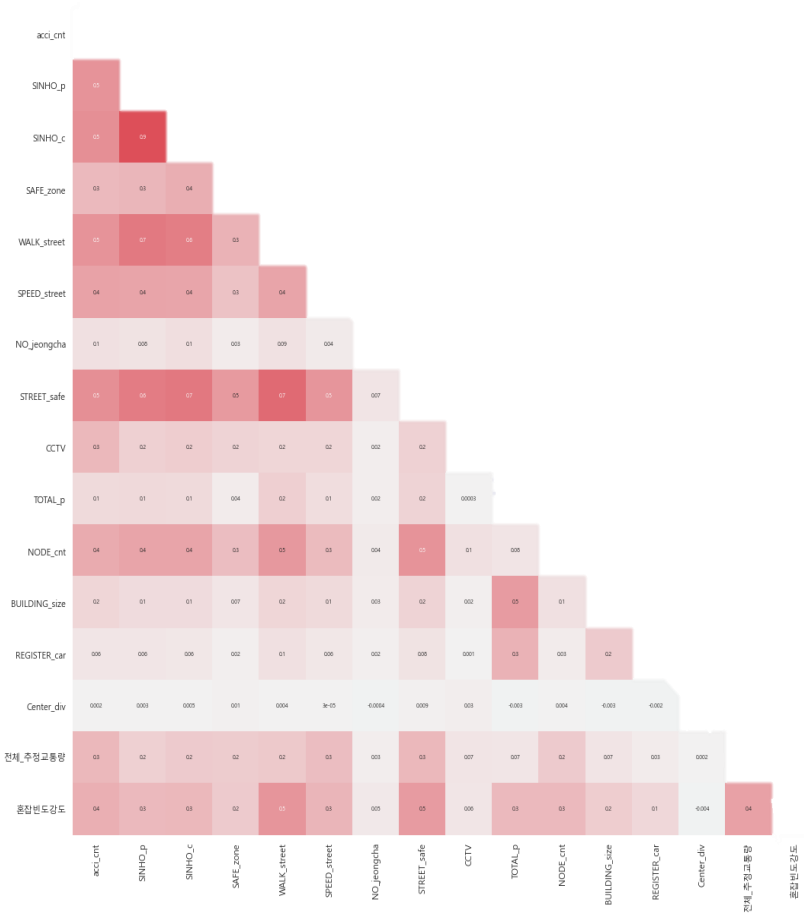
→ 여전히 높은VIF, 전체_추정교통량사용 → **전체_추정교통량**

- Traf_var
- Traf_mean
- 혼잡시간강도

VIF_Factor

변수	VIF_Factor
acci_cnt	1.8594
SINHO_p	5.4343
SINHO_c	5.5126
SAFE_zone	1.3650
WALK_street	3.0373
SPEED_street	1.5108
NO_jeongcha	1.0188
STREET_safe	3.4376
CCTV	1.1331
TOTAL_p	1.5817
NODE_cnt	1.4995
BUILDING_size	1.3580
REGISTER_car	1.1594
Center_div	1.0010
전체_추정교통량	1.4760
혼잡빈도강도	2.0002

최종 데이터 Correlation Plot



최종 데이터셋 분류

도로환경변수	보행자 신호등 수: SINHO_p 차량 신호등 수 : SINHO_c 안전지대 개수 : SAFE_zone 횡단보도 개수 : WALK_street 도로 속도 표지판 개수 : SPEED_street 정차금지지대 개수 : NO_jeongcha 교통안전 표지 개수 : STREET_safe CCTV 수 : CCTV 교통노드(교차로,다리) 개수 : NODE_cnt 중앙분리대 개수 : Center_div
교통환경변수	전체_추정교통량 혼잡빈도강도 일강수량(mm) 일 최심적설(cm) 안개 계속시간(hr)
도시환경변수	건물 연면적 : BUILDING_size 차량등록 현황 : REGISTER_car
인구환경변수	인구정보(총인구) :TOTAL_p

01 서론

/ 분석배경
/ 분석주제

02 전처리

/ 데이터분류
/ 데이터통합
/ 데이터정리

03 모델링

/ 인공지능 모델
/ 통계학적 분석

04 분석결과

/ 모델 비교
/ 시각화
/ 분석의의

모델링 흐름


인공지능 모델
데이터셋 설정
• 최종 데이터셋 활용
효율적 예측 모델 선별
• 앙상블 기법활용
• 과적합 방지
최종 모델 선정
• XGB Regressor
• ECLO가중치 부여
통계 모델
데이터셋 설정
• 추가 데이터 전처리
다양한 모델 사용
• OLS
• 포아송
• 음이항
• SLM
• SEM
• GWR
최종 모델 선정
• GWR
• PSI 가중치 부여
두 모델의 R^2비교

종속변수 설정

사고건수(acci_cnt) 변수에 nearby가중치 부여

- 사고지점이 2개 이상의 격자에 **중첩될 가능성 고려**
  -> 종속변수 acci_cnt에 nearby가중치 부여 결정
- 특정 지역 쏠림 완화
- 사고 건수 X 가중치를 통해 종속변수 도출

기준 범위

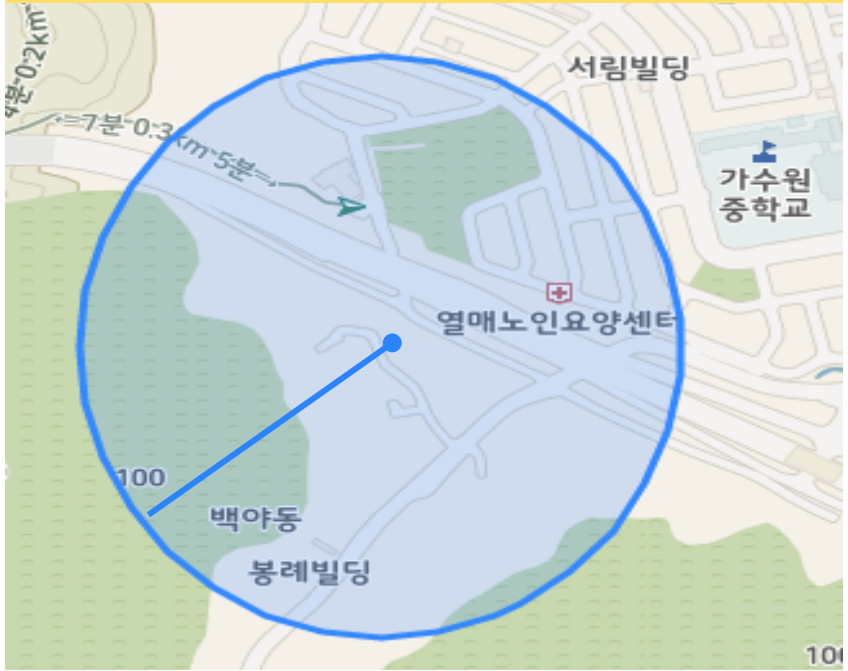
교통사고분석시스템(TAAS)의
사고다발지 기준 -> [반경300M]반영

가중치 수식

$$\frac{x['inverse\_d'] - x['inverse\_d'].min}{x['inverse\_d'].\max() - x['inverse\_d'].min}$$

격자주변 반경300m 가중치 부여

다바866110 기준



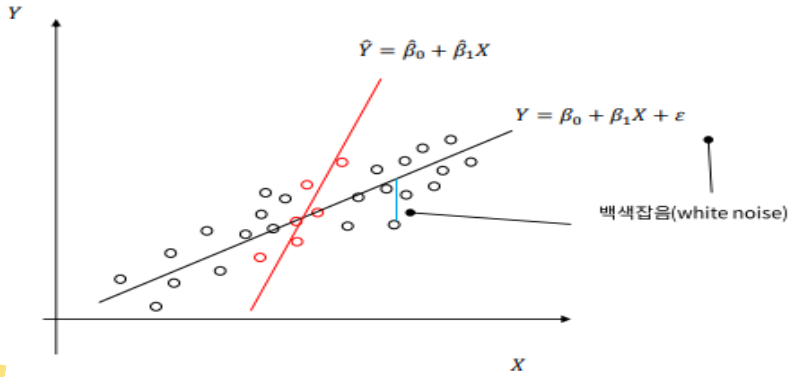
인공지능모델 구성

종속변수	**거리 가중치가 부여된 acci_cnt**
독립변수	• SINHO_p • SINHO_c • SAFE_zone • WALK_street • SPEED_street • NO_jeongcha • STREET_safe • CCTV • TOTAL_p • NODE_cnt • BUILDING_size • REGISTER_car • Center_div • 전체_추정교통량 • 혼잡빈도강도 • 사망자수 • 중상자수 • 경상자수 • 부상신고자수 • 일강수량 • 일 최심 적설 • 안개계속시간
인공지능모델 6가지	✓ Linear Regression ✓ Random Forest Regressor ✓ XGB regressor ✓ SVM Regressor ✓ Voting ✓ LIGHT GBM

인공지능 모델 (회귀변수)

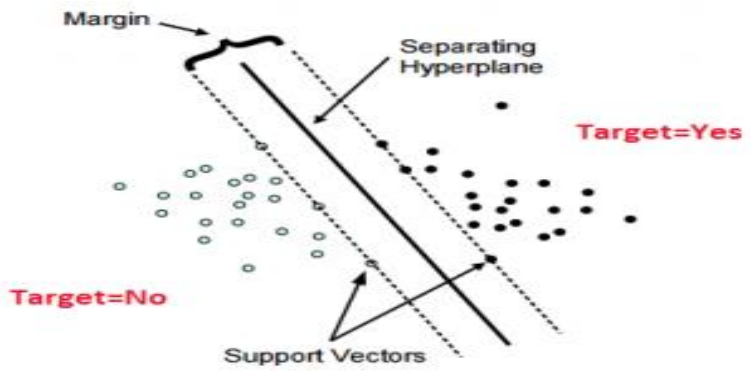
Linear Regression

: 독립변수-종속변수가 선형적 관계일때 가능한 모델.
직선을 통해 종속변수를 예측함.



SVM

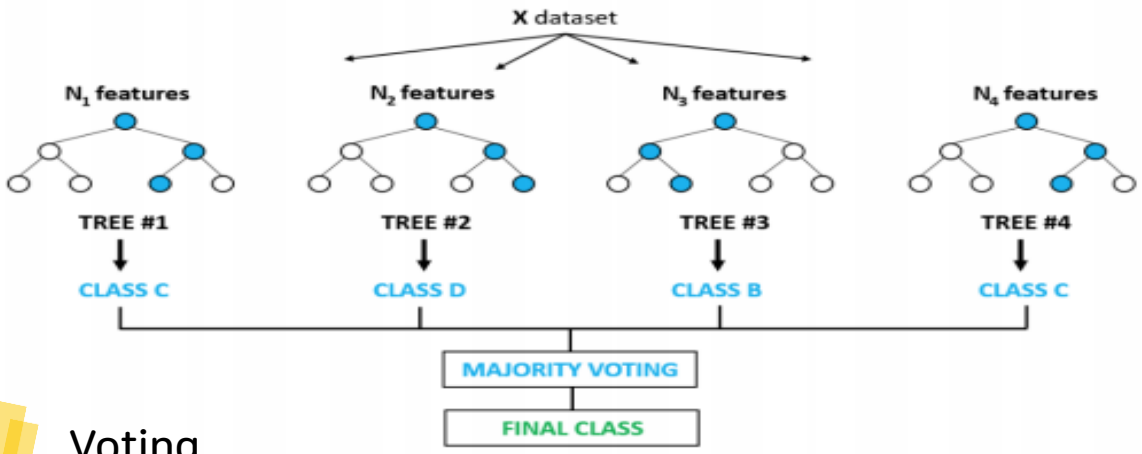
: class간의 거리가 최대가 되도록 decision boundary를 만드는 방법.



인공지능 모델 (앙상블)

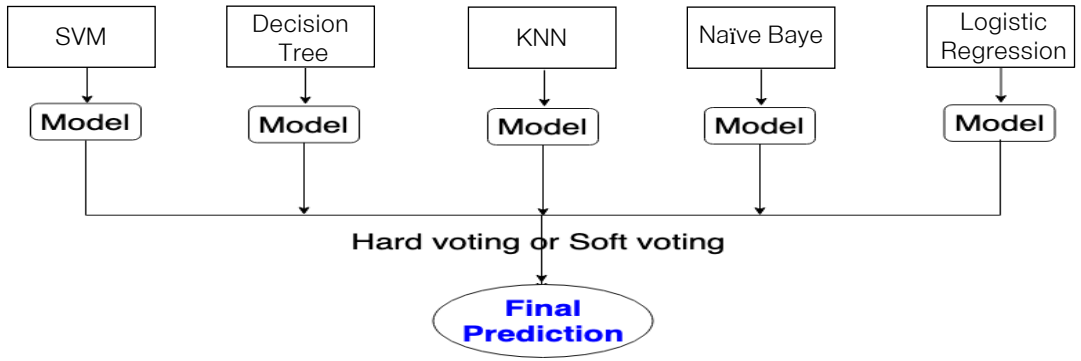
XGBoost, LightGBM, RandomForest

: 약한 학습기를 순차적으로 학습을 하되, 이전 학습에 대하여 잘못 예측된 데이터에 가중치를 부여해 오차를 보완해 나가는 방식



Voting

: 여러 모델의 앙상블에서 다같이 훈련하고 투표를 통해 결과 도출



모델 적합 과정

Regressor모델 사용

그룹이 나뉘어져 있는 상태에서 맞춰가는 분류모델 보다 회귀 등을 사용하는 예측모델이 정확도면에서 더 뛰어나다고 판단

앙상블기법 3가지 포함

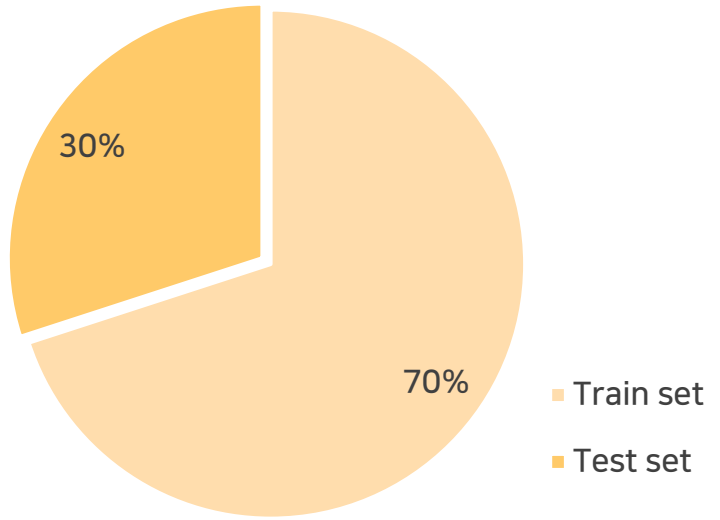
앙상블 : 여러가지 모델을 생성하고, 예측들을 결합하여 예측력을 결합함으로써 정확한 예측을 도출하는 학습법

앙상블은 일반적으로 보팅(Voting), 배깅(Bagging), 부스팅(Boosting) 세 가지의 유형으로 구분

2가지(LGBM, XGBM) 부스팅 기법과 보팅(Voting)사용

Train Test split

- Train 데이터를 100% 학습시키면 성능이 좋지 않음.
- Overfitting 발생 가능성
- 오차를 고려하여 Train data를 70%설정



모델 결과 분석

	Linear Regression		Random Forest Regressor		XGB Regressor		Light GBM		SVM Regressor		Voting	
	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST	TRAIN	TEST
R^2	0.6650	**0.6381**	0.9670	**0.7497**	0.9833	**0.7868**	0.8972	**0.7655**	0.5100	**0.4905**	0.9441	**0.7787**
Adjusted R^2	0.6648	**0.6376**	0.9670	**0.7494**	0.9833	**0.7865**	0.8972	**0.7651**	0.5097	**0.4898**	0.9440	**0.7784**
MAE	1.4741	1.4970	0.3842	1.0330	0.3404	1.0240	0.8297	1.1076	1.2251	1.2748	0.6089	1.0323
MSE	12.9656	13.6063	1.2756	9.4075	0.6432	**8.0154**	3.9765	8.8170	18.9685	19.1564	2.1631	8.3182
RMSE	3.6007	3.6886	1.1294	3.0671	0.8020	2.8311	1.9941	2.969	4.3552	4.3768	1.4707	2.8841

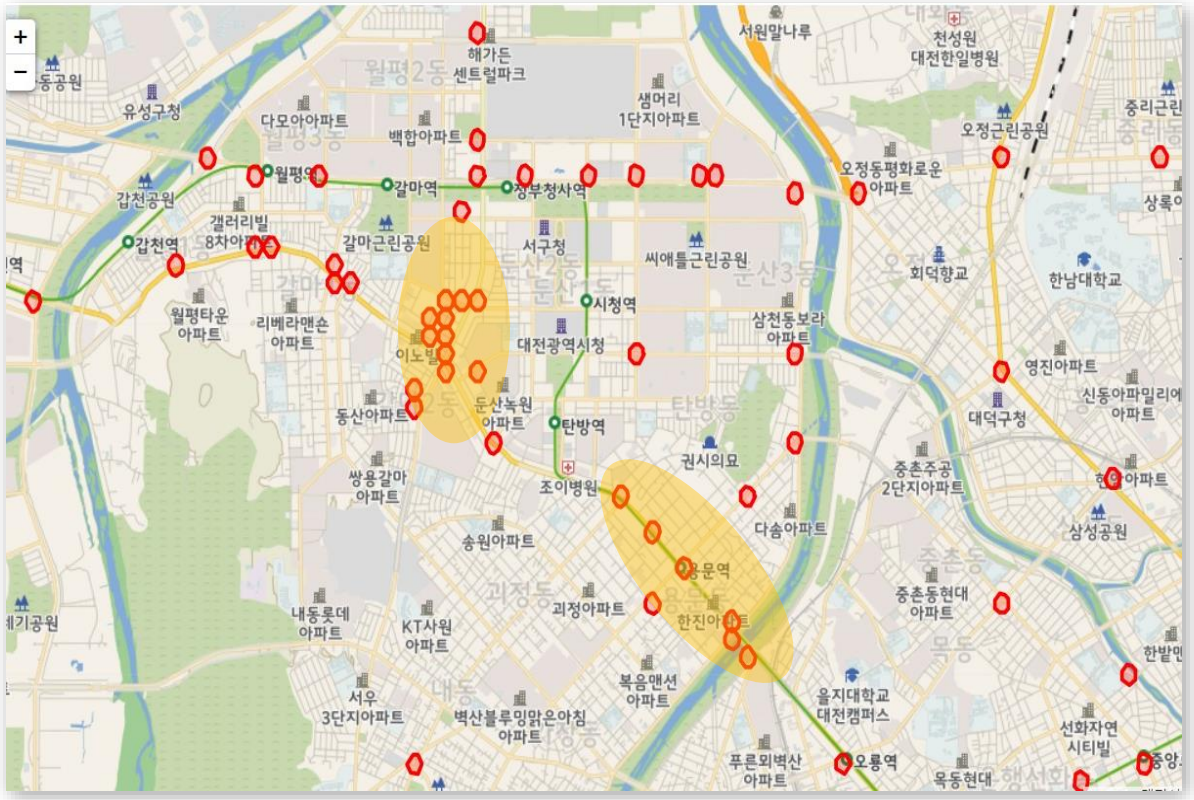
R Square 값이 가장 높고 MSE 값이 가장 낮은
XGB Regressor를 최종 모델로 선정

결과도출 : 교통사고 위험지역 예측 100개소

ECLO minmax 가중치 부여

- 교통사고의 경중을 고려하여 가중치 부여
- 구간별 ECLO차이를 조정하기 위해
  Min-Max 정규화값 적용
- 가중치 이전과 특정장소 차이를 보임.
  특히 갈마네거리, 용문역 등은 인명피해를
  반영했을 때 큰 사고가 발생할 가능성이 높음

100개 장소 샘플



통계모델 분석 개요(OLS,포아송,음이항)

독립변수 조정

: 모델의 효율성을 위해 일부변수 변동

- 상관계수가 낮은 기상데이터 제거 후 교통환경변수 추가, PSI 지수 도입
- 사망자 및 부상자는 ECLO지수 대체

종속변수

공간모델과 일관성있는 분석을 위해
기존 종속변수인 사고건수(acci_cnt)사용

사고유형	법규위반	사고내용	가해운전자 차종
가해운전자 연령대	피해운전자 차종	피해운전자 연령대	세부사고유형
SINHO_c	SAFE_zone	WALK_street	STREET_speed
TOTAL_p	CCTV	NO_jeongcha	STREET_safe
SINHO_p	Center_div	혼잡빈도강도	BUILDING_size
NODE_cnt	ECLO	전체_추정교통량	REGISTER_car

인구환경변수	총 인구수:Total_p	인명피해가중치: ECLO
도시환경변수	건물 연면적 : BUILDING_size	차량등록 현황 : REGISTER_car
교통환경변수	전체_추정교통량	혼잡빈도강도
도로환경변수	사고유형 법규위반 사고내용 가해운전자.차종 가해운전자.연령대 피해운전자.차종 피해운전자.연령대 세부사고유형 CCTV 수 : CCTV	보행자 신호등:SINHO_p 차량 신호등:SINHO_c 안전지대 개수 :SAFE_zone 횡단보도 개수 : WALK_street 중앙분리대 개수 : Center_div 도로 속도 표지판 개수 : SPEED_street 정차금지지대 개수 : NO_jeongcha 교통노드(교차로,다리) 개수 : NODE_cnt 교통안전 표지 개수 : STREET_safe

통계모델 분석 개요(전역,국지적 지리모델)

독립변수 조정

: 모델의 효율성을 위해 일부변수 변동

- 상관계수가 낮은 기상데이터,차종, 성별, 사고유형 제거 후 교통환경변수 추가, PSI 지수 도입
- 사망자 및 부상자는 ECLO지수 대체

종속변수

: 기존 종속변수 사고건수(acci_cnt)사용

버퍼가중치와 공간모델의 공간자기상관성의 중복가능성 배제

가동 데이터셋

SINHO_c	SAFE_zone	WALK_street	STREET_speed
TOTAL_p	CCTV	NO_jeongcha	STREET_safe
SINHO_p	Center_div	혼잡빈도강도	BUILDING_size
NODE_cnt	ECLO	전체_추정교통량	REGISTER_car

인구환경변수	총 인구수:Total_p	인명피해가중치: ECLO
도시환경변수	건물 연면적 : BUILDING_size	차량등록 현황 : REGISTER_car
교통환경변수	전체_추정교통량	혼잡빈도강도
도로환경변수	보행자 신호등:SINHO_p 차량 신호등:SINHO_c 안전지대 개수 :SAFE_zone 횡단보도 개수 : WALK_street 중앙분리대 개수 : Center_div	도로 속도 표지판 개수 : SPEED_street 정차금지지대 개수 : NO_jeongcha 교통노드(교차로,다리)개수 : NODE_cnt 교통안전 표지 개수 : STREET_safe CCTV 수 : CCTV

최소제곱법 (OLS: Ordinary Least Squares)

OLS모델 개념

: 가장 기본적인 결정론적 회귀방법으로 **잔차제곱합을 최소화**하는 가중치 벡터를 행렬미분으로 구하는 방법

$$\hat{\beta} = (X^T X)^{-1} X^T y = \left(\sum X_i X_i^T\right)^{-1} \left(\sum X_i y_i\right)$$

추정하고자 하는 모수 β에 대한 표현식을 다음과 같이 구할 수 있다.

적합도가 가장 큰 표본회귀선이란 오차항의 합이 가장 작은 회귀선이다. 이러한 최적의 표본회귀선을 구하는 방법중 가장 많이 사용되는 추정법이 최소자승법(OLS: Ordinary Least Squares)이다.

산출식

종속변수: 격자별 교통사고 건수

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{24} x_{24} + \varepsilon$$

단계적 회귀분석

독립변수를 하나씩 추가/제거하여 **종속변수를 잘 예측하는 변수**들을 선택하는 기법.
통계적으로 예측력이 유의미한 예측변수들만을 골라준다.

전진 방식과 후진 방식의 결과가 항상 같은것이 아니라 둘 다 비교해 보았다.
가장 낮은 AIC가 도출 된 총 24개의 변수들의 조합을 모델링에 사용했다.

최종 변수_24개와 베타 계수(표시없으면 ***)

사고유형	2.291142	SAFE_zone	0.261570
법규위반	0.064715	WALK_street	0.220501
사고내용*	0.252499	STREET_speed	0.186160
사망자수	1.952419	NO_jeongcha	7.683925
경상자수*	0.140580	STREET_safe	0.430693
가해운전자 차종	0.199281	CCTV	5.602762
가해운전자 연령대	0.035748	TOTAL_p	0.006995
피해운전자 차종**	0.074441	NODE_cnt	1.856949
피해운전자 연령대	0.041322	BUILDING_size	0.000232
세부사고유형	0.144070	Center_div*	6.137108
SINHO_p	0.235052	전체_추정교통량	0.000303
SINHO_c	0.873390	혼잡빈도강도	0.026400

변수 중요도 표시 :
0 '***' (표시없으면 ***) 0.001 '**' 0.01 '*' .'

OLS모델 검증

OLS 기본 가정 (Gausss-Markov 정리)

1. 선형성 (Linear)
2. 오차항 독립성
3. 등분산성 (homoskedasticity)
4. 오차항 정규성 (normality)

4가지 기본가정을 만족해야
유의한 회귀모델이 나온다.

다중공선성 확인 (VIF 10 이상이면 다중공선성)

단계적 회귀분석은 다중공선성 문제에 대처불가
→ **다중공선성** 문제 없음.

등분산성 검정 (Breusch-Pagan test)

BP = 2371.9, df = 24, p-value < 0.00000000000000022

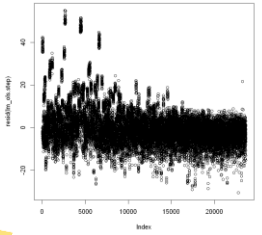
오차항의 모든 분산은 모든 관찰치에서 일정할 것**(등분산성)**
귀무가설 위배 → **이분산성**을 지님을 알 수 있음

p-value < 0.05 ⇒ 이분산성 (Heteroskedasticity)

독립성 검정 (Durbin-Watson test)

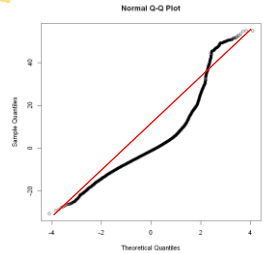
DW = 0.21647, p-value < 0.00000000000000022

d값 이 0에 가까우므로 오차항 사이 상관이 있음



역메가폰 형태 + DWtest
→ 독립성 가정 위배된다.

정규성 검정 (qqplot)



어느정도 정규성을 지녔으나 부족
→ 정규성 가정 위배된다.

OLS모델 검증 후 개선방안

OLS 검정결과

1. 선형성 (Linear)
2. 오차항 독립성
3. 등분산성 (homoskedasticity)
4. 오차항 정규성 (normality)

독립성, 등분산성, 정규성
3가지 조건 위배
→ 다른모델의 필요성

개선방안으로 제시된 2가지 모델

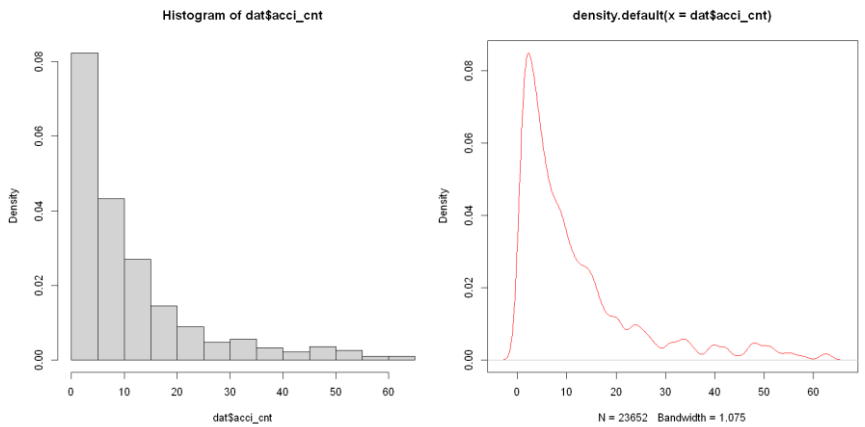
포아송 회귀모델 (Poisson Regression)

acci_cnt(교통사고건수) 같은 도수데이터를 종속변수이자 오차가 정규분포, 등분산이 아닌 회귀모형에서는 **포아송 회귀모형**이 널리 활용된다.

음이항 회귀모델 (Negative Binomial Regression)

포아송은 평균과 분산이 동일해야하는 조건이 있으며, 실제자료에선 과대산포문제 가 일어나기 때문에 이를 조절할 수 있는 **음이항 회귀모형**을 활용

반응변수 분포



0에 치우쳐진 롱테일 형태의 도수데이터

특정 시간 동안 발생한 교통사고 건수에 대한 도수 자료(count data)를 목표변수(acci_cnt)로, 특히 평균이 10 미만일 경우, 최소제곱법(OLS) 회귀모형을 적합하면 **표준 오차와 유의수준이 편향되는 문제**가 발생한다. OLS 회귀모형은 오차의 조건부 정규성(normality), 등분산성(homoscedasiticity), 독립성(independence)을 가정한다. 하지만 평균이 작은 도수 데이터는 0 이상의 정수만 있고, 작은 계급에 많은 관측치가 몰려있으며, 롱테일 형태를 가져 **회귀모형의 가정을 위배하고 음수값도 결과값으로 반환하는 문제**가 있다.

포아송 회귀모델 (Poisson Regression)

포아송 회귀모델 개념

: 포아송 회귀모형은 일반화선형모형(GLM)의 한 종류로서, 최대가능도추정(Maximum Likelihood Estimation)을 통해 모수를 추정

$$ln(\hat{\mu}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

X_1이 한단위 증가하면 도수의 추정치 $ln(\hat{\mu})$ 는 β_1만큼 승법적으로 변화

포아송 회귀모델 평가방식

: 설명변수의 추가에 따른 **이탈도의 감소 비율(proportional reduction in deviance)로 표현되는 pseudo-R^2**를 사용하여 완전모델(perfect model)에 얼마나 가깝게 적합이 되었는지를 평가한다.

$$pseudo-R^2 = R^2_{deviance} = 1 - \frac{deviance(fitted_{model})}{deviance(intercept_{only})}$$

단계적 회귀분석 후 결정된 최종 변수 26개와 베타 계수(표시없으면 ***)

사고유형	법규위반	사고내용	사망자수	중상자수
-0.20449	-0.0056	-0.02974	-0.22377	0.023397
경상자수	부상신고자수 **	가해운전자.차종	가해운전자.연령대	피해운전자.차종
0.012103	0.016609	-0.01927	-0.00268	-0.008
피해운전자.연령대	세부사고유형	SINHO_p	SINHO_c	SAFE_zone
-0.00358	-0.01029	0.004514	0.052767	-0.00748
WALK_street	STREET_speed	NO_jeongcha	STREET_safe	CCTV
0.02482	0.026219	0.355799	0.028685	0.179942
TOTAL_p	NODE_cnt	BUILDING_size	Center_div*	전체_추정교통량
-0.00088	0.175846	0.000016	-0.20091	0.000031
혼잡빈도강도				
0.004182				

변수 중요도 표시
0 '***' (표시없으면 ***) 0.001 '**' 0.01 '*'

포아송 회귀모델 과대산포 검정

포아송 회귀모델은 평균과 분산이 동일해야하는 조건을 가진다.

0.05 이하 :과대산포

z = 29.218, p-value <0.00000000000000022

dispersion : 7.233085

과대산포의 경우 적절하지 않으며, 2개의 모수를 가져서 과대산포분포를 적합할 수 있는 음이항 분포 사용가능.

대안 모델

- 음이항 회귀
- 과대산포 포아송 회귀

→ **음이항 회귀모델 채택**

음이항 회귀모델 (Negative Binomial Regression)

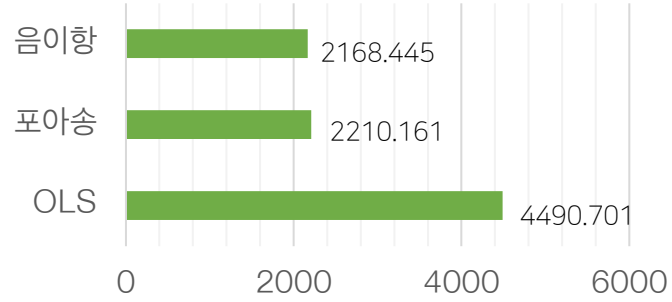
음이항 회귀모델 개념

: 음이항회귀모형은 평균에는 영향이 없으면서 과대산포를 유발하는, 설명이 되지 않는 추가적인 가변성이 있다고 가정한다. 음이항 회귀모형은 똑같은 설명변수 값을 가지는 관측치가 다른 평균 모수를 가지고 포아송 회귀모형에 적합될 수 있도록 해준다.

$$P(y \mid X) = \frac{\Gamma(y+\alpha^{-1})}{y!\,\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y}$$

α 는 과대산포 범위값을 표기한다. 오차 함수는 포아송 분포와 감마분포의 혼합 분포이다.

OLS, 포아송, 음이항 회귀모델 모형적합도 비교



모델 선택을 위해서는 AIC(Akaike Information Criterion)를 사용하며, AIC의 값이 작은 모델을 선택하는데, 이에 따라 음이항모델이 가장 적합하다 판정.

단계적 회귀분석 후 결정된 최종 변수 26개와 베타 계수(표시없으면 ***)

사고유형	법규위반*	사고내용	사망자수	경상자수	가해운전자.차종
-0.25164	-0.00607	-0.02653	-0.26311	0.015966	-0.023268552
가해운전자.연령대	피해운전자.차종	피해운전자.연령대	세부사고유형	SINHO_p ' '	SINHO_c
-0.00348	-0.01031	-0.00456	-0.01292	0.004874	0.05240075
WALK_street	STREET_speed	NO_jeongcha	STREET_safe	CCTV	TOTAL_p
0.037795	0.042809	0.641886	0.026855	0.192204	0.000824047
NODE_cnt	BUILDING_size	전체_추정교통량	혼잡빈도강도		
0.196449	1.77E-05	2.94E-05	0.004245		

변수 중요도 표시 :
0 '***' (표시없으면 ***) 0.001 '**'
0.01 '*' 1 ' '

한계점

" 지리의 제1법칙- 공간상의 객체들은 공간상에 무작위하게 있지 않고 서로간에 영향을 주고받으며 존재한다 " 이와 같이 지리적 공간상에서 **공간객체간 상호의존성과 상호작용을 공간적자기상관**이라고 할 수 있다.

변수들이 무작위적. 오차항이 독립적, 등분산한다는 가정을 하는 기존 선형분석 방법들로 공간데이터를 분석할 경우 많은 인구현상이 공간상에서 나타나는 특성인 **공간적자기상관을 통제하지 못한다.**

기준모형(OLS)을 기준으로 모형 설명력과 모형 오차의 크기 비교 → 교통사고 발생 특성을 가장 잘 나타낸 **전역적 모델**과 **국지적 모델**을 선정

공간계량모델

공간계량모델의 필요성

: 시간의 흐름에 따라 수집되는 시계열 데이터는 자기상관성을 갖게 되는데 공간단위로 수집되는 공간데이터 역시 공간적 종속성(spatial dependence)과 공간적 이질성(spatial heterogeneity)을 갖게 된다.

이러한 공간데이터를 일반선형회귀로 추정시 OLS의 기본가정을 충족시 키지 못해 추정결과에 오류가 발생할 가능성이 있으므로 공간데이터를 설명변수로 하는 공간계량모형의 적용이 필요하다.

기준모델	일반회귀모델		OLS기본가정충족
대안모델	전역적 공간회귀모델	공간시차모델(SLM)	공간상관성 고려
		공간오차모델(SEM)	
	국지적 공간회귀모델	지리적 가중회귀모델 (GWR)	공간상관성 & 공간이질성 고려

공간계량모델 흐름도

공간데이터 전환 후 Queen 방식에 의한 인접지역(이웃) 정의

공간가중치 행렬 생성

전역적 자기상관성 파악 (Morans I 분석)

국지적 자기상관성 파악 (LISA 분석)

공간 자기상관성 존재 시 공간계량모델 3가지 분석

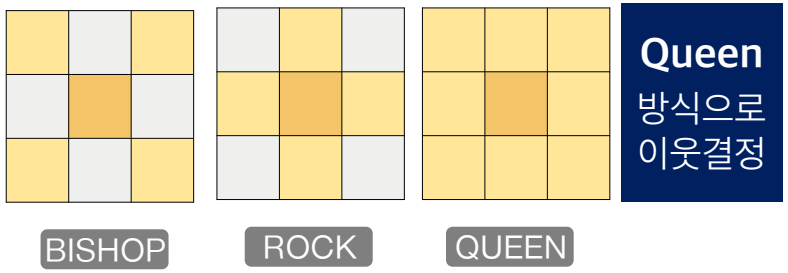
공간 자기 상관성 분석

공간 가중 행렬 생성

: 공간적 상관성을 판단하기 위해서는 공간상에서의 상호작용을 수리적으로 표현한 공간가중행렬이 필요

교통사고건수 시각화
공간적 패턴 확인

→공간데이터로 변환



"인접성 및 거리에 따른 공간 가중치 행렬 생성 후 종속변수(사고건수)의 공간적 자기 상관성 측정"

공간상관성 진단 – 전역적 자기상관성 (Moran's I test)

: 공간가중행렬을 토대로 Moran's I 값을 분석하고 이를 통해 공간적 상관성 여부를 판단한다. Moran's I 값은 -1에서 1 사이의 값을 가지며, 1에 가까울수록 공간적 상관성이 크다.

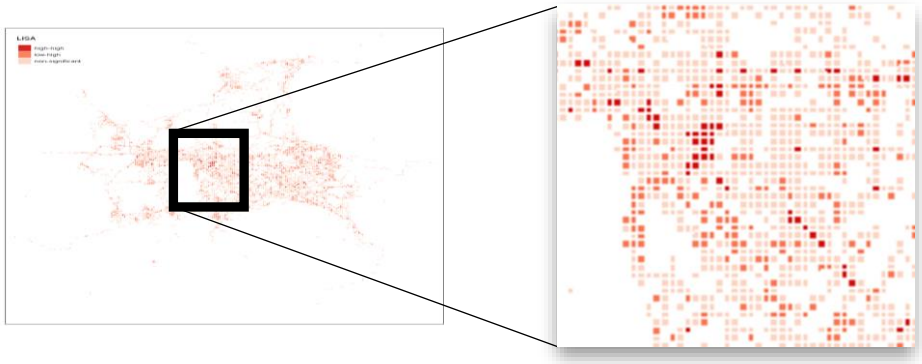
Monte-Carlo simulation of Moran I

statistic = 0.59468, observed rank = 100, p-value = 0.01

0.59468으로 높은 수준의 공간적 자기 상관성이 존재함 .
교통사고가 발생한 곳에 더욱 사고가 많이 발생하며 공간적 군집을 이루는 것 확인가능.

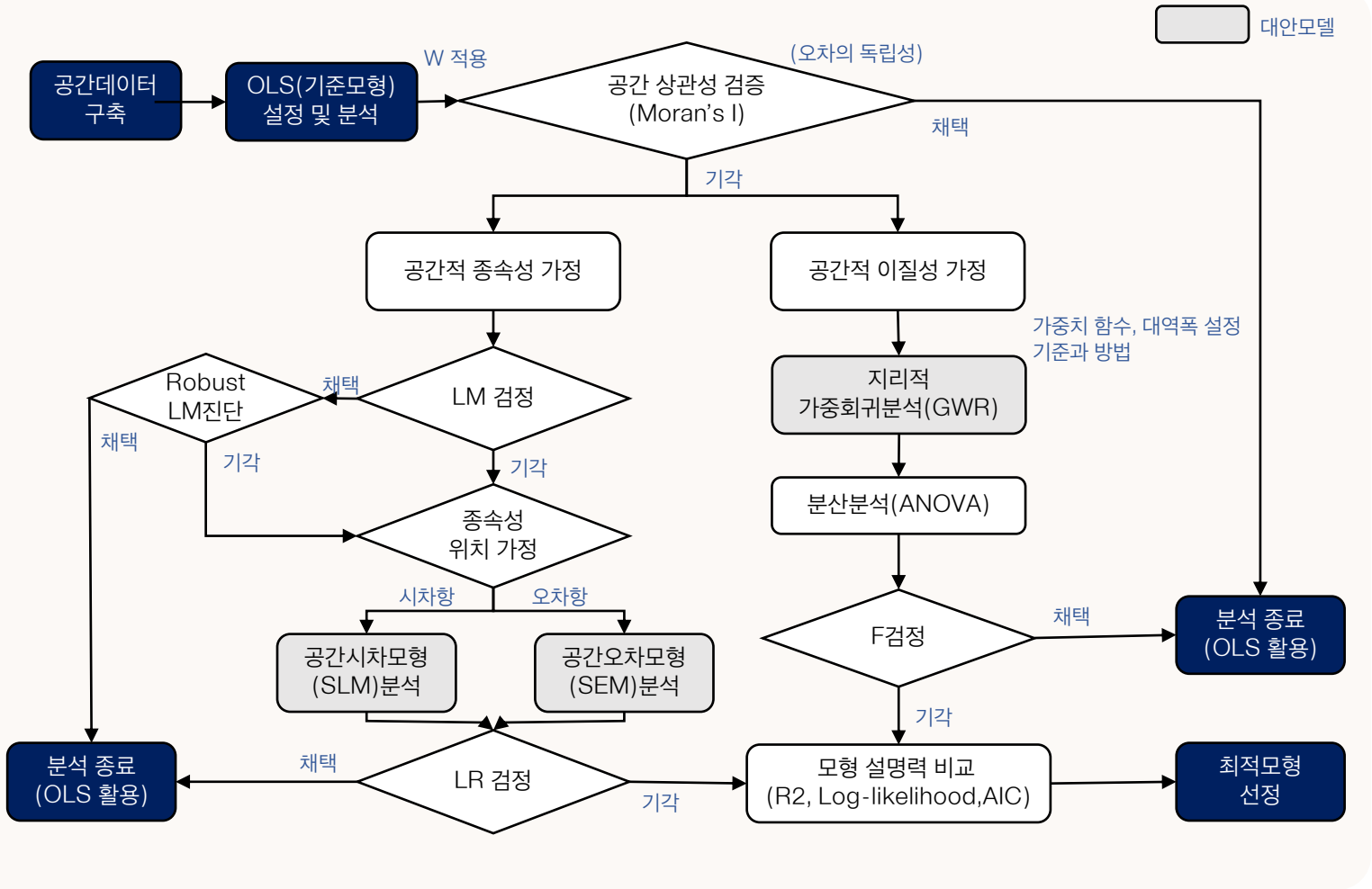
공간상관성 진단 – 국지적 자기상관성 (LISA test)

: 이때 Moran's I는 전역적인 값이며, 보다 국지적으로 어떤 지역들이 서로 유사한 값들을 가지며 클러스터하고 있는지 분석하기 위해 국지적 자기상관성 LISA 분석을 실시



공간 자기 상관성
있음을 확인
→ 공간계량모델 적용

공간계량모델 검증 다이어그램



공간시차모델 (SLM)

: 시차항의 공간적 종속성을 통제하는 공간시차모형

$$Y = \rho W_1 Y + X\beta + \epsilon$$

W_1 : 설명변수로 활용 | ρ: W_1 의 계수

공간오차모델 (SEM)

: 오차항의 공간적 종속성을 통제하는 공간오차모형
공간시차모형과 달리 공간오차모형은 공간시차항이 설명변수항이 아닌 오차항 부분에 포함되어 있다.

$$Y = X\beta + u \ (W_1 = 0)$$

라그랑지 승수 검정 (Lagrange Multiplier, LM 검정)

LM검정 개념

: OLS 분석 이후 LM검정을 실시한다. LM검정 둘다 유의한 결과면 Robust LM검정 실시하고 이를 통해 OLS 모형의 문제점을 진단, 적합한 대안모형을 설정하여 분석을 수행하게 된다.

독립변수간 다중공선성	모형 설명력, 안정성
오차의 정규성	설명 변수의 효율성
잔차의 공간적 자기상관성, 공간적 종속성의 위치 확인	

LM검정 결과 및 해석

적합한 대안모형 선택	LM(lag) 검정	유의	공간시차 모형이 회귀모형에 비해 더 적합
	LM(error) 검정	유의	공간오차 모형이 회귀모형에 비해 더 적합
	Robust LM(lag) 검정	유의	공간시차 모형이 회귀모형에 비해 더 적합
	Robust LM(error) 검정	유의	공간오차 모형이 회귀모형에 비해 더 적합
	LM(SARMA) 검정	유의	공간오차항과 공간시차항의 동시 적용시 유의

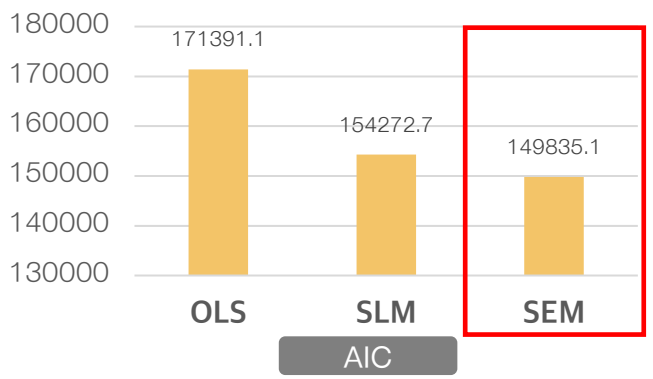
모든 검정에서 유의 값 나왔으므로 → **모델 전부 활용 후 R-square 및 AIC 값 비교**

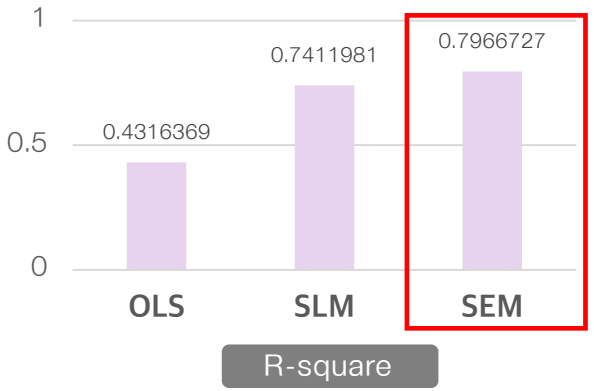
SLM(공간시차모델) vs SEM(공간오차모델) 비교분석

	Rho(p)	p-value
SLM	0.73032	2.22E-16

	Lambda(λ)	p-value(lambda)
SEM	0.86713	2.22E-16

" 각 공간시차항(ρ)과 공간오차항(λ)을 통해 교통사고에 대한 **인접지역의 영향력**을 알 수 있다 "





공간오차모델이 가장 적합한 모델임을 알 수 있다

지리적가중회귀모델 (GWR, Geographically Weighted Regression)

지리적가중 회귀모델 개념

: 지리가중 회귀모형의 원리는 회귀점으로부터 인접한 지역일수록 더 영향을 받아 상대적으로 큰 가중치를 주는 원리로 공간적 이질성 반영, 이분산성을 해결하기 위해 개별 지역마다의 회귀계수를 추정하는 모델이다.

$$Y_i = \beta_{0i} + \sum_{k=1}^{m} \beta k_0 X_{ki} + \varepsilon_i$$

지역적 영향력을 파악할 수 있다는 장점

최적의 대역폭 (bandwidth) 결정

회귀계수를 추정함에 있어 공간가중행렬을 정의하기 위해 최적의 대역폭 결정이 우선적으로 실시되어야 한다. 교통사고건수 실제값과 예측값과의 차이의 제곱 함수를 대역폭에 관한 함수로 표현하며 이 값을 최소화하는 최적의 대역폭을 찾는다.

AIC

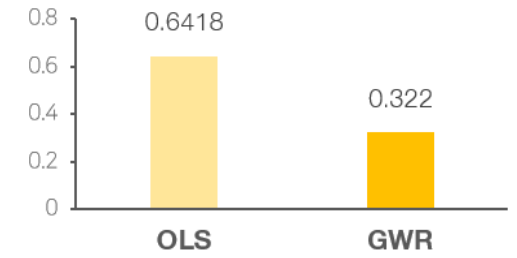
CV

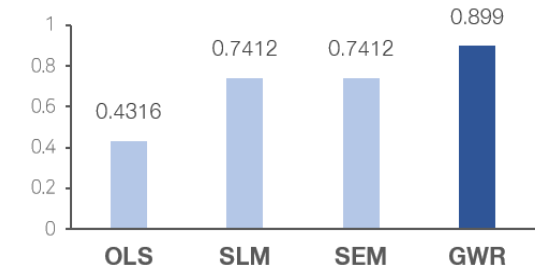
Cross-Validation 으로 찾은

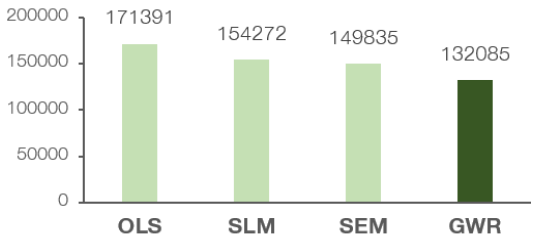
bw = 0.005585065904786666

* 대역폭이 커질 수록 동일 거리 가중치는 1에 가까움

모델 성능 비교







잔차분석 결과(Moran I)
기준모형과 비교해 GWR모델 개선되었음 확인

R-squared GWR이 제일 높다
→ GWR이 모델설명력 높음

AIC = - 2 LogLikelihood + 2p

AIC 가 GWR이 제일 낮다
→ GWR이 모델적합도 높음

지리적가중회귀모델 채택

결과도출

PSI 기준으로 위험지역 100개소 추출

: 위험도 산출 (PSI 지수 개념 활용)

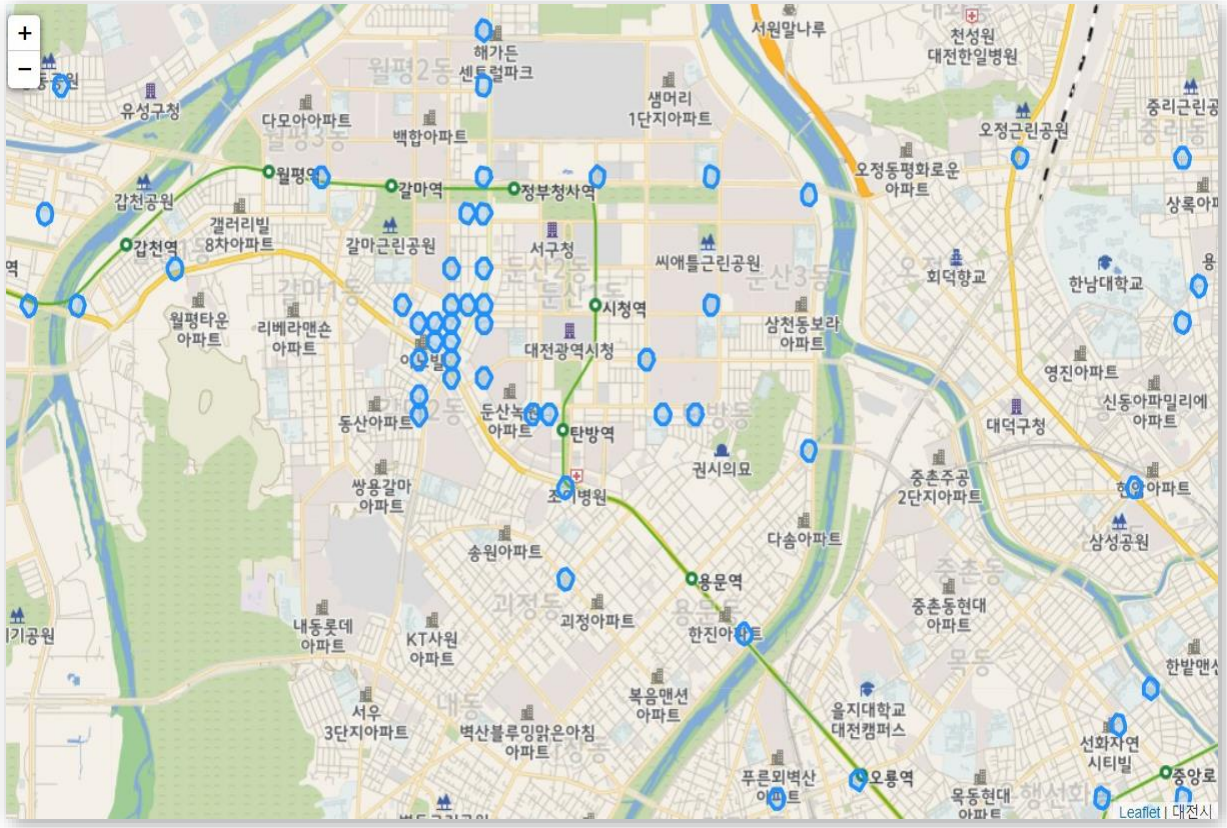
$$T = \frac{PSI \times Y}{max(Y)}$$

T : 교통사고 위험지역 우선순위 지수
PSI = 실제 사고건수 - 예측된 사고건수
Y : 실제 사고 발생횟수
max(Y) : 교통사고 최대 발생 횟수

위험도 점수는 PSI(Potential for Safety Improvement, 잠재적 교통안전개선지수) 개념을 활용

→ PSI 값이 높다는 말은 위험지수가 높아 개선할 경우 큰 효과를 볼 수 있는 지역이란 의미이다. 만일 PSI 값이 동일한 격자들 중 실제로 사고건수가 더 높은 지역에 가중치를 부여해 최종 위험도 점수를 산정

통계모델 위험지역 100소 샘플 도출



서론 01

/ 분석배경
/ 분석주제

02 전처리

/ 데이터분류
/ 데이터통합
/ 데이터정리

모델링 03

/ 인공지능 모델
/ 통계학적 분석

04 분석결과

/ 모델 비교
/ 시각화
/ 분석의의

모델간 성능비교

최적 인공지능 모델 vs 최적 통계 모델

인공지능 모델	
Best Model : XGBoost	
R^2	0.7868
Adjusted R^2	0.7865
MAE	1.0240
MSE	8.0154
RMSE	2.8311

통계 모델	
Best Model : GWR	
R^2	0.8989
AIC	132085.788
잔차의 공간자기상관성	0.3220634

R-squared 비교 시 통계모델인
GWR(공간가중회귀모델)이 설명력이 높아
최종모델로 선택되었다!

결정계수 R-squared의 의미

: 결정계수는 회귀모델에서 독립변수가 종속변수를 얼마만큼 설명해 주는지를 가지키는 지표로 모델의 설명력 이라고 불린다.

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

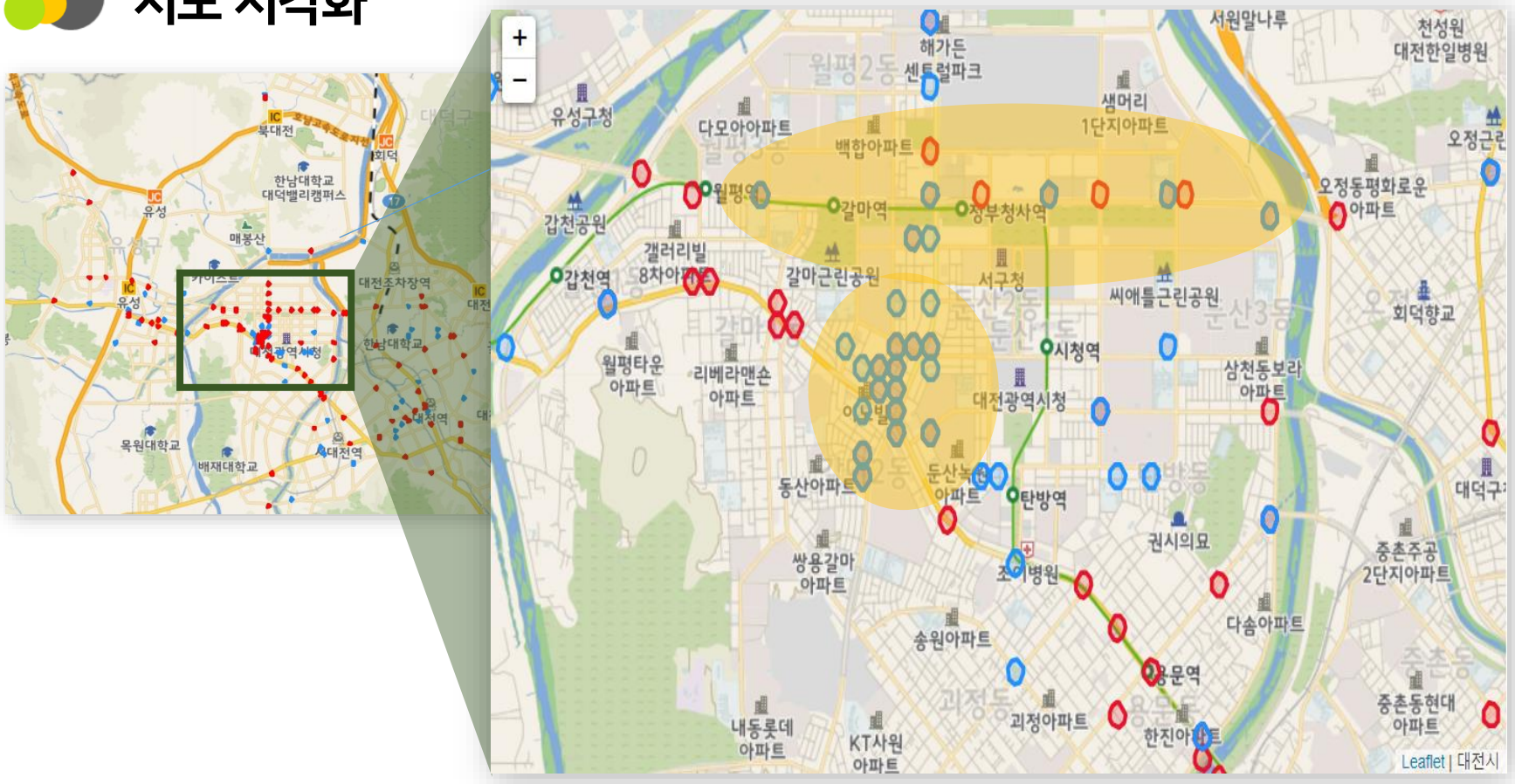
결정계수가 높을수록 종속변수의 독립변수에 대한 설명력이 높다는 의미

SST : 관측값에서 관측값의 평균을 뺀 결과의 총합

SSE : 추정값에서 관측값의 평균을 뺀 결과의 총합

04 결과분석_시각화

지도 시각화



- 통계모델GWR(인공지능모델과 겹치는 부분도 파란색으로 표시)
- ECLO가중치 적용된 인공지능 모델

인명피해가 크고
사고가 자주 발생하는 지역 도출

갈마 네거리 부근

터널직후 내리막길이 공존하는
지리적특성

한밭대교 부근

한밭대교에서 정부 청사쪽 도로 역시
인명사고 빈도가 높음

위 두 지역은 심층분석이 필요

EX) 사고유형과 법규위반, 연령대, 구체적인 지리적 특성, 사망자와 부상자 비율 등

분석의 의의와 한계

의의

종합적 분석

하나의 모델만 사용한 데이터분석에서 더 나아가
통계 모델 6가지 와 인공지능 모델 6가지(앙상블 포함)를
모두 활용하여 **다각적인 분석**을 함

선행연구(공간 계량 모델 3가지)를 토대로
공간데이터에 적합한 분석을 적용함

Depth 분석

인공지능 모델의 경우 단순 건수도출이 아닌
ECLO가중치를 활용하여 교통사고의 경중고려

버퍼 300M 가중치를 통해 사고범위 중복방지

PSI지수를 활용하여 실제로 사고건수가 더 높은 지역에
가중치를 주어 위험도 점수를 산정

한계

필요 데이터의 부재

킥보드 등의 개인이동수단 관련 사고 데이터가 적어
좀 더 폭넓은 분석을 하지 못함

교통 시설물의 설치시기를 알 수 있는
데이터가 있었다면 시간 경과에 따른
효과를 반영할 수 있었을 것.

시간상의 한계

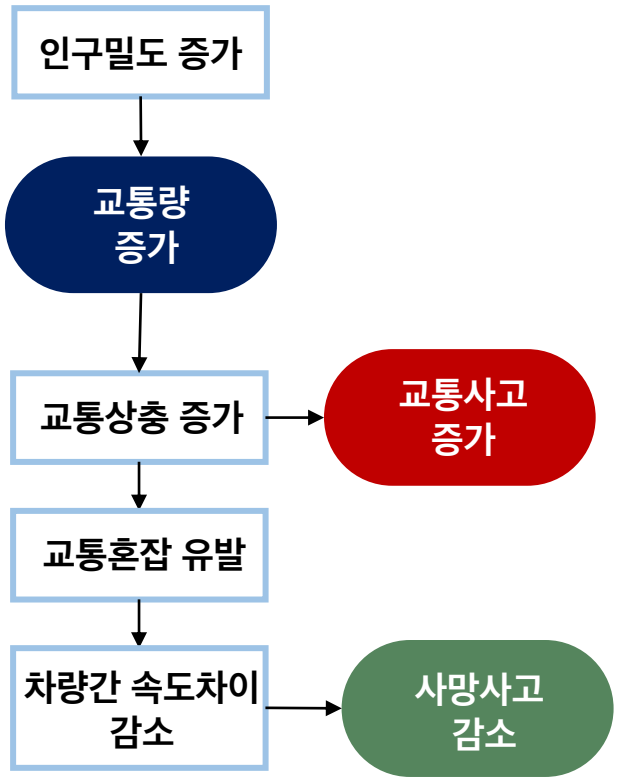
늘어나는 보행자 사고를 가중한 세밀한 분석을 하지 못함

갈마네거리, 한밭대교 등 인명피해가
큰 지역의 심층분석 부재

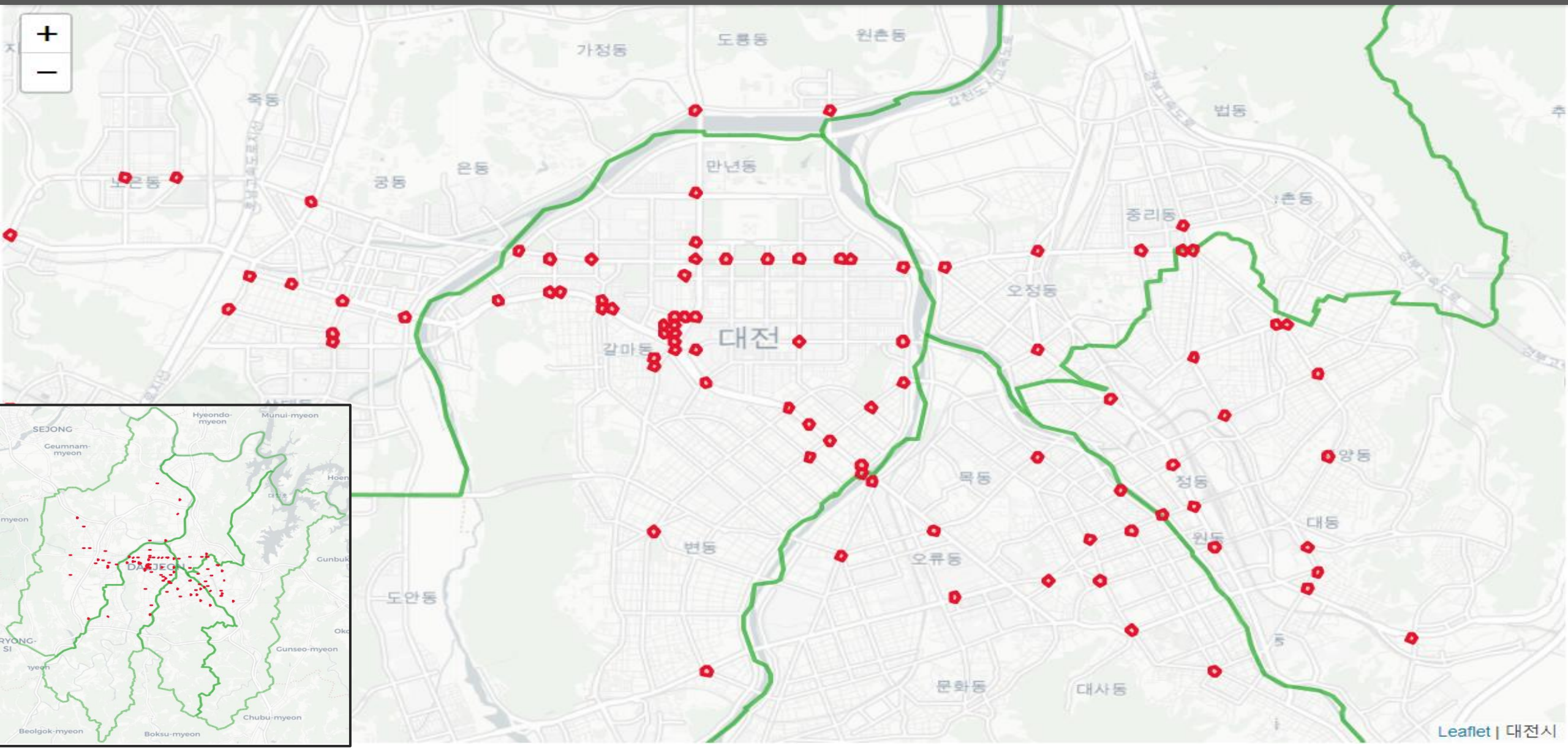
교통사고 분석결과 및 정책적 시사점

모델 분석결과	정책적 시사점
교통사고의 특성을 이해하기 위해서 인구밀도,교통량,교통사고 간 관계 이해가 필요함	교통사고가 이 변수들과 상호영향을 미치기 때문에 교통모형 분석시 다양한 변수들에 대한 충분한 고려가 필요함
제한속도가 30km 이하인 도로연장비율은 교통사고 감소영향	도로의 제한속도를 제한하는 정책이 교통사고에 안전한 도시환경을 조성하는데 기여할 수 있음을 제시하는 토대마련
교차로 지점에서 차대차 사고발생 위험 높음(차대차사고 4배 더 높음)	교차로 지점에서의 차대차 사고가 많이 일어남으로 차량사고 방지시 교차로를 우선적으로 고려해야함
CCTV 유무의 영향력	CCTV설치 유무가 사고발생에 큰 영향을 미치는 것을 알 수 있음. 사고위험 지역 CCTV설치 대수 증가를 제안
교통사고가 나타나게 된 공간적 이질성에 대한 검토 필요	교통사고건수는 동일하다 하더라도 잠재적 위험도는 주변환경에 따라 달라지기 때문. 이를 통해 안전개선 효과 및 예산 투자의 효율성 제고

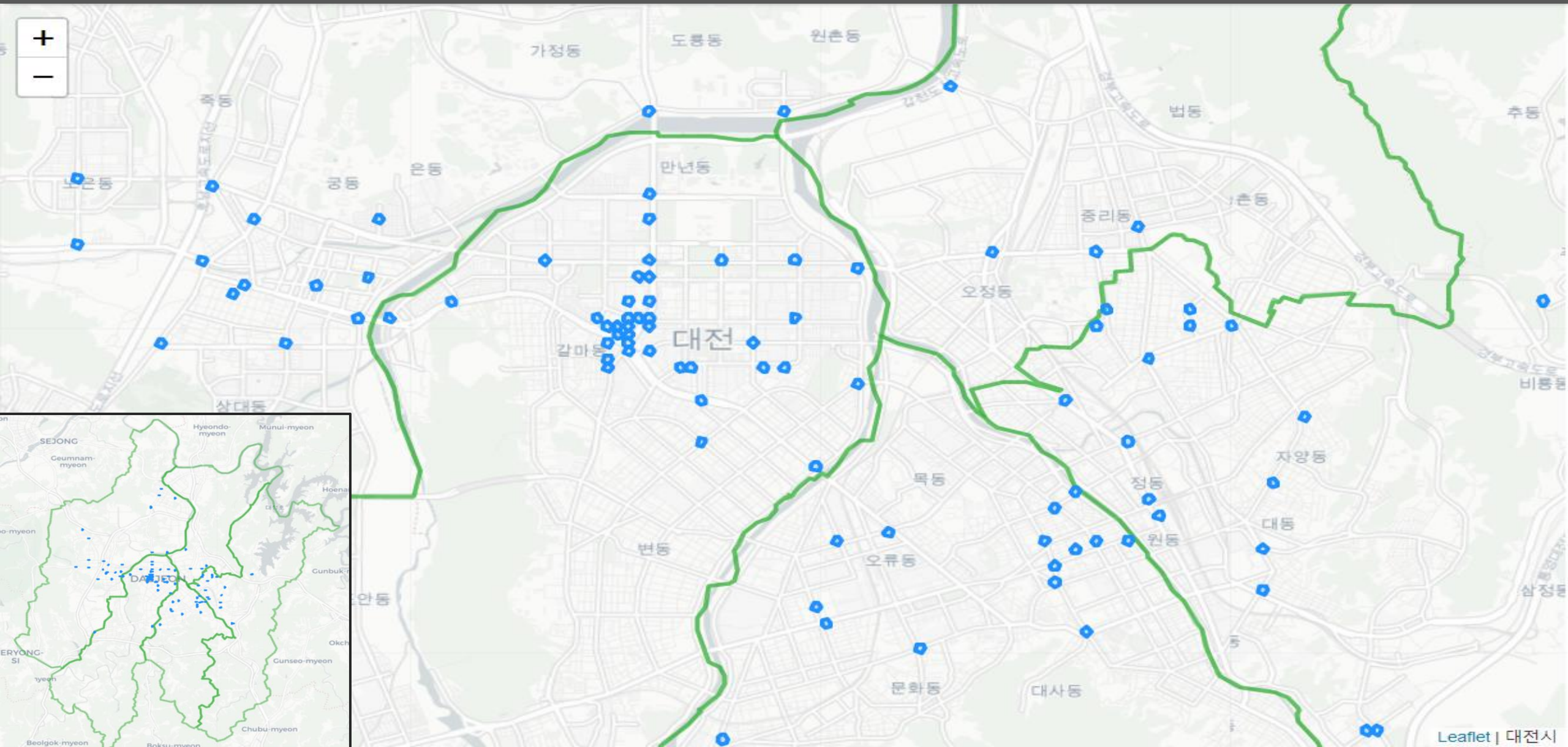
인구밀도,교통량,교통사고 관계



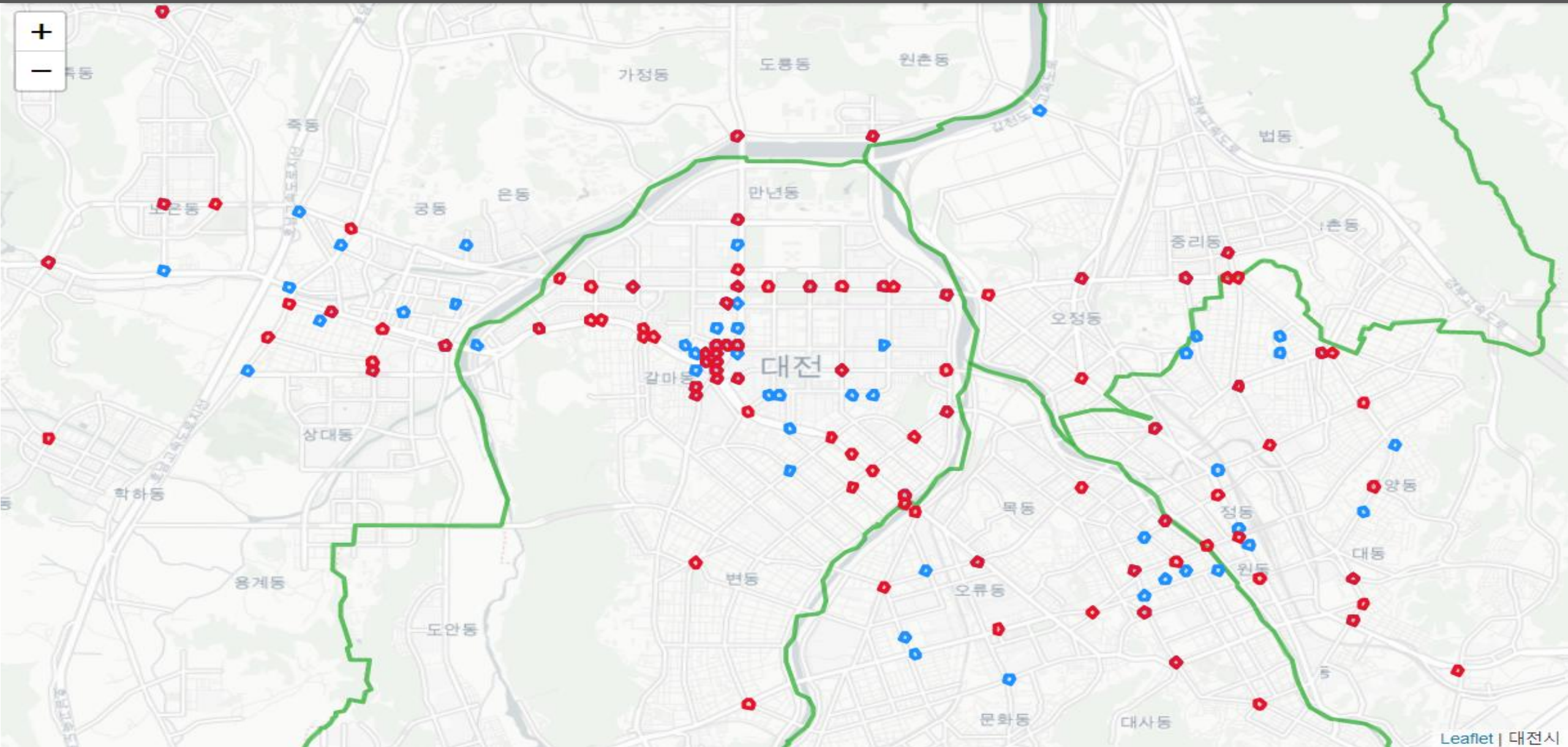
부록_ 인공지능 모델 100개소



부록_통계 모델 100개소



부록_ 인공지능,통계 100개소 중복



부록

부록_1번 데이터 묶음정리

	10대미만	10대	20대	30대	40대	50대	60대	70대	80대	합계
차대차 - 측면충돌	11	194	1206	1517	1551	1661	1046	347	41	1411
차대차 - 기타	11	141	840	1105	1153	1199	662	185	37	992
차대차 - 추돌	0	38	643	912	1061	1013	593	124	16	681
차대사람 - 횡단중	126	245	283	199	248	316	329	267	86	654
차대사람 - 기타	66	134	328	181	213	302	257	172	96	528
차대사람 - 길가장자리구역통행중	6	38	109	36	33	40	27	28	10	153
차대차 - 정면충돌	2	16	96	112	132	126	91	16	2	114
차대사람 - 차도통행중	15	17	72	31	42	49	36	47	16	104
차대차 - 후진중충돌	0	6	95	97	72	57	20	3	3	101
차대사람 - 보도통행중	11	20	27	21	21	38	25	13	11	58
차량단독 - 전도전복	0	0	0	0	0	25	0	0	0	0
차량단독 - 전도전복 - 전도	0	0	0	0	0	13	0	0	0	0
차량단독 - 전도전복 - 전복	0	0	0	0	0	3	0	0	0	0

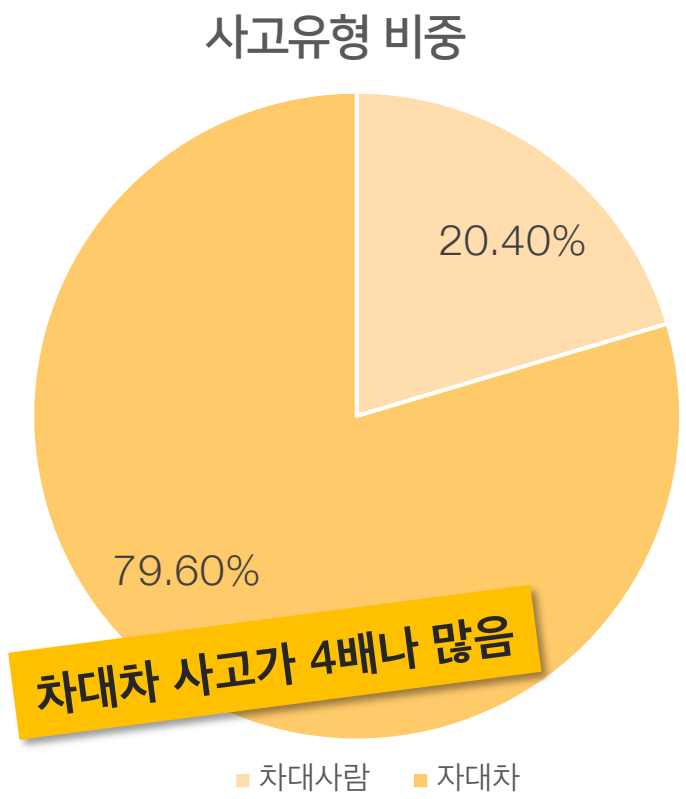
- 피해운전자 기준 사망자수,중상자,경상자,부상신고자 합
- 90대는 80대, 미분류는 가장많은 50대로 재분류

	10대미만	10대	20대	30대	40대	50대	60대	70대	80대	합계
승용	1	63	1987	2930	2905	2866	1778	363	28	12921
보행자	224	454	819	468	557	745	674	527	219	4687
이륜	0	113	520	392	312	213	105	59	13	1727
자전거	21	178	133	63	107	197	199	168	44	1110
화물	0	1	71	161	253	330	172	40	3	1031
승합	0	1	21	84	258	328	106	17	0	815
원동기	0	35	109	74	89	66	38	20	6	437
개인형이동수단(PM)	1	3	20	14	17	10	2	3	2	72
특수	0	0	12	24	16	8	1	0	0	61
미분류	0	0	0	0	0	41	0	0	0	41
건설기계	0	0	1	1	6	18	7	1	0	34
기타불명	1	0	2	0	3	18	2	2	2	30
사륜오토바이(ATV)	0	1	4	0	3	2	0	0	1	11
농기계	0	0	0	0	0	0	2	2	0	4

- 연령대별 피해운전자 차량종류

부록

부록_1번 데이터 묶음정리



구	10대미만	10대	20대	30대	40대	50대	60대	70대	80대	합계
서구	95	311	1371	1448	1339	1429	879	280	77	7229
유성구	55	187	922	1199	1197	1224	650	214	46	5694
동구	23	138	553	564	669	766	595	272	66	3646
중구	47	128	488	523	699	747	571	276	81	3560
대덕구	28	85	365	477	622	676	391	160	48	2852

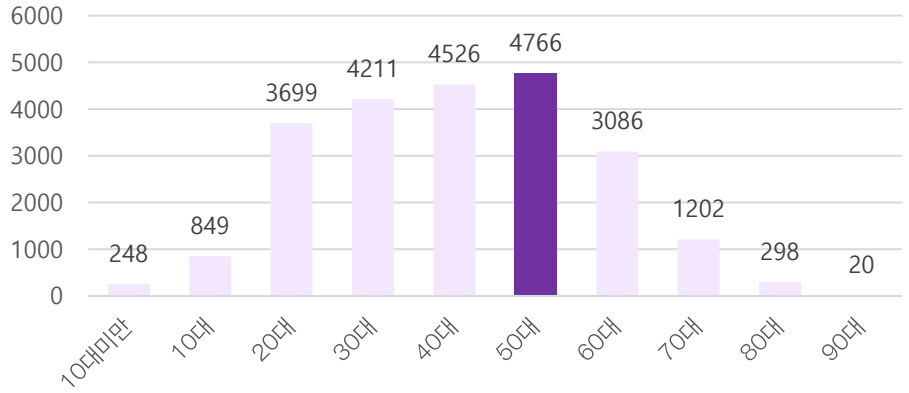
연령대별 사고유형 그래프는 코드 부록 참조

- 구별 연령대사고
- 동별 데이터는 Ipynb파일에 구 - 동 데이터 정리파일 존재

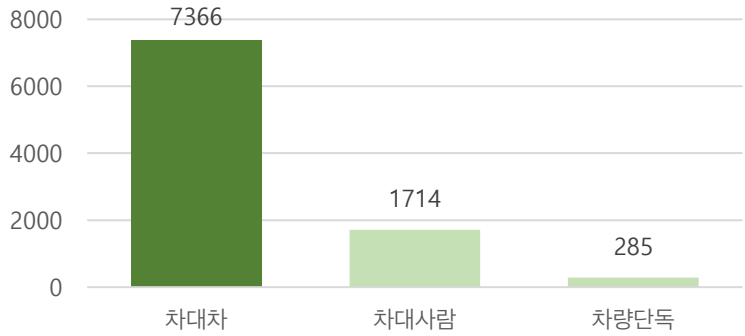
부록

부록_1번 데이터 묶음정리

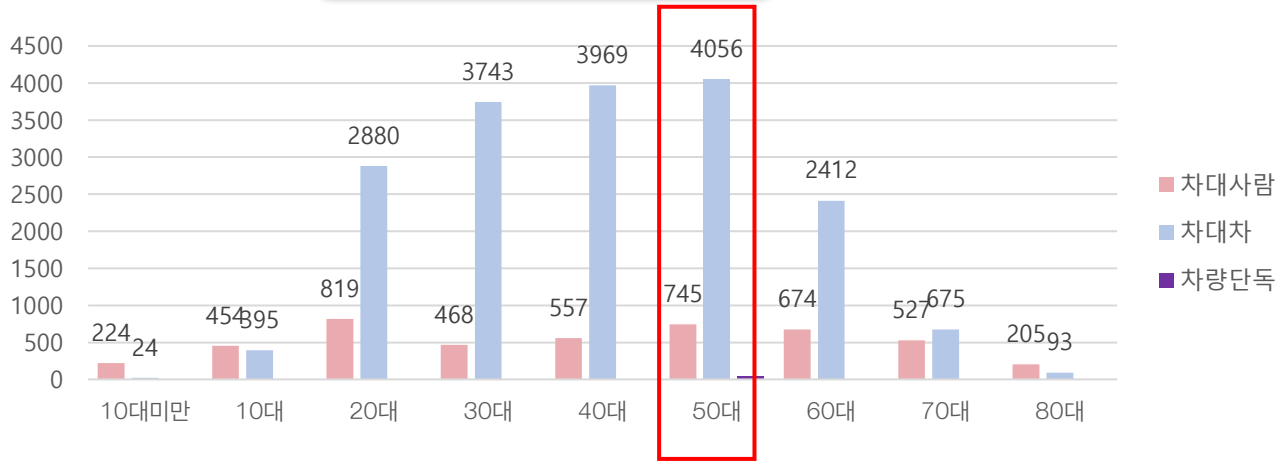
연령대별 사고횟수



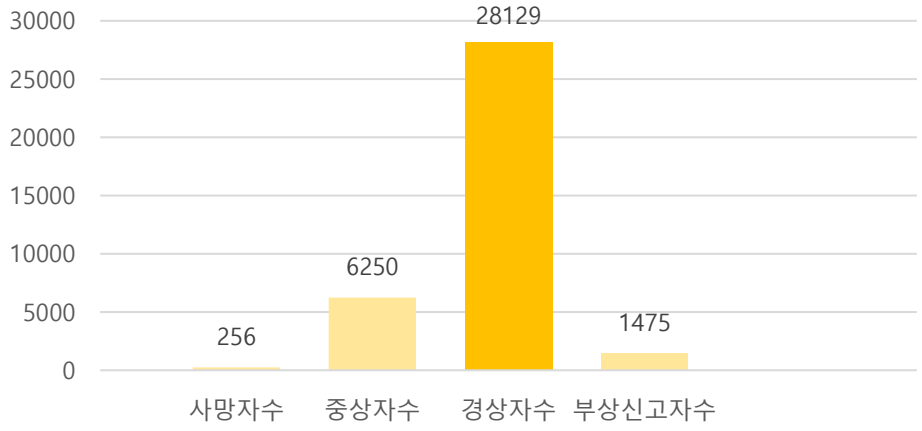
사고 유형별 교통사고 횟수



연령대별 사고유형



부상 심각도별 사고횟수



참고문헌

참고자료

(1) 도로교통공단/서울시 교통안전은 우리가 해결한다/2013
(2) 행정자치부/공공 빅데이터 분석사업 성과 공유대회/2016
(3) 도로교통공단/교통사고 빅데이터를 활용한 위험도로예보시스템/2016
(4) 조준한,박웅원,김민우/교통 기술과 정책, 제11권 제2호/기상정보와 연계한 교통안전예보지수 개발 및 활용방안/2014
(5) 박나영,박병호/산업과학기술 논문집 제32권 제1호/교통사고비용과 EPDO에 근거한 사고밀도 청주사례 분석/2018
(6) 백태헌, 손슬기, 박병호/대한교통학회지 34(6)/공간적 특성을 고려한 어린이 교통사고 모형 개발/2016
(7) 경향신문/ 교차로 조명 밝게 하면 교통사고 줄어들까?/2020.12.28
(8) 뉴스티앤티/ [2019 국감] 세종·대전, 교통사고 증가율...'나란히 1·2위/2019.10.08
(9) 연합뉴스/ 대전 교통사고 사망자 감소율 특·광역시 중 꼴찌…광주의 절반/2021.02.28
(10) 충청투데이/ 대전시, 교통체증 해소 등 시민체감 교통분야 사업에 538억원 투입/2021.02.16
(11) 오마이뉴스/대전시, 자동차 1만대 당 교통사고 발생 전국 1위/2020.10.2
(12) 박세희,김민수,백장선/지리시간가중 회귀모형을 이용한 주택가 | 격 영향요인 분석/2017
(13) 김지우,이건학/대한지리학회지 제54권 제5호/지리가중회귀 모델을 이용한 학교급별 학업성취도 영향 요인 분석/2019
(14) 이삼청춘/오산시 COPASS공모전 최우수 수상작/통계적 모델 부분 참조/2021
(15) 이경아/공간계량모형을 활용한 교통사고 유형별 발생 특성 분석/2016.08
(16) 정재풍,최종후/Journal of The Korean Data Analysis Society/교통사고건수에 대한 포아송회귀와 음이항 회귀모형 적합/2014
(17) 서민송,유환희/Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography/시설물 유형에 따른 화재 발생의 공간계량분석
(18) 성현곤/충북대학교 도시공학과/06.R정형데이터 분석 02 분류와 예측
(19) 전해정/공간계량분석기법과 GIS를 이용한 주택가격모형 비교에 관한 연구/2016.02.15
(20) 박진옥,최일수,나명환/공간 자료를 이용한 대기오염이 순환기계 건강에 미치는 영향 분석/2016.11.08
(21)최상열/고급통계분석론 제7장 상관분석과 공간적 자기상관, 13장 공간계량모델/p634~657
(22) 국토연구원/교통사고에 안전한 국토구현/p119~128

감사합니다!

(TEAM. 호반우)