Traffic Accident Analysis Using Big Data

SeungYup Baek, JangHun Kwon, JuYeon Kim, AYoung Jo, HyunChul JungJungeun Kim

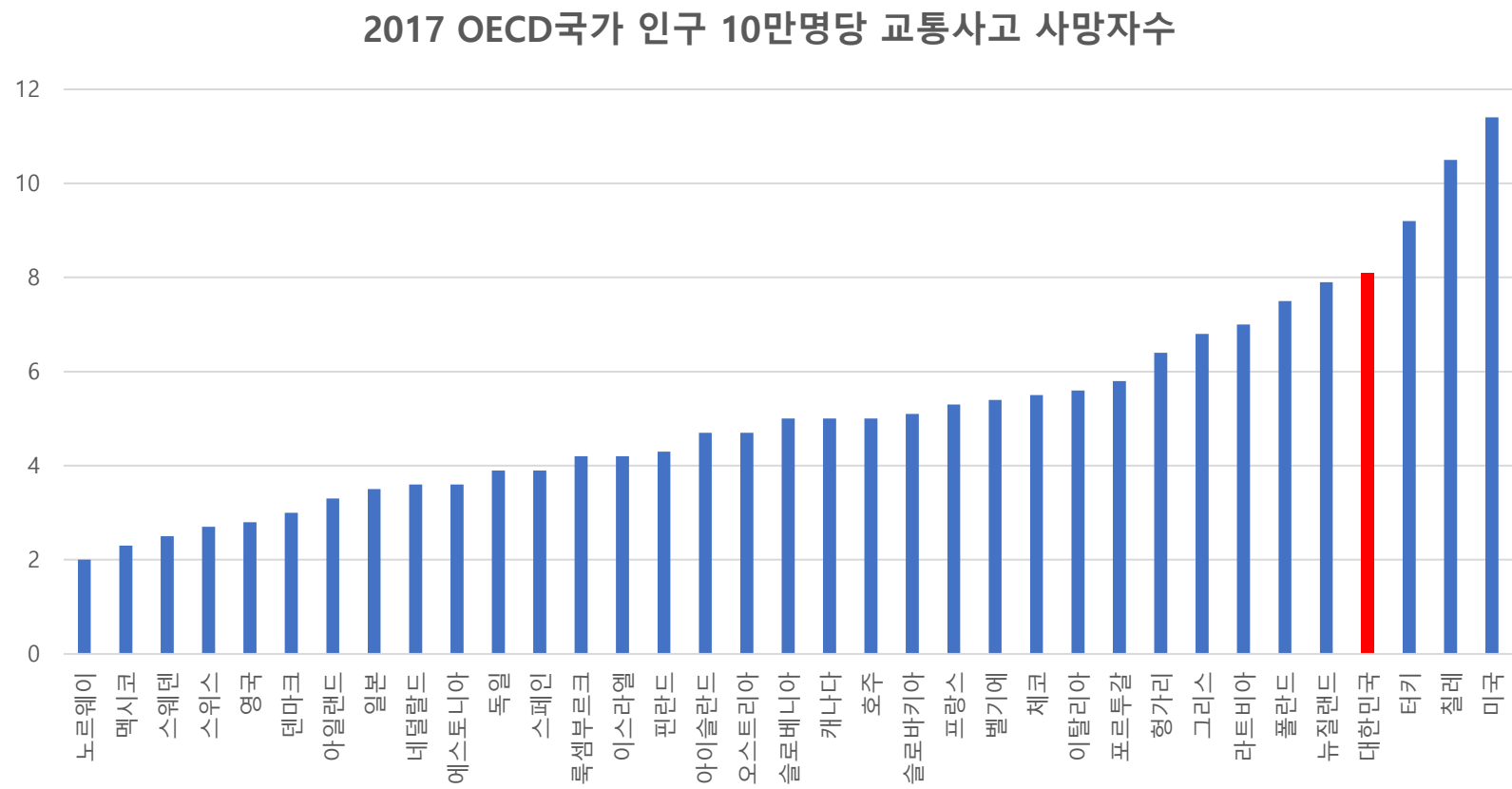
Department of Computer Science and Engineering, Kongju National University

CONTENTS

서론

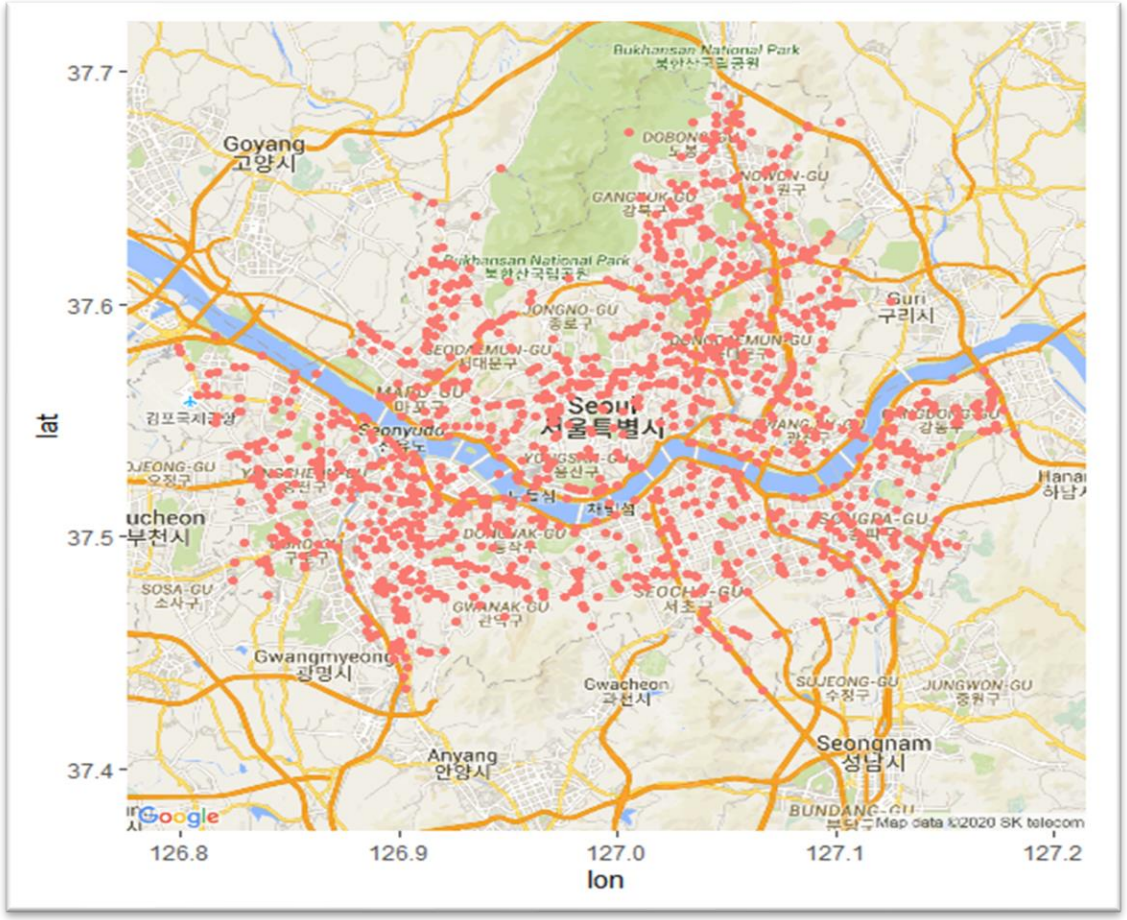
- 교통사고는 막대한 인적, 물적 비용을 초래하는 사회적 문제이다
- 제103회 국전현안점검조정회의에서「2020 교통사고 사망자 줄이기 대책」을 심의 확정 하였다.

서론



- OECD 국가 통계를 보면 우리나라의 인구 10만명당 교통사고 사망자수는 상위권에 위치한 것을 알 수 있다

데이터셋 : k-means



서울시 교통사고 가시화

서울시 교통사고 중 사망 교통사고를 지도위에 표시

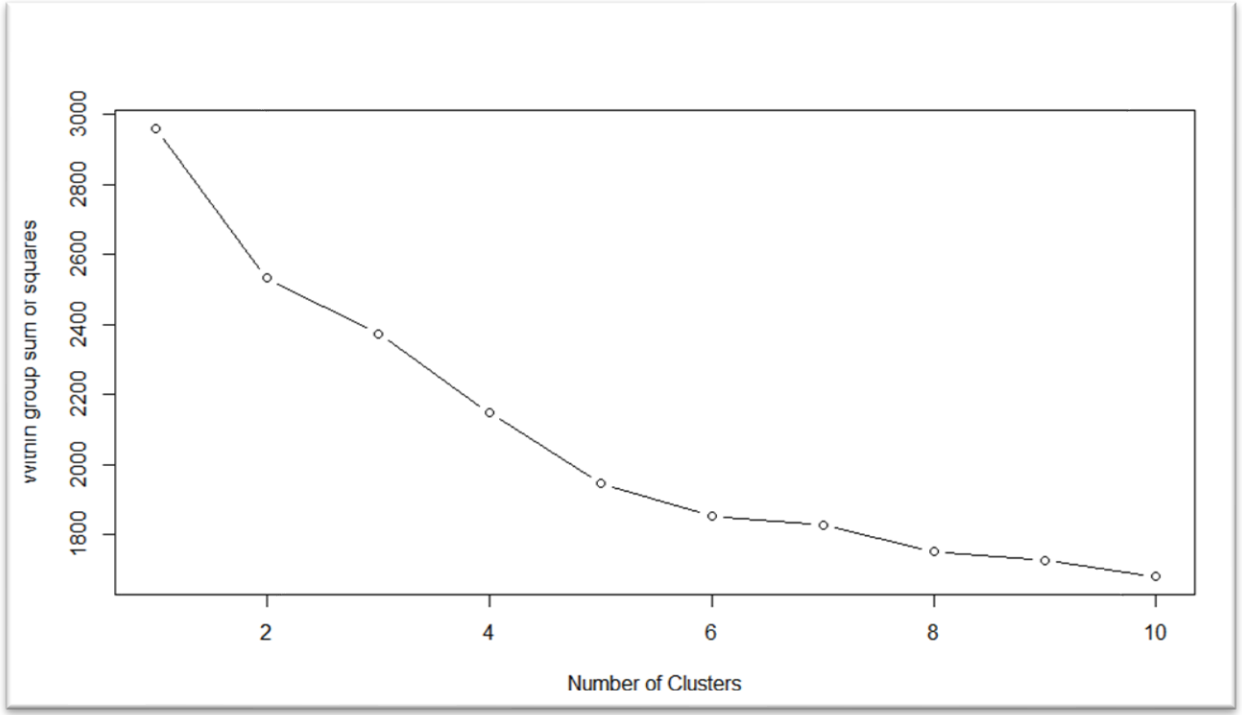
- 공공 데이터포털에서 제공하는 2015년부터 2018년 까지 전국 사망 교통사고 자료 중 서울시에서만 발생한 교통사고 데이터를 이용
- 문헌 조사를 통해 총 26개의 변수 중 요일, 사고유형, 법규위반, 도로형태 4가지 변수를 유의미한 영향을 미친다고 가정

데이터셋 : k-means

데이터셋 전처리

- 분석에 사용된 데이터가 모두 범주형이여서 k-means에 사용하기 적합하지 않았다
- 전처리를 통해 요일 변수는 평일과 주말을 구분하기 위해 평일은 0 주말은 1로 값을 변경
- 전처리를 통해 사고유형, 법규위반, 도로형태 변수는 더미 코딩을 통해 더미변수를 생성

요인분석 : k-means



엘보우 그래프

- k-means 알고리즘을 통해 클러스터를 도출하고 각 클러스터별 특성 분석을 통해 사망 교통사고의 주요 요인을 도출
- k-means 클러스터링에 가장 적합한 클러스터 수 k를 도출하기 위해 엘보우 기법을 사용
- 좌측 엘보우 그래프에서 가장 급격한 값의 변화를 나타내는 클러스터 5를 최적개수 k로 결정

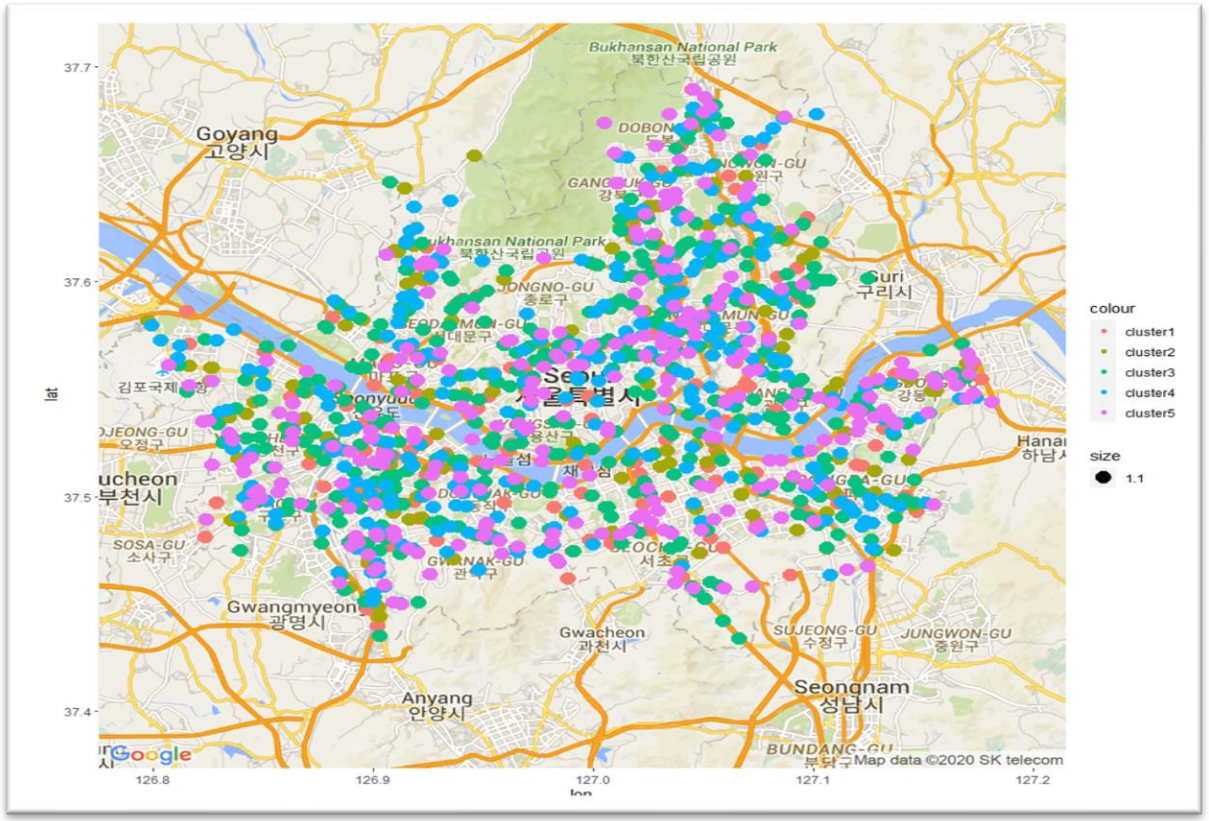
요인분석 : k-means

	클러스터 1	클러스터 2	클러스터 3	클러스터 4	클러스터 5
데이터 개수	181개	176개	406개	299개	286개
요일	평일(70%)	평일(73%)	평일(73%)	평일(78%)	평일(69%)
사고유형	횡단중(30%)	측면충돌(31%)	횡단중(42%)	횡단중(57%)	기타(100%)
법규위반	중앙선 침범(30%)	신호위반(84%)	안전운전 의무 불이행(100%)	안전운전 의무 불이행(84%)	안전운전 의무 불이행(88%)
도로형태	기타단일로(76%)	교차로내(81%)	기타단일로(100%)	교차로내(34%)	기타단일로(59%)

클러스터 별 변수의 최빈값

- 분석결과 요일 변수는 평일, 사고유형 변수는 횡단 중, 법규 위반 변수는 안전운전 의무 불이행, 도로형태 변수는 기타 단일로가 가장 높은 빈도를 보였다.
- 교통량이 많은 주말이 평일보다 교통사고가 더 많이 일어날 것이라는 기존의 인식과 달리 평일에 사망 교통사고 발생률이 더 높게 관측된 점은 특이할 만한 사항으로 판단됨

요인분석 : k-means



클러스터 별 지도 시각화

- 사망 교통사고를 클러스터별로 컬러코딩 하여 지도에 가시화
- 특정 장소에서 동일하거나 유사한 원인들로 인해 교통사고가 자주 발생하는 것이 아니라, 각기 다른 장소에서 다양한 원인들로 인해 교통사고가 발생한다는 것을 알 수 있다

결론 : k-means

- 교통량이 많은 주말이 평일보다 교통사고가 더 많이 일어날 것이라는 기존의 인식과 달리 평일에 사망 교통사고 발생률이 더 높게 관측된 점은 특이할 만한 사항으로 판단됨
- 교통사고의 주요 요인과 유형을 분석한 결과 평일에 기타 단일로에서 안전운전 의무를 불이행하는 횡단 중 교통사고가 가장 빈번하게 발생한다는 것을 확인하였다.

사망사고의 가장 빈번한 법규위반의 내용이 운전자의 안전의무 불이행이라는 점은 시사하는 바가 크며 운전자의 안전의무를 더욱 강화하는 노력이 필요하다고 판단된다.

데이터셋 : EM 알고리즘

발생일	발생시간	요일	발생지_시	사고내용	법정동명	사망자수	중상자수	경상자수	부상신고자	사고유형_	사고유형_	사고유형	가해자법규	노면상태_	노면상태	기상상태	도로형태_	도로형태	가해자차종	가해성별	가해자연령	가해자신체	피해자차종	피해자성별	피해자연령	피해자신고
20100101	00시	금	마포구	경상	서교동	0	0	1	0	차대차	기타	기타	기타	포장	젖음/습기	맑음	단일로	기타단일로	승용차	남	54세	상해없음	이륜차	남	19세	경상
20100101	00시	금	동작구	중상	대방동	0	1	3	0	차대차	추돌	진행중 추	안전거리	포장	서리/결빙	맑음	단일로	기타단일로	승합차	남	57세	상해없음	승용차	남	39세	중상
20100101	00시	금	관악구	경상	신림동	0	0	2	0	차대차	추돌	진행중 추	기타	포장	서리/결빙	흐림	단일로	기타단일로	승용차	남	51세	경상	승용차	남	58세	경상
20100101	00시	금	강서구	경상	화곡동	0	0	1	0	차대차	기타	기타	안전거리	포장	건조	맑음	교차로	교차로부근	승용차	남	56세	상해없음	승용차	남	55세	경상
20100101	00시	금	구로구	경상	구로동	0	0	2	0	차대차	정면충돌	정면충돌	신호위반	포장	건조	맑음	교차로	교차로내	승용차	남	33세	상해없음	승용차	남	55세	경상
20100101	00시	금	서초구	중상	내곡동	0	1	0	0	차대차	추돌	진행중 추	안전거리	포장	건조	맑음	단일로	터널안	승용차	남	58세	상해없음	승용차	여	32세	중상
20100101	00시	금	양천구	중상	신월동	0	1	0	0	차대차	측면충돌	측면충돌	교차로 통	포장	건조	맑음	교차로	교차로내	승용차	남	26세	상해없음	이륜차	남	20세	중상
20100101	00시	금	도봉구	경상	쌍문동	0	0	1	0	차대차	측면충돌	측면충돌	안전거리	포장	건조	맑음	교차로	교차로부근	승용차	남	30세	상해없음	원동기장치	남	17세	경상
20100101	00시	금	은평구	사망	갈현동	1	0	0	0	차대사람	길가장자리	길가장자리	안전운전	포장	건조	맑음	단일로	기타단일로	화물차	남	30세	상해없음	보행자	남	27세	사망
20100101	01시	금	종로구	경상	세종로	0	0	1	0	차대차	추돌	주정차중	안전운전	포장	건조	맑음	단일로	기타단일로	승용차	남	34세	상해없음	승용차	남	29세	경상
20100101	01시	금	서대문구	중상	연희동	0	1	1	0	차대차	추돌	주정차중	안전거리	포장	건조	맑음	교차로	교차로부근	승용차	남	53세	상해없음	승용차	남	48세	경상
20100101	01시	금	강북구	경상	미아동	0	0	2	0	차대차	추돌	진행중 추	안전운전	포장	젖음/습기	흐림	단일로	기타단일로	승용차	남	27세	상해없음	승용차	남	47세	경상
20100101	01시	금	동대문구	경상	장안동	0	0	2	0	차대차	추돌	진행중 추	안전거리	포장	서리/결빙	흐림	단일로	횡단보도부	승용차	남	50세	상해없음	승용차	남	44세	경상
20100101	01시	금	마포구	경상	공덕동	0	0	1	3	차대차	측면충돌	측면충돌	안전거리	포장	건조	맑음	단일로	기타단일로	승용차	남	50세	상해없음	승용차	남	40세	경상
20100101	01시	금	영등포구	중상	문래동6가	0	2	1	0	차대차	측면충돌	측면충돌	안전운전	포장	서리/결빙	흐림	교차로	교차로내	승용차	남	30세	상해없음	승용차	남	53세	경상
20100101	01시	금	금천구	경상	독산동	0	0	5	0	차대차	추돌	주정차중	안전운전	포장	젖음/습기	흐림	교차로	교차로내	승용차	남	33세	상해없음	승용차	남	52세	경상
20100101	01시	금	관악구	경상	봉천동	0	0	2	0	차대차	추돌	진행중 추	안전운전	포장	건조	맑음	단일로	기타단일로	승용차	남	66세	상해없음	승용차	남	36세	경상
20100101	01시	금	강동구	경상	천호동	0	0	7	1	차대차	추돌	진행중 추	안전운전	포장	서리/결빙	맑음	단일로	기타단일로	승용차	여	23세	상해없음	승용차	남	53세	부상신고
20100101	01시	금	서초구	경상	서초동	0	0	1	0	차대사람	길가장자리	길가장자리	구역통행	포장	건조	맑음	교차로	교차로부근	승용차	남	51세	상해없음	보행자	여	18세	경상
20100101	01시	금	노원구	중상	상계동	0	2	0	0	차대차	측면충돌	측면충돌	중앙선 침	포장	건조	맑음	교차로	교차로부근	승용차	남	52세	상해없음	이륜차	남	17세	중상
20100101	02시	금	중구	경상	의주로1가	0	0	1	0	차대차	추돌	주정차중	안전거리	포장	적설	눈	단일로	기타단일로	승용차	남	51세	상해없음	승용차	남	54세	상해없음
20100101	02시	금	강북구	경상	수유동	0	0	1	2	차대차	측면충돌	측면충돌	중앙선 침	포장	건조	맑음	기타/불명	기타/불명	승용차	남	44세	상해없음	승용차	남	47세	경상
20100101	02시	금	광진구	경상	자양동	0	0	1	0	차대차	측면충돌	측면충돌	교차로 통	포장	건조	맑음	교차로	교차로내	승용차	남	59세	상해없음	승용차	남	38세	경상
20100101	02시	금	마포구	중상	성산동	0	1	1	0	차대차	기타	기타	안전운전	포장	건조	맑음	단일로	기타단일로	승용차	남	20세	상해없음	승용차	남	33세	경상
20100101	02시	금	관악구	중상	신림동	0	1	0	0	차대차	추돌	주정차중	기타	포장	젖음/습기	맑음	교차로	교차로부근	승용차	남	40세	상해없음	이륜차	남	23세	중상
20100101	02시	금	관악구	경상	봉천동	0	0	3	0	차대차	추돌	진행중 추	안전거리	포장	젖음/습기	맑음	교차로	교차로부근	승용차	남	54세	상해없음	승용차	남	49세	경상
20100101	02시	금	성북구	경상	장위동	0	0	2	0	차대차	추돌	진행중 추	안전운전	포장	건조	맑음	단일로	기타단일로	승용차	남	64세	상해없음	승용차	남	55세	경상
20100101	02시	금	도봉구	경상	창동	0	0	2	0	차대차	측면충돌	측면충돌	신호위반	포장	건조	맑음	교차로	교차로부근	승용차	남	55세	상해없음	승용차	남	38세	상해없음
20100101	03시	금	종로구	경상	내자동	0	0	3	0	차대차	측면충돌	측면충돌	교차로 통	포장	건조	맑음	교차로	교차로내	승용차	남	54세	상해없음	승용차	남	49세	경상
20100101	03시	금	은평구	중상	녹번동	0	1	0	0	차대사람	횡단중	횡단중	보행자 보	포장	서리/결빙	맑음	단일로	기타단일로	승용차	남	41세	상해없음	보행자	남	30세	중상
20100101	03시	금	영등포구	경상	양평동4가	0	0	1	0	차대차	측면충돌	측면충돌	안전거리	포장	건조	맑음	단일로	교량위	승용차	남	31세	상해없음	승용차	남	46세	경상

- 공공데이터포털의 서울시 개별교통사고 정보 데이터를 사용함

데이터셋 : EM 알고리즘

발생시간	
주간	야간
1	2

요일						
월	화	수	목	금	토	일
1	2	3	4	5	6	7

사고내용			
경상	중상	사망	부상신고
1	2	3	4

사고유형_대분류			
차대차	차대사람	차량단독	철길건널목
1	2	3	4

- 데이터의 결측값이나 이상값이 포함된 경우는 데이터 전처리를 통해 제외하였다.
- 사고유형_대분류, 사고유형_중분류와 같이 의미상 중복인 변수들은 대표변수 하나만을 채택하였다.
- 발생시간 데이터에서 06시~18시는 주간 18시~06시는 야간으로 범주화 하였다.

데이터셋 : EM 알고리즘

가해자법규위반						
안전운전의무불이행	안전거리미확보	신호위반	교차로통행방법위반	중앙선침범	보행자보호의무위반	과속
1	2	3	4	5	6	7

노면상태						
건조	젖음습기	서리결빙	적설	침수	해빙	기타불명
1	2	3	4	5	6	0

기상상태					
맑음	흐림	비	안개	눈	기타불명
1	2	3	4	5	0

도로형태_대분류			
단일로	교차로	철길건널목	기타불명
1	2	3	0

- 데이터셋의 데이터 타입을 정수형 데이터 타입으로 변경하였다.
- 기타불명같이 여러 데이터에 속하는 항목은 최대한 같은 정수를 가지도록 변경하였다.

요인분석 : EM 알고리즘

방법론

데이터 정제

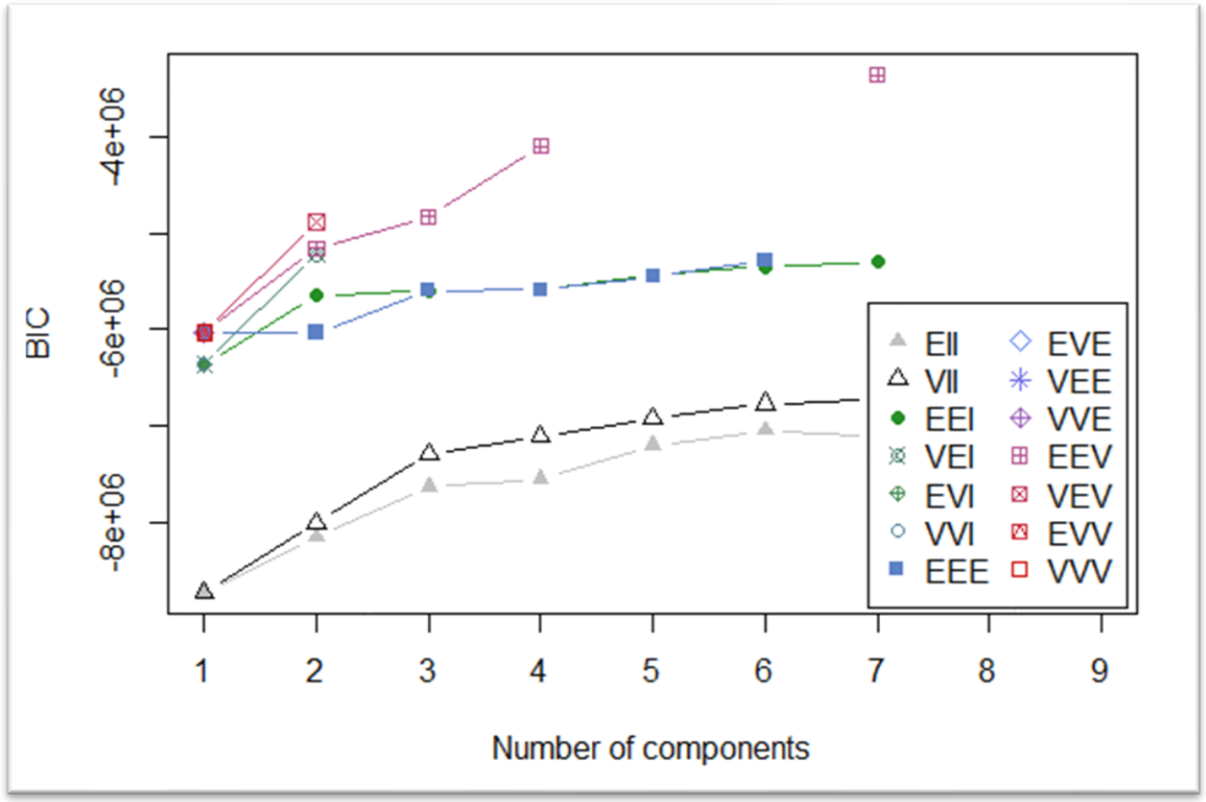
기댓값 최대화 알고리즘 (Expectation-Maximization Algorithm)을 이용하여 데이터를 클러스터링

주성분 분석을 활용하여 교통사고의 주요 요인 파악

- EM알고리즘을 이용하여 데이터를 클러스터링 하고 클러스터별 규칙을 찾는다.
- 최적의 클러스터 수를 결정하기위해 베이즈 정보기준을 활용하였다.
- 주성분 분석을 통해 변수간의 상관관계와 변수가 교통사고에 끼치는 영향의 정도를 분석

요인분석 : EM 알고리즘

EM알고리즘 결과



첫 글자는 군집의 크기 두 번째 글자는 군집의 형태 세 번째 글자는 방향을 의미함

(E는 equal V는 variable I는 Coordinate axes의 약자)

- BIC (Bayesian Information Criterion) 값을 확인한 결과 가장 높은 BIC 값으로 EEV 7값이 나타났다.
- EEV 7은 군집의 크기가 같고 형태가 같고 방향은 다양하다는 것을 의미한다.
- BIC를 통해 7개의 클러스터를 사용하기로 결정했다.

요인분석 : EM 알고리즘

EM알고리즘 결과

클러스터	1	2	3	4	5	6	7
발생시간	1.9764	1.9717	1.9677	1.9576	1.9384	1.9732	1.9521
요일	1.5074	3.4882	2.1115	5.5259	4.0386	5.0245	6.425
사고내용	1.4323	1.3519	1.5414	1.5356	1.9074	1.3041	1.441
사고유형_대분류	1.3402	1.2925	1.2888	1.2532	1.5719	1.2648	1.3292
가해자법규위반	1.0773	1.0957	3.9919	3.9336	1.1957	1.0926	1.0845
노면상태	1.0174	0.9978	1.0901	1.0812	1.71	0.9967	1.0172
기상상태	1.05	1.0225	1.1861	1.1709	2.2593	1.0187	1.0514
도로형태_대분류	1.2729	1.2767	1.6392	1.6488	1.2881	1.2752	1.2697

- 클러스터별 변수평균을 정리하였다.
- 변수의 특징이 비슷한 클러스터끼리 묶어 몇 가지 규칙을 예측 할 수 있었다.

요인분석 : EM 알고리즘

EM알고리즘 결과

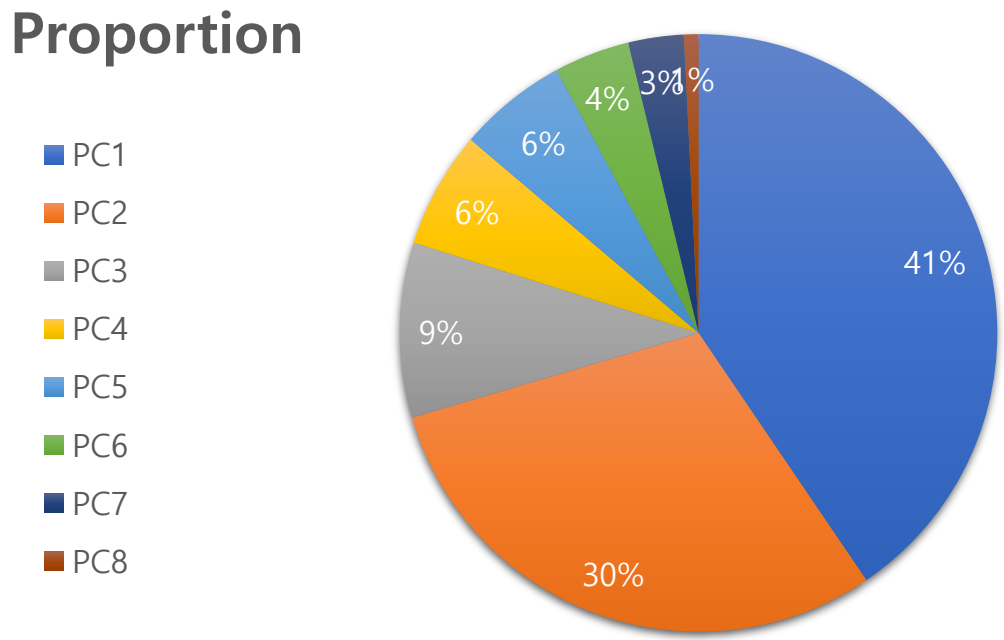
1. 도로의 상태가 건조하고 맑은 날씨 야간에 안전운전 의무 불이행 차대차사고가 단일로에서 일어났을 때 경상이 많았다. (클러스터 1, 2, 6, 7)

2. 도로의 상태가 건조하고 맑은 날씨 야간에 교차로 통행 방법 위반 차대차사고가 교차로에서 일어났을 때 경상, 중상의 사고가 일어났다. (클러스터 3, 4)

3. 도로의 상태가 습하고 흐림, 비오는 날씨 야간에 안전운전 의무 불이행 차대차사고가 단일로에서 일어났을 때 중상의 사고가 일어났다. (클러스터 5)

요인분석 : EM 알고리즘

주성분 분석 결과



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard Deviation	1.6764	1.4407	0.81018	0.66184	0.64094	0.53307	0.45597	0.23391
Proportion of Variance	0.4051	0.2992	0.09461	0.06314	0.05921	0.04096	0.02997	0.00789
Cumulative Proportion	0.4051	0.7042	0.79884	0.86198	0.92119	0.96215	0.99211	1

- 클러스터 5의 Proportion 그래프와 주성분 분석 요약 표
- 제1주성분의 기여율이 전체 데이터를 대표할 만큼 결과가 나오지 못하였기에 제4주성분까지 이용

요인분석 : EM 알고리즘

주성분 분석 결과

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
발생시간	0.00322	-0.00525	0.00725	0.00435	-0.01647	0.0026	-0.00095	-0.9998
요일	-0.9948	0.10061	0.01152	0.00784	0.00146	-0.00137	0.00462	-0.0365
사고내용	-0.3771	-0.49382	0.61531	0.6072	-0.01775	0.00567	0.08341	0.00981
사고유형_대분류	-0.0196	-0.17804	-0.5617	0.44059	0.62135	-0.26721	0.02524	-0.01223
가해자법규위반	0.01909	0.14941	0.50305	-36081	0.76941	0.33628	-0.02831	-0.0112
노면상태	0.04523	0.38967	0.03346	0.16153	0.00862	0.03861	0.90405	-0.00185
기상상태	0.07802	0.73372	0.14616	0.51037	-0.01446	-0.03318	-0.4152	0.00023
도로형태_대분류	0.00644	0.04194	0.1736	-0.14258	-0.14505	-0.96168	0.0431	0.00027

- 클러스터 5의 주성분 값
- 요일 기상상태 노면상태가 교통사고에 가장 많은 영향을 미친 변수라는 결과를 얻었다.
- 나머지 클러스터와 종합했을 때 어떤 변수를 교통사고와 연관 있는 주요 요인으로 특정하기 어려웠다.

결론 : EM 알고리즘

- EM알고리즘을 사용하여 분석한 결과 3가지 규칙들을 찾아낼 수 있었다.
- 하지만 주성분 분석을 이용한 실험에서는 만족스러운 결과를 얻지 못했다.
- 교통사고에 영향을 미치는 요인이 매우 다양하고 복합적으로 작용한다고 생각됨
- 또한 특정 요인과는 무관하게 교통사고가 발생할 수 있기 때문이라고 생각됨

더 폭넓고 구체적인 정보를 포괄하는 데이터셋의 구축과 공개가 필요

데이터셋 : 의사결정트리

데이터의 분포가 불균형한 경우 좋은 모델을 만들기 어렵기 때문에 전체 데이터에서 데이터 빈도가 너무 적은 값은 제외하거나 그룹화 시키는 단계를 거침

- 사망 건수는 1%정도로 매우 적기 때문에 중상과 함께 그룹화하여 중상으로 처리
- 철길건널목 사고는 1회 발생하였으므로 삭제
- 발생시간은 새벽(0:00~6:00), 낮(6:00~18:00), 야간(18:00~24:00)으로 구분
- 나이는 우리나라 건강보험 기준에 따라 미성년(0~18세), 청년(19~34), 중년(35~49), 장년(50~64),노년(65~)으로 구분

데이터셋 : 의사결정트리

- 전처리 후 변수별 데이터들의 분류기준과 분포비율

피처	분류기준	가짓수	경상 (%)	중상 (%)
사고내용	경상	193387	64.7	35.3
	중상	105418	-	-
계절	봄	78545	64.2	35.8
	여름	75948	65	35
	가을	75905	64.4	35.6
	겨울	68407	65.4	34.6
시간	낮	152061	64.8	35.2
	밤	93772	65.5	34.5
	새벽	52972	63.1	36.9
요일	월	40767	64.4	35.6
	화	42735	65	35
	수	43732	64.4	35.6
	목	43672	64.3	35.7
	금	47497	64.5	35.5
	토	46122	65.7	34.3
	일	34280	64.7	35.3

피처	분류기준	가짓수	경상(%)	중상(%)
사고 유형	측면충돌	93841	68	32
	후진중충돌	574	88.2	11.8
	정면충돌	10444	58.1	41.9
	추돌	66062	74.7	25.3
	횡단중	33684	42	58
	차도통행중	7261	58.3	41.7
	길가장자리구역통행중	5399	64.4	35.6
	보도통행중	3998	53.5	46.5
	기타	77542	64.8	35.2
도로 상태	건조	264085	64.9	35.1
	젖음/습기	31596	63	37
	서리/결빙	2041	68.5	31.5
	적설	1083	66.6	33.4

피처	분류기준	가짓수	경상(%)	중상(%)
날씨	맑음	258349	65	35
	비	22294	63.2	36.8
	흐림	16084	62.2	37.8
	눈	2078	66.7	33.3
가해자 법규위반	안전운전 의무 불이행	158966	65	35
	안전거리 미확보	44245	75.5	24.5
	신호위반	39416	65.4	34.6
	교차로 통행방법 위반	12384	70.4	29.6
	중앙선 침범	11297	56.5	43.5
	보행자 보호의무 위반	10754	48.3	51.7
	과속	455	21.3	78.7
	기타	21288	67.8	32.2

데이터셋 : 의사결정트리

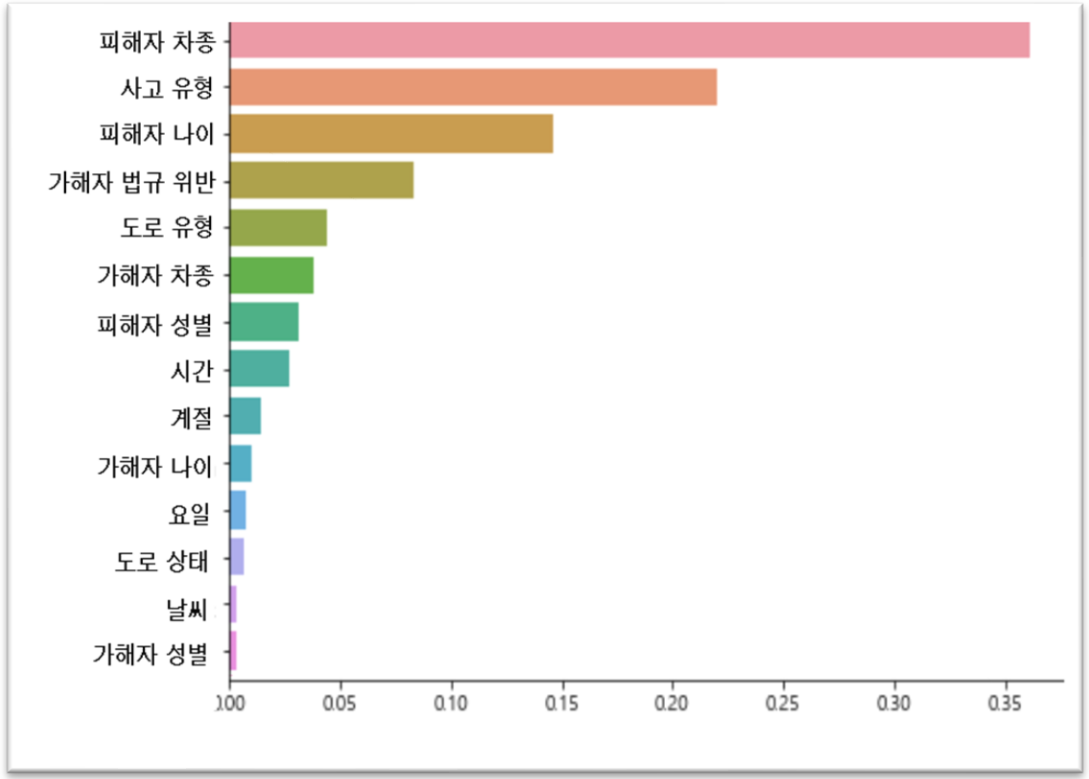
- 전처리 후 변수별 데이터들의 분류기준과 분포비율

피처	분류기준	가짓수	경상(%)	중상(%)
도로 유형	교차로내	70541	62.9	37.1
	교차로부근	50830	68.1	31.9
	횡단보도상	11889	45.5	54.5
	횡단보도부근	6037	52.5	47.5
	교량위	2916	67.2	32.8
	기타단일로	151963	66.3	33.7
	교차로횡단보도내	2508	58.3	41.7
	터널안	773	66.4	33.6
	고가도로위	755	70.6	29.4
	지하차도(도로)내	593	67.1	32.9

피처	분류기준	가짓수	경상(%)	중상(%)
가해자 차종	승용차	215661	66.7	33.3
	화물차	24215	59	41
	이륜차	19822	60.7	39.3
	승합차	19772	57.8	42.2
	자전거	9037	62.2	37.8
	원동기장치자전거	8251	61.4	38.6
	건설기계			
	특수차	1530	54.8	45.2
		517	55.1	44.9
가해자 성별	남	254924	64.6	35.4
	여	43881	65.3	34.7
가해자 나이	미성년	10424	62.6	37.4
	청년	62835	64.5	35.5
	중년	106285	63.9	36.1
	장년	89769	65.6	34.4
	노년	29492	65.3	34.7

피처	분류기준	가짓수	경상(%)	중상(%)
피해자 차종	승용차	137014	76.1	33.9
	화물차	9781	72.9	37.1
	이륜차	31428	53.9	46.1
	승합차	13235	67	33
	자전거	14368	58.2	41.8
	원동기장치자전거	12696	53.7	46.3
	건설기계			
	특수차	472	69.1	30.9
	보행자	277	71.8	28.2
		79534	50.2	49.8
피해자 성별	남	224221	66.8	33.2
	여	74584	58.3	41.7
피해자 나이	미성년	20436	62.7	37.3
	청년	70421	68.6	31.4
	중년	88565	68.5	31.5
	장년	89212	64	36
	노년	30171	47.9	52.1

데이터셋 : 의사결정트리



변수 중요도

- 랜덤포레스트를 사용하여 변수들의 중요도를 확인
- 대분류와 소분류로 나뉜 값은 중요도에 따라 하나를 선택하고 나머지를 삭제
- 예를 들어 도로유형_대분류, 도로유형_소분류에는 비슷한 유형의 변수가 포함됨

요인분석 : 의사결정트리

의사결정트리 실험과정

의사결정트리는 범주형 데이터도 분류할 수 있지만 변수 값 각각의 중요도를 확인하기 위해 원핫인코딩을 사용함

의사결정트리의 과적합을 막고 정확도를 높이기 위해 하이퍼 파라미터를 최적화

의사결정트리가 너무 깊을 경우 트리가 과적합 될 수 있어 반복 실험을 통해 최대 깊이를 4로 설정, Gini Index를 분리기준으로 설정

리프 노드가 되기 위한 최소한의 샘플 데이터 수와 노드를 분할하기 위한 최소한의 데이터 수를 조정하였다

그 결과 리프 노드가 되기 위한 최소한의 샘플 데이터 수가 30, 노드를 분할하기 위한 최소한의 데이터 수가 30일 때 가장 높은 정확도를 나타냈으며 분류 정확도는 67.83%로 나타났다

요인분석 : 의사결정트리

의사결정트리 결과

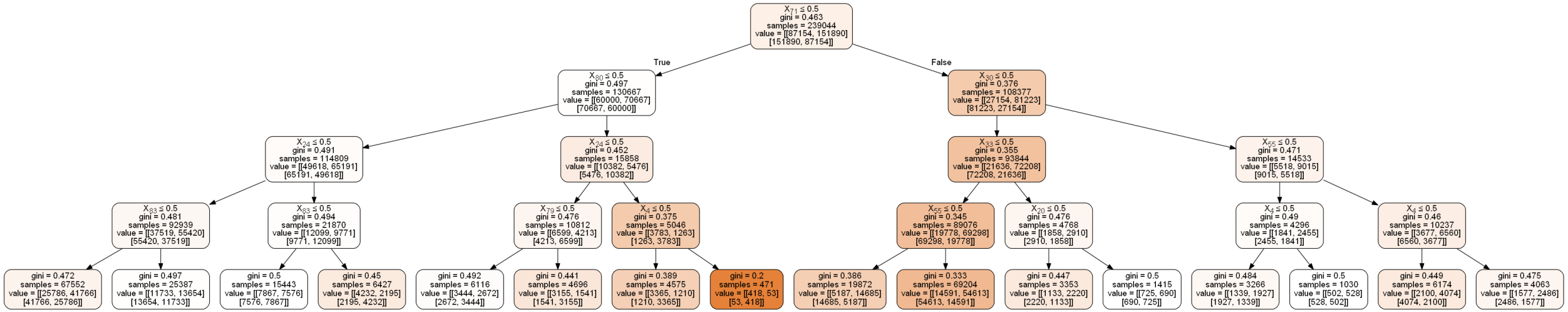
독립변수	중요도
피해자차종_승용차	0.22
사고유형_횡단중	0.11
피해자차종_보행자	0.08
피해자나이_노년	0.07
피해자차종_이륜차	0.04
피해자법규위반_신호위반	0.03
사고유형_추돌	0.03
가해자법규위반_안전거리 미확보	0.02

- 의사결정트리에서 교통사고 심각도를 설명하는데 가장 중요한 독립변수는 피해자차종_승용차로 나타남
- 전처리후 변수별 데이터들의 분류기준 표와 비교해보면 피해자차종_승용차의 경우 경상 76.1%, 중 상 33.9%이므로 피해자차종이 승용차일 경우 경상사고 가 일어날 확률이 높다고 할 수 있다.

요인분석 : 의사결정트리

의사결정트리 결과

- 구축된 의사결정트리



요인분석 : 의사결정트리

1. 피해자 차종이 승용차가 아니고, 피해자 나이가 노년이고, 도로유형이 횡단중, 사고발생시간이 새벽일 경우 가장 잘 분류되었으며 대부분 중상사고로 분류되었다.

2. 피해자 차종이 승용차이고, 가해자 법규위반이 신호위반이거나 중앙선 침범이 아니고, 가해자 차종이 승용차의 경우, 두번째로 잘 분류되었으며 대부분 경상사고로 분류되었다.

3. 피해자 차종이 승용차가 아니고, 피해자 나이가 노년이고, 도로유형이 횡단중, 피해자 성별이 여성일 경우 세번째로 잘 분류되었으며 대부분 중상사고로 분류 되었다.

4. 피해자 차종이 승용차이고, 가해자 법규위반이 중앙선 침범이고, 사고유형이 정면충돌이 아닌 경우 네번째로 잘 분류되었으며 대부분 경상사고로 분류되었다.

결론 : 의사결정트리

- 피해자 차종이 승용차인지 아닌지 여부가 교통사고의 심각도에 영향을 가장 많이 준 것으로 나타났으며 횡단보도에서 횡단중인 고령 보행자에게서 사고 가 일어날 경우 심각한 교통사고가 발생할 확률이 높았다.
- 교통사고의 주요 요인을 발견하는데 있어서 주요한 시사점을 도출할 수 있었지만 분명한 한계 도 존재했다. 첫째, 의사결정트리 분류결과가 대부분 명확히 나뉘지 않았다. 둘째, 교통사고에서 계절, 요일, 도로상태, 날씨, 가해자 성별, 가해자 나이 등 대부분의 변 수들에 크게 영향이 없는 것으로 나타났다.
- 데이터 셋의 변수만으로 교통사고 규칙을 찾는데 제한이 있으며 실제로는 교통사고에 영향을 주는 더 다양한 요인이 있거나 특정 요인과는 무관하게 교통사고가 발생할 수 있기 때문이라고 생각해볼 수 있다.

더 다양한 정보를 포함한 데이터와 적극적인 데이터 개방과 연결 필요

연구성과 및 활용계획

- 서울시 교통사고 데이터를 각각의 기법들을 활용하여 분석해 기법별로 결과를 정리하였다.
- 공통적으로 교통사고 분석을 위해서는 현재 제공되는 데이터보다 더 다양한 정보를 포함한 데이터가 필요하며 적극적인 데이터 개방과 연결 또한 필요하다고 판단된다.

전체적인 분석 결과는 교통사고율을 감소시킬 수 있는 근거로 활용되기를 기대한다.