KOSDI
ISSN : 1976-2208 (Print)
ISSN : 2671-5287 (On-line)
www.sodi.or.kr

Journal of the Society of Disaster Information
Vol. 20, No. 1, pp. 156-168, March 2024
https://doi.org/10.15683/kosdi.2024.3.31.156

Original Article

Random Forest를 활용한 고속도로 교통사고 심각도 비교분석에 관한 연구

Studying the Comparative Analysis of Highway Traffic Accident Severity Using the Random Forest Method.

이선민[1] · 윤병조[2]* · 웃위린[3]
Byoung-Jo Yoon[1]*, Sun-min Lee[2], WutYeeLwin[3]

[1]Researcher, College of Urban Science, Incheon National University, Incheon, Republic of Korea
[2]Professor, College of Urban Science, Incheon National University, Incheon, Republic of Korea
[3]Researcher, College of Urban Science, Incheon National University, Incheon, Republic of Korea

*Corresponding author: Byoung-Jo Yoon, bjyoon63@inu.ac.kr

ABSTRACT

Purpose: The trend of highway traffic accidents shows a repeating pattern of increase and decrease, with the fatality rate being highest on highways among all road types. Therefore, there is a need to establish improvement measures that reflect the situation within the country. **Method:** We conducted accident severity analysis using Random Forest on data from accidents occurring on 10 specific routes with high accident rates among national highways from 2019 to 2021. Factors influencing accident severity were identified. **Result:** The analysis, conducted using the SHAP package to determine the top 10 variable importance, revealed that among highway traffic accidents, the variables with a significant impact on accident severity are the age of the perpetrator being between 20 and less than 39 years, the time period being daytime (06:00-18:00), occurrence on weekends (Sat-Sun), seasons being summer and winter, violation of traffic regulations (failure to comply with safe driving), road type being a tunnel, geometric structure having a high number of lanes and a high speed limit. We identified a total of 10 independent variables that showed a positive correlation with highway traffic accident severity. **Conclusion:** As accidents on highways occur due to the complex interaction of various factors, predicting accidents poses significant challenges. However, utilizing the results obtained from this study, there is a need for in-depth analysis of the factors influencing the severity of highway traffic accidents. Efforts should be made to establish efficient and rational response measures based on the findings of this research.

Keywords: Highway, Traffic Accidents, Accident Severity, Machine Learning, Random Forest, Feature Importance

Received | 30 January, 2024
Revised | 14 March, 2024
Accepted | 14 March, 2024

요 약

연구목적: 고속도로 교통사고의 추세는 증감을 반복하며 도로 종류 중 고속도로에서의 치사율은 최고치를 나타내고 있다. 따라서 국내 실정을 반영한 개선대책 수립이 필요하다. **연구방법:** Random Forest를 활용해 2019년부터 2021년까지 전국 고속도로 노선 중 사고 다발 10개 노선에서 발생한 교통사고 자료로 사고 심각도 분석 및 사고 심각도에 미치는 영향요인을 도출하였다. **연구결과:** SHAP 패키지를 활용해 상위 10개의 변수 중요도를 분석한 결과, 고속도로 교통사고 중 사고 심각도에 높은 영향을 미치는 변수는 가해자 연령이 20세 이상 39세 미만, 시간대가 주간(06:00-18:00), 주말(토~일), 계절이 여름과 겨울, 법규위반이 안전운전불이행, 도로 형태가 터널, 기하구조상 차로 수가 많고 제한속도가 높은 경우로 총 10개의 독립변수에서 고속도로 교통사고 심각도와 양(+)의 상관관계를 가지는 것으로 분석되었다. **결론:** 고속도로에서의 사고 발생은 매우 다양한 요인의 복합적인 작용으로 인해 발생하므로 사고 예측에 많은 어려움이 있지만 본 연구로 도출된 결과를 활용해 고속도로 교통사고 심각도에 영향을 주는 요인을 심층적으로 분석해 효율적이고 합리적인 대응책 수립을 위한 노력이 필요하다.

핵심용어: 고속도로, 교통사고, 사고 심각도, 머신러닝, 랜덤 포레스트, 변수 중요도

서 론

현재 우리나라 교통망의 주요 축(軸)을 이루며 주요 도시를 연결하는 고속도로 교통사고 건수는 감소추세를 나타내고 있지만, 여전히 증감을 반복하고 있고 도로 종류 중 고속도로에서의 교통사고 치사율[1]은 최고치를 나타내고 있다.

2021년 기준, 고속도로 노선은 총연장 4,866km에 이르고 연간 1,698,097 천 대가 고속도로를 이용한 것으로 나타났다. 2021년 한 해 동안 전국 고속도로에서는 총 4,883건의 사고 발생과 그로 인한 사망자 수는 총 191명, 부상자 수는 총 9,708명으로 분석되었다(한국도로공사, 2022). 또한, OECD 국가별 인구 10만 명당 교통사고 건수는 회원국 평균 27.0건에 비해 우리나라의 수치는 74.2건으로 나타나 OECD 회원국 평균보다 약 2.8배 높은 사고가 발생해 현재까지도 여전히 높은 수치를 나타내고 있는 것으로 분석되었다.

이는 국내 교통사고의 감소를 위한 개선의 여지가 분명히 남아있다고 판단되며 특히 고속도로에서의 사고 발생 시 가장 높은 치사율을 보이고 있는 만큼 고속도로 사고 감소를 위한 사고 심각도 영향요인에 대한 정확한 분석과 국내의 실정을 반영한 여러 고민이 필요하다. 또한, 최근 머신러닝 및 딥러닝과 같은 빅 데이터를 활용한 새로운 분석방법이 주목을 받기 시작한 만큼, 예측 성공률을 높이는 데에 중점을 두고 이종(異種)의 대량 자료를 활용하여 분석하는 데에 장점을 보이는 빅 데이터 분석 기법을 활용하여 사고 심각도를 분석하고자 한다.

따라서 본 연구는 고속도로 노선 중 사고 다발 10개의 노선을 선정해 사고 심각도를 분석하고 이에 영향을 미치는 요인을 도출하는 데에 목적이 있다. 이때, 고속도로 교통사고 자료와 고속도로의 주요 분석단위인 콘 존(Congestion Zone, Conzone)을 결합한 자료를 활용해 분석을 진행했으며 머신러닝의 PyCaret 라이브러리를 활용해 채택된 모델인 랜덤 포레스트 회귀분석(Random Forest Regressor)으로 고속도로 사고 심각도에 미치는 영향요인 분석과 변수 중요도를 도출하였다.

연구의 공간적 범위는 전국의 고속도로 노선 중 많은 사고 건수와 다양한 사고 정보를 수집하여 일관성 있고 신뢰성 확보가 가능한 상위 10개의 노선을 선정하여 연구를 수행했다. 선정된 노선은 경부선, 영동선, 서해안선, 남해선, 수도권제1순환선, 중부선, 중부내륙선, 중앙선, 호남선, 서울양양선의 총 10개 노선이다. 시간적 범위는 2019년에서 2021년의 3년간 발생한 교통사고 자료로 분석을 진행했으며, 교통사고 심각도 분석을 위한 사고 정보는 교통사고분석시스템(Traffic Accident Analysis System, TAAS)의 자료를 활용하였다.

관련 이론 및 연구고찰

머신러닝 알고리즘 – 랜덤 포레스트(Random Forest)

머신러닝(Machine Learning)은 규칙을 일일이 프로그래밍하지 않아도 자동으로 데이터에서 규칙을 학습할 수 있도록 알고리즘과 기술을 연구하는 분야로서 인공지능의 하위 분야 중 지능을 구현하기 위한 소프트웨어를 담당하는 핵심 분야이다. 머신러닝 알고리즘은 크게 지도학습(Supervised Learning)과 비지도학습(Unsupervised Learning)으로 구분하고 지도학습은 실제 출력값과 정확한 출력값을 비교하고 오류를 검출하며 알고리즘 학습을 하고 과거의 데이터를 기반으로 미래에 있을 사건을 예측하는 데에 보편적으로 사용되고 있다.

랜덤 포레스트(Random Forest)는 분류와 회귀에 사용되는 지도학습 알고리즘으로 여러 개의 의사결정나무(Decision

1) 2021년 기준 도로별 치사율(명/100건) : 고속도로 3.9 > 군도 3.0 > 일반국도 2.8 > 지방도 2.4

Tree)를 조합한 모델로, 의사결정나무에 배깅(Bootstrap aggregation, bagging)이라는 앙상블 학습(Ensemble learning)을 적용한 모델이다. 이는 분류와 회귀에서 높은 성능을 보여 데이터 분석 시 많이 활용되고 있으며 과대 적합(Over-fitting)을 방지하기 위해 최적의 기준 변수를 랜덤하게 선택할 수 있다(Breiman, 2001). 랜덤 포레스트는 과대 적합을 해소하고 분산을 감소시켜 단일 결정 트리보다 더욱 안정적인 예측 성능을 보이고 정확도가 높다는 장점이 있지만 계산비용이 높고 규칙이 많아 추론 로직을 설명하기에는 어렵다는 단점이 있다.

Kang et al.(2022)는 인구 대비 교통사고 사망자 비율이 높은 대전시를 대상으로 보행자 교통사고 자료를 수집한 후, 기계학습을 통한 최적의 알고리즘과 심각도에 영향을 미치는 요인을 도출하고자 하였다. 분석결과, Ada Boost와 Random Forest 기법이 최적의 성능을 나타내었으며, 대전시 보행자 교통사고 심각도에 영향을 미치는 요인으로는 보행자 연령대가 70대 또는 20대인 경우, 사고유형이 횡단사고인 경우, 교통사고 심각도에 영향을 주는 것으로 나타났으며 이에 따른 사고 저감 대책을 제안하였다.

Kwon et al.(2021)은 XGBoost를 활용하여 이륜자동차 교통사고 심각도에 영향을 미치는 요인을 도출하고 이륜자동차로 인해 발생하는 심각한 교통사고 예방을 위한 법규 개편방안을 제시하였다. 분석결과, 신호위반인 경우, 운전자 연령대가 60대 이상인 경우, 이륜자동차 단독사고인 경우, 중앙선 침범 사고인 경우가 이륜자동차 교통사고 심각도에 영향을 주는 변수로 나타났다. 이를 토대로 이륜자동차 법규위반 감소를 위한 노력과 이륜자동차 안전 교육의 필요성을 강조하였다.

Kim et al.(2021)은 기계학습을 기반으로 다양한 알고리즘을 활용하여 고령 운전자에 의해 발생하는 보행자 피해사고 심각도에 미치는 요인을 분석하고자 하였다. 분석 결과, 로지스틱 모형과 SVM 모형이 상대적으로 높은 예측력을 보였고 정확도 측면에서는 Random Forest가 뛰어난 것으로 분석되어 보행자의 부상 정도를 정확히 예측하기 위해서는 Random Forest 모형의 이용을 권장한다고 주장하였다.

Rabia Emhamed AlMamlook et al.(2019)는 교통사고로 인한 경제적 손실이 가중됨에 따라 교통사고 심각도를 예측하는 모델은 교통시스템에 중요한 작업이라고 주장하였다. AdaBoost, LR(Logistic Regression), NB(Naive Bayes), RF(Random Forests)의 머신러닝 알고리즘을 활용해 사고 심각도를 분석한 결과, Random Forest 모델의 정확도가 75.5%로 가장 높게 나타났고, LR(74.5%), AdaBoost(74.5%), NB(73.1%)의 정확도를 나타내는 것으로 분석되었다.

교통사고 심각도 추정을 위한 기존 연구고찰

교통사고 심각도에 미치는 요인과 관련하여 다양한 연구가 수행되었다.

Yoon et al.(2017)은 고속도로 교통사고의 사고 건당 심각도(EPDO)를 계산하여 선형회귀 분석을 통한 사고 심각도와 EPDO에 미치는 요인을 분석하였다. 분석결과, 사고유형 중 차대 차 사고인 경우 EPDO가 2.006, 사고에 직접적인 영향을 미치는 사고요인 중 역주행 운행인 경우 EPDO가 3.142, 주간 발생 사고보다 야간 발생 사고의 경우 EPDO가 0.221로 나타났으며 운전자 연령대가 20대 미만이거나 60대 이상일 경우에서 유의확률이 나타났다.

Lee et al.(2015)는 서울, 수도권, 부산광역시의 4지 신호 교차로를 대상으로 도로의 기하구조, 교통 특성 및 환경 특성 등 다양한 요인을 고려한 교통사고 예측모형을 구축하고 교차로에서 발생하는 사고와의 상관관계를 규명하고자 하였다. 분석결과, 기존 음이항 모형보다 확률적 음이항 모형에서 설명력이 높게 나타났으며 종속변수를 총 교통사고 건수가 아닌 사고심각도별로 사고 건수를 적용한다면 각 변수가 단순 사고 발생이 아닌 사고 심각도에 미치는 변수를 파악할 수 있다고 주장하였다.

Park(2011)은 신호 교차로에서 보행자 사고 심각성을 인지하고 순서형 프로빗 모형을 이용해 횡단보행자 사고 심각도에 영향을 미치는 요인을 도출하고자 하였다. 분석결과, 사고 심각도에 영향을 미치는 변수로 보행자 연령대, 토지이용, 시간대, 차종, 제한속도 등으로 유의수준이 0.05 이하로 95%의 신뢰수준에서 통계적 유의성이 존재하는 것으로 나타났다.

고속도로 교통사고 심각도 분석방안

데이터 수집 및 전처리

연구에 사용된 고속도로 사고 자료는 도로교통안전공단 TAAS에서 제공하는 2019~2021년 자료로 고속도로 전체 노선의 사고 중 71.1%를 차지하고 있는 사고 다발 10개 노선을 선정하여 분석하였다. 10개 노선에서 발생한 사고는 총 13,130건으로 내용 미기입 등의 결측치를 제외하고 총 9,227건을 모형분석에 활용하였으며 사고 다발 10개 노선의 사고 건수는 Table 1과 같다. 또한, 한국도로공사에서 제공하는 콘 존(Conzone)[2] 자료와 콘 존 전자지도(Shp) 파일을 수집하여 분석에 활용할 변수를 선정한 후 가공하여 콘 존 마스터 테이블로 구축하였다. 이때 선정한 연속형 변수는 해당 고속도로 노선의 차로 수, 제한속도, 곡선반경 길이의 평균 및 종단경사 데이터이다. 노선의 사고분석을 위한 정확한 위치 정보 구득을 위해 사고지점의 위치 좌표를 Q-GIS 3.32.3을 활용해 모두 수집하였다.

Table 1. Number of highway traffic accidents in 2019-2021

Rank	Route	Number of Accidents	Scope of Accident
1	Gyeongbu Expressway	3,149	24.0
2	Yeongdong Expressway	1,349	34.3
3	Seohaean Expressway	1,028	42.1
4	Namhae Expressway	744	47.8
5	Sudogwon Je1sunhwan Expressway	663	52.8
6	Jungbu Expressway	593	57.3
7	Jungbunaeryuk Expressway	496	61.1
8	Jungang Expressway	494	64.9
9	Honam Expressway	420	68.1
10	Seoul-Yangyang Expressway	394	71.1

데이터 기술통계분석

고속도로 교통사고 심각도 모형분석에 앞서 수집한 자료를 통해 기술통계분석을 수행하였다. Table 2를 살펴보면 가해자 성별이 남성일 때 7,825건의 사고를 유발하여 전체 교통사고 중 86%를 차지하였고 연령대의 경우 60세 이상일 때 1,843건으로 가장 적었으나 사고 심각도가 평균적으로 7.95로 가장 높게 나타났다. 야간(18:00~06:00)보다 주간(06:00~18:00)에 발생한 사고의 비율이 약 1.7배 많았으며 주중(월~금)의 사고 발생비율은 69%로 주말(토~일)보다 약 2.2배 높았다. 계절별 고속도로 교통사고는 봄 2,209건, 여름 2,318건, 가을 2,603건, 겨울 1,966건으로 나타났으며 기상 조건은 맑은 날이 7,924건

2) 콘 존(Conzone) : 한국도로공사에서 고속도로 구간을 IC, Jct, TG 등 통행하는 차량 수가 일정한 고속도로 구간으로 분류한 개념을 의미.

으로 87%를 차지하였고 안개가 꼈을 때 사고 심각도 평균이 8.64로 가장 높았다. 노면 상태는 건조할 때가 5,390건으로 전체 사고 중 89%를 차지하였다. 사고요인 중 가해 운전자의 차종의 경우 승용차가 5,793건으로 64%를 차지하였고 승합차의 사고 심각도 평균이 14.09로 가장 높게 나타났다. 법규위반의 유형 중 안전운전불이행이 5,864건으로 전체 교통사고의 65%를 차지해 가장 높은 수치를 나타내고 있는데 과속의 경우 사고 건수는 101건이나 사고 심각도 평균이 10.72로 나타나 가장 높은 수치로 분석되었다. 도로 형태는 사고 심각도 평균이 터널에서 10.07로 가장 높게 나타났고 지하차도(도로)내에서 9.78로 다음으로 높게 나타났다. 또한, 기하구조 요인 중 고속도로 2차로 노선에서 3,202건으로 가장 많은 사고가 발생했고 심각도

Table 2. Descriptive statistics table in 2019-2021

Variables		Count	Mean	Standard Deviation
Driver gender	Male	7,825	7.87	0.092
	Female	1,271	7.30	0.175
Driver age range	under 19	105	7.02	3.954
	20-39	3,099	7.71	7.869
	40-59	4,049	7.81	7.652
	over 60	1,843	7.95	8.539
Time of day	Daytime(06:00-18:00)	5,761	8.17	8.355
	Nighttime(18:00-06:00)	3,335	7.15	6.942
Day of Week	Weekday	6,284	7.26	7.445
	Weekend	2,812	8.98	8.663
Season	Spring	2,209	7.79	6.930
	Summer	2,318	7.74	7.376
	Fall	2,603	7.75	8.393
	Winter	1,966	7.90	8.724
Weather Condition	Sunny	7,924	7.77	7.926
	Rainy	797	7.70	7.417
	Snow	65	7.77	7.399
	Fog	11	8.64	7.047
	Cloudy	284	8.44	8.137
Road surface condition	Dry	5,390	7.99	8.377
	Humid	684	7.88	7.935
	Frost	16	8.19	10.913
	Snow	8	10.50	5.855
	Etc.	5	6.40	7.733
Vehicle Type	Car	5,793	7.40	6.743
	Van	504	14.09	17.519
	Truck	2,330	7.54	6.738
	Special vehicle	332	6.54	5.289
	Construction machinery	128	8.50	8.089
	Other unknown	9	7.11	4.314

Table 2. Continue

Variables		Count	Mean	Standard Deviation
Accident type	Car to car	5,739	7.81	7.899
	Car only	256	7.50	7.963
	Car to pedestrian	101	7.27	5.948
Violation type	Speeding to fast	101	10.72	7.812
	Violation of intersection driving rules	2	21.0	0
	Signal violation	209	6.37	6.526
	Inadequate safety distance	2,405	6.96	6.589
	Failure to observe	5,864	8.22	8.460
	Violation of the center line	59	7.83	4.658
	Violation of traffic lane	293	5.90	5.838
	Etc.	163	7.80	7.605
Road type	Intersection	42	5.52	4.312
	Overpass	16	7.12	6.365
	Bridge	385	7.08	6.450
	Underpass	203	9.78	10.290
	Tunnel	325	10.07	8.876
	Etc.	8,125	7.70	7.827
Number of lanes	2 Lanes	3,202	8.60	8.123
	3 Lanes	2,245	7.41	7.110
	4 Lanes	2,170	7.40	7.284
	5 Lanes	1,236	7.92	9.551
	6 Lanes	52	5.92	4.648
Speed limit	80km/h	114	8.36	6.005
	100km/h	5,299	7.48	6.896
	110km/h	3,492	8.29	9.268

평균 역시도 2차로인 경우 8.60으로 분석되어 고속도로 2차로 노선에서의 발생한 사고가 심각도에 큰 영향을 미치는 것으로 분석되었다. 고속도로 제한속도는 100km/h에서 5,299건의 사고가 발생해 전체 사고 중 60%를 차지한 것으로 나타났으나 사고 심각도 평균은 80km/h일 때 8.36으로 가장 높게 나타났다.

구축 데이터 샘플

구축 자료의 예시는 아래 Fig. 1과 같다. 교통사고는 시간과 장소 등이 불특정하게 발생하기 때문에 사고 발생 시의 기상 상태, 도로 유형, 법규위반 등 대표되는 유형이 없는 경우 특성을 '기타'로 분류하는 경우가 존재한다(Kwon et al., 2021). 따라서 사고 심각도에 영향을 주는 요인을 파악하기 위해 사고 심각도에 영향을 미치는 요인을 명확히 알 수 없는 데이터의 열은 제외하고 분석을 진행하였으며 총 4,475행×48열의 데이터를 구축해 연구에 활용하였다.

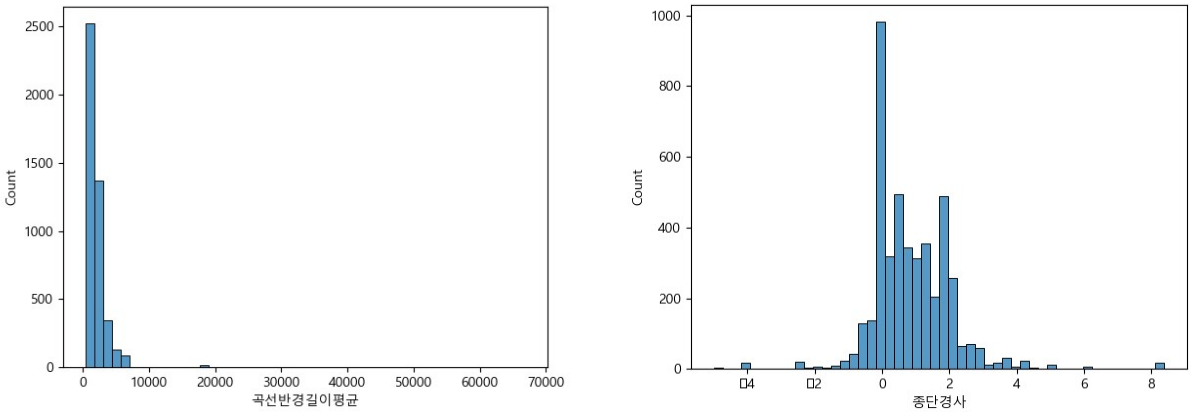
	법규위반_과속	법규위반_교차로운행방법위반	법규위반_기타	법규위반_신호위반	법규위반_안전거리미확보	법규위반_안전운전불이행	법규위반_중앙선침범	법규위반_차로위반	기상상태_기타	기상상태_눈	...	계절_여름	시간대_야간	시간대_주간	요일_주말	요일_주중	노면상태_건조	노면상태_결빙	노면상태_기타	노면상태_적설	노면상태_젖음/습기
0	0	0	0	0	0	1	0	0	0	0	...	0	0	1	0	1	1	0	0	0	0
1	0	0	0	0	0	1	0	0	0	0	...	1	0	1	1	0	1	0	0	0	0
2	0	0	0	0	1	0	0	0	0	0	...	0	0	1	1	0	0	0	0	0	1
3	0	0	0	0	0	1	0	0	0	0	...	0	0	1	0	1	1	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0	...	0	1	0	1	0	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
4470	0	0	0	0	0	1	0	0	0	0	...	0	0	1	0	1	1	0	0	0	0
4471	0	0	0	0	0	0	1	0	0	0	...	0	1	0	1	0	1	0	0	0	0
4472	0	0	0	0	0	1	0	0	0	0	...	0	1	0	0	1	1	0	0	0	0
4473	0	0	0	0	1	0	0	0	0	0	...	0	1	0	1	0	0	0	0	0	1
4474	0	0	0	1	0	0	0	0	0	0	...	0	1	0	0	1	0	0	0	0	1

4475 rows × 48 columns

Fig 1. Constructed data sample

데이터 스케일링 - StandardScaler

고속도로 교통사고 심각도 분석을 위해 사용된 수치형 변수인 곡선반경 길이의 평균과 종단경사 변수는 Fig. 2와 같이 값의 크기가 제각각으로 머신러닝 모델의 성능을 떨어뜨릴 수 있는 문제가 발생한다. 따라서 서로 다른 변수의 범위를 일정 수준으로 맞춰주는 작업이 필요하다. 이때 활용한 표준화(Standardization) 기법은 변수의 범위를 일정 수준으로 맞추어 각각의 평균을 0, 분산을 1인 정규분포를 가진 값으로 변환하고 최솟값과 최댓값의 크기를 제한하지 않아 이상치(outlier) 파악에 유리하다. 또한, Z-score를 산출하여 데이터가 평균으로부터 얼마나 떨어져 있는지 계산한 후 특정 범위를 벗어난 데이터는 이상치로 간주해 제거할 수 있다. 이때, 훈련 데이터와 시험 데이터에는 계산된 평균과 표준편차를 그대로 이용해 표준화를 진행해야 한다. 변수의 값을 표준화시키는 수식은 Equation 1과 같으며 z는 표본의 표준점수(Z-score)를, μ는 평균, σ는 평균으로부터의 표준편차이다. 본 연구에서는 표준화 기법 중 StandardScaler 기법을 활용해 표준화를 진행했으며 Fig. 3은 스케일링을 완료한 데이터의 샘플이다.



(a) Average curvature length (b) Slope gradient

Fig. 2. Distribution of numerical variable

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

	법규위반_과속	법규위반_교차로운행방법위반	법규위반_신호위반	법규위반_안전거리미확보	법규위반_안전운전불이행	법규위반_중앙선침범	법규위반_차로위반	기상상태_눈	기상상태_맑음	기상상태_비	...	요일_주말	요일_주중	노면상태_건조	노면상태_결빙	노면상태_기타	노면상태_적설	노면상태_젖음/습기	곡선반경길이평균	종단경사	EPDO
0	-0.10	-0.02	-0.14	-0.60	0.72	-0.09	-0.16	-0.07	0.39	-0.31	...	-0.67	0.67	0.36	-0.06	-0.02	-0.03	-0.35	0.41	-0.73	3
1	-0.10	-0.02	-0.14	-0.60	0.72	-0.09	-0.16	-0.07	0.39	-0.31	...	1.48	-1.48	0.36	-0.06	-0.02	-0.03	-0.35	0.41	-0.73	3
2	-0.10	-0.02	-0.14	1.68	-1.38	-0.09	-0.16	-0.07	-2.56	3.18	...	1.48	-1.48	-2.76	-0.06	-0.02	-0.03	2.83	0.41	-0.73	6
3	-0.10	-0.02	-0.14	-0.60	0.72	-0.09	-0.16	-0.07	0.39	-0.31	...	-0.67	0.67	0.36	-0.06	-0.02	-0.03	-0.35	0.41	-0.73	6
4	10.15	-0.02	-0.14	-0.60	-1.38	-0.09	-0.16	-0.07	0.39	-0.31	...	1.48	-1.48	0.36	-0.06	-0.02	-0.03	-0.35	0.41	-0.24	11
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
6069	-0.10	-0.02	-0.14	-0.60	0.72	-0.09	-0.16	-0.07	0.39	-0.31	...	-0.67	0.67	0.36	-0.06	-0.02	-0.03	-0.35	-0.35	-0.31	10
6070	-0.10	-0.02	-0.14	-0.60	-1.38	10.67	-0.16	-0.07	0.39	-0.31	...	1.48	-1.48	0.36	-0.06	-0.02	-0.03	-0.35	-0.35	-0.31	3
6071	-0.10	-0.02	-0.14	-0.60	0.72	-0.09	-0.16	-0.07	0.39	-0.31	...	-0.67	0.67	0.36	-0.06	-0.02	-0.03	-0.35	-0.35	-0.31	6
6072	-0.10	-0.02	-0.14	1.68	-1.38	-0.09	-0.16	-0.07	-2.56	3.18	...	1.48	-1.48	-2.76	-0.06	-0.02	-0.03	2.83	-0.32	-1.81	6
6073	-0.10	-0.02	7.32	-0.60	-1.38	-0.09	-0.16	-0.07	-2.56	-0.31	...	-0.67	0.67	-2.76	-0.06	-0.02	-0.03	2.83	-0.32	-1.81	3

4475 rows × 47 columns

Fig. 3. Data sample utilizing StandardScaler technique

모형성능 평가지표 선정 – RMSE

개발한 모형의 성능 평가를 위해 성능 평가지표 선정이 필요하다. 회귀모델에 사용되는 성능 평가지표에는 MAE, MSE, RMSE, R2, MAPE 등이 있으며 본 연구에서는 예측 모델에 주로 사용되는 회귀모형 성능 평가지표인 RMSE를 선정하였다.

RMSE는 MSE 모형에 루트를 씌운 모형으로 RMSE 모형을 사용해 오류 지표를 실제 값과 유사한 단위로 다시 변환하여 해석을 보다 쉽게 하고자 하였다. RMSE의 수식은 아래와 같으며, 여기서 n은 데이터의 수, $\hat{y_i}$는 예측값, y_i는 실제 값을 의미한다.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y_i} - y_i)^2}{n}} \tag{2}$$

모형 선정 - PyCaret

파이캐럿(PyCaret)은 scikit-learn 패키지를 기반으로 머신러닝 모델 워크플로우(Workflow)를 자동화해주는 파이썬 라이브러리로 분류(Classification), 회귀(Regression), 군집(Clustering) 등의 분석에 사용이 가능하다. 파이캐럿을 활용하여 여러 모델을 같은 환경에서 한 번에 비교, 분석 및 튜닝까지 가능하고 compare_models() 함수로 머신러닝 모델을 자동으로 학습한 결과를 나타내주어 각 metric별로 성능이 가장 좋은 위치에 노란색으로 표시된다. Fig. 4는 파이캐럿을 활용해 모델의 성능을 비교한 결과이다. RMSE 값을 기준으로 성능이 좋은 다섯개의 모델을 Top 5 변수로 출력하여 나타낸 결과, 최종 채택되어 고속도로 교통사고 심각도 분석에 사용된 모형은 RMSE 값이 6.8963인 랜덤 포레스트 회귀모형(RandomForestRegressor)이다.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
rf	Random Forest Regressor	4.2897	48.0194	6.8763	0.1920	0.6173	0.8582	0.3540
lightgbm	Light Gradient Boosting Machine	4.6313	50.3225	7.0455	0.1550	0.6405	0.8836	1.1670
gbr	Gradient Boosting Regressor	4.8078	54.4204	7.3206	0.0877	0.6482	0.9353	0.3050
br	Bayesian Ridge	4.8574	56.0309	7.4268	0.0664	0.6539	0.9435	0.0350
ridge	Ridge Regression	4.8746	56.2573	7.4445	0.0606	0.6575	0.9434	0.0230
omp	Orthogonal Matching Pursuit	4.8976	56.8055	7.4809	0.0512	0.6584	0.9548	0.0220
en	Elastic Net	4.9447	57.6769	7.5261	0.0451	0.6622	0.9763	0.0200
llar	Lasso Least Angle Regression	4.9679	58.0690	7.5513	0.0388	0.6652	0.9828	0.0240
lasso	Lasso Regression	4.9679	58.0690	7.5513	0.0388	0.6652	0.9828	0.0200
et	Extra Trees Regressor	4.0918	59.1592	7.6512	-0.0415	0.6492	0.8151	0.2870
dummy	Dummy Regressor	5.0160	60.6834	7.7168	-0.0027	0.6744	0.9918	0.0140
huber	Huber Regressor	4.4289	61.2037	7.7490	-0.0110	0.6196	0.6706	0.0670
knn	K Neighbors Regressor	5.0148	61.1660	7.7745	-0.0331	0.6837	0.9399	0.1060
dt	Decision Tree Regressor	4.6504	76.7264	8.7060	-0.3475	0.7247	0.8651	0.0320
par	Passive Aggressive Regressor	7.3138	112.9857	10.5648	-0.9403	0.9535	1.4034	0.0310
ada	AdaBoost Regressor	11.6414	163.6196	12.7696	-1.9254	1.1689	3.0033	0.2000
lr	Linear Regression	141.3988	29167374.7648	1714.4466	-366584.5136	0.6923	10.5615	0.0250
lar	Least Angle Regression	304.2158	132844808.1455	3678.3185	-1669712.2987	0.8933	23.4213	0.0310

Fig. 4. Results of model performance comparison using PyCaret

하이퍼 파라미터(Hyper Parameter) 성능 평가

랜덤 포레스트 모델의 성능을 평가하기 전, 모형의 높은 정확도를 얻기 위해 최적의 하이퍼 파라미터(Hyper Parameter) 설정이 필요하다. 하이퍼 파라미터는 모델을 생성할 때 연구자가 직접 설정해야 하는 변수로 높은 성능을 낼 수 있는 하이퍼 파라미터의 값은 정확히 알려진 바가 없으며 데이터마다 최적의 하이퍼 파라미터의 값이 달라질 수 있어서 매번 실험을 통해 찾아야 한다(Lee et al., 2020). 본 연구는 주어진 범위 내에서 임의의 조합을 추출해 최적의 조합을 탐색하는 랜덤서치(Random Search)[3] 방법을 활용하였다. 랜덤 포레스트의 최적 상태를 이용하기 위해 설정한 하이퍼 파라미터는 다음 Table 3과 같으며 이외의 나머지 하이퍼 파라미터는 RandomForestRegressor 기본값으로 자동 설정된다. 훈련 데이터를 활용하여 랜덤 포레스트 하이퍼 파라미터를 튜닝한 결과, 파라미터값은 'n_estimators' 1000, 'max_depth'는 None, 'bootstrap'은

Table 3. Random forest configuration hyper parameters

Division	Value
n_estimators	200, 400, 600, 800, 1,000, 1,200, 1,400, 1,600, 1,800, 2,000
max_depth	1, 3, 5, 7, 9, 11, 12, None
min_samples_leaf	1, 2, 4
min_samples_split	2, 5, 10
bootstrap	True, False
max_features	auto, sqrt

3) 랜덤서치(Random Search) : 연속된 매개변수의 값을 탐색할 때 유용한 방법으로, 탐색할 값을 직접 나열하는 것이 아니라 탐색 값을 샘플링할 수 있는 확률 분포 객체를 전달한다. 지정된 횟수만큼 샘플링하여 교차 검증을 수행하기 때문에 시스템 자원이 허락하는 만큼 탐색량 조절이 가능하다.

True, 'min_samples_leaf'는 2, 'min_samples_split'은 2, 'max_features'는 auto일 때 RMSE의 성능이 7.167로 가장 최적의 상태인 것으로 나타났다.

분석결과

최종적으로 Random Forest 활용한 고속도로 사고 다발 10개 노선(경부선, 영동선, 서해안선, 남해선, 수도권제1순환선, 중부선, 중부내륙선, 중앙선, 호남선, 서울양양선)의 교통사고 심각도 영향요인을 분석한 결과, RMSE 값이 8.093으로 측정되었다. Random Forest를 활용한 사고 심각도 모형을 SHAP 패키지를 활용해 영향을 미치는 요인을 수치화하여 분석했으며 Table 4와 같다.

인적요인 중 가해자의 성별은 여성보다 남성이 교통사고 가해자일 경우 사고 심각도에 미치는 영향이 크다. 또한, 가해자 연령대가 20대 이상에서 39세 이하일 경우 변수 중요도가 0.05로 가장 높게 나타났다. 시간·환경 요인 중 주·야간의 경우 주간(06:00~18:00)의 사고 발생이 0.11로 분석되어 야간(18:00~06:00) 발생보다 사고 심각도에 더 높은 영향을 주는 것으로 분석되었고 주중(월~금)보다 주말(토~일)의 사고 심각도가 더 높다. 계절의 경우 겨울이 0.06으로 사고 심각도에 가장 큰 영향을 주고 있으며 기상 상태의 경우 맑음, 흐림인 경우가 0.03으로 다른 변수보다 사고 심각도에 더 큰 영향을 주는 요인으로 분석되었다. 노면 상태는 젖은 상태(0.02)가 가장 높았다. 사고유형은 차량 단독(0.01)으로 발생한 경우가 사고 심각도에 영향을 미치고 있는 것으로 분석되었고 법규위반의 경우, 안전운전 불이행(0.14)이 가장 높게 나타났으며 안전거리 미확보(0.06), 과속(0.02), 차로위반(0.01)의 순서로 사고 심각도가 높게 분석되었다. 도로 형태는 터널(0.04)에서 발생한 사고가 사고 심각도에 가장 큰 영향을 미쳤고 지하차도(0.02)가 그 뒤를 이었다. 사고요인 고속도로 기하구조가 전체적인 교통사고 심각도 중 큰 영향을 미치고 있는 것으로 분석되었는데, 제한속도의 변수 중요도가 0.24로 가장 높게 나타났고 차로 수(0.15), 곡선반경 길이의 평균값(0.12), 종단경사(0.09)의 순서로 사고 심각도에 높은 영향을 미치고 있어 이를 낮추기 위한 기하구조 요소의 개선방안이 필요하다.

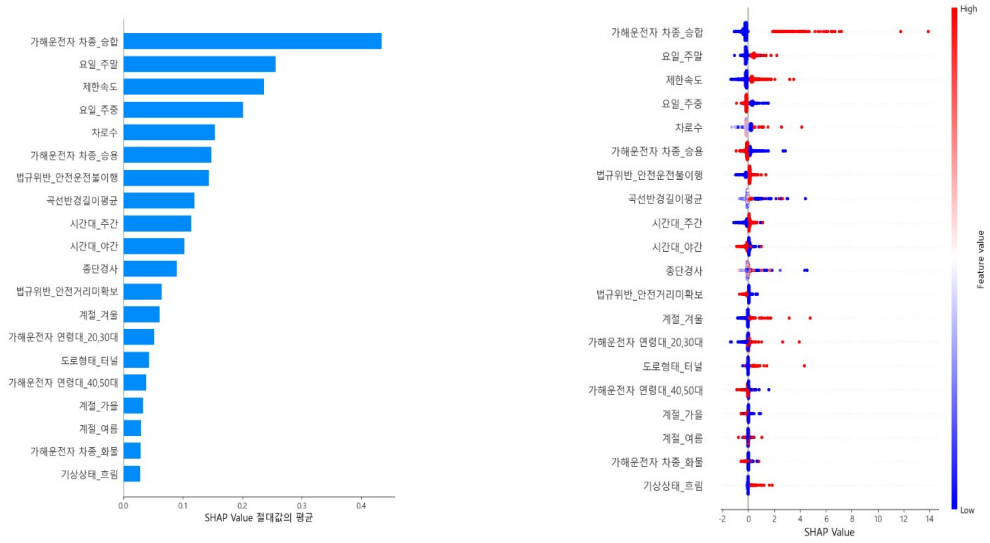
RandomForestRegressor을 활용해 분석된 고속도로 교통사고 심각도에 따른 영향 변수 중 중요도 평가를 위한 SHAP 패키지를 활용한 결과는 다음 Fig. 5의 (a) SHAP Feature Importance와 같다. 이때, 변수 중요도는 Shapley value의 절대값의 평균으로 정의되기 때문에 각각의 변수가 고속도로 교통사고 심각도에 미치는 절대 영향도라고 할 수 있으며, 가해 운전자 차종이 승합차인 경우가 가장 큰 영향을 주는 것으로 분석되었으며 주말에 사고가 발생했을 경우가 뒤를 이었다. 이후 제한속도, 요일_주중, 차로 수, 가해 운전자 차종_승용차, 법규위반_안전운전 불이행, 곡선반경 길이의 평균 순서이다. 다만, Fig 5의 (a)의 경우 절대 영향도는 각 변수가 교통사고 심각도에 어떠한 방향으로 기여했는지 그 정도를 확인할 수 없어 특성값에 따른 고속도로 교통사고 심각도에 미치는 영향력 분석이 필요하며, 그 결과는 Fig 5의 (b) SHAP Summary Plot과 같다.

SHAP Summary Plot은 변수 중요도를 통해 어떠한 변수가 고속도로 교통사고 심각도에 얼마나 영향을 주는지 평가할 수 있으며 절대 영향도가 높은 변수가 상단에 위치하게 된다. 그래프의 x축은 변수 당 Shapley value의 분포를, y축은 변수를 나타낸다. 본 연구의 feature value에 해당하는 값 중 범주형 변수는 0 또는 1로 구성되어 있어 feature 값이 존재하거나 높은 경우 붉은색으로, feature 값이 존재하지 않거나 낮으면 파란색으로 표시된다. 또한, feature가 존재하고 Shapley value가 양수(+)일 경우 고속도로 교통사고 심각도에 높은 경향을 보인다는 의미이고 feature가 존재하고 Shapley value가 음수(-)로 나타나는 경우 사고 심각도에 낮은 경향을 보인다는 의미이며 인과관계가 아닌 기여 방향의 경향성으로 해석해야 한다.

먼저, 인적요인에서 가해자의 연령대가 20세 이상 39세 미만일 때 feature value가 1일 때 Shapley value가 양수(+)로 나타나 사고 심각도가 높은 경향이 있고 40대 이상 59세 미만일 경우 음수(-)이므로 사고 발생 시 심각도가 낮은 경향이 있다. 이와 같은 해석방법으로 양(+)의 상관성을 띄어 사고 심각도가 높은 경향이 있는 변수는 주간(06:00~18:00), 주말(토~일), 계절이 여름과 겨울인 경우, 기상상태가 흐림인 경우, 가해 운전자 차종이 승합차인 경우, 안전운전불이행인 경우, 도로 형태가 터널인 경우, 차로 수, 제한속도가 높은 경우로 분석되었다.

Table 4. Feature importance

Variables		Feature Importance	Variables		Feature Importance
Driver gender	Male	0	Vehicle type	Car	0.15
	Female	0.01		Van	0.43
Driver age range	under 19	0		Truck	0.03
	20-39	0.05		Special vehicle	0.01
	40-59	0.04		Construction machinery	0
	over 60	0.01		Other unknown	0
Time of day	Daytime(06:00-18:00)	0.11	Accident type	Car to car	0
	Nighttime(18:00-06:00)	0.10		Car only	0.01
Day of week	Weekday	0.20		Car to pedestrian	0
	Weekend	0.26	Violation type	Speeding to fast	0.02
Season	Spring	0.01		Violation of intersection driving rules	0
	Summer	0.03		Signal violation	0
	Fall	0.03		Inadequate safety distance	0.06
	Winter	0.06		Failure to observe	0.14
Weather condition	Sunny	0.03		Violation of the center line	0
	Rainy	0.01		Violation of traffic lane	0.01
	Snow	0		Etc.	0
	Fog	0	Road type	Intersection	0
	Cloudy	0.03		Overpass	0
Road surface condition	Dry	0.01		Bridge	0
	Humid	0.02		Underpass	0.02
	Frost	0		Tunnel	0.04
	Snow	0		Etc.	0.02
	Etc.	0		Number of Lanes	0.15
			Geometric structure	Speed limit	0.24
				Average length of curve radius	0.12
				Longitudinal Slope	0.09



(a) SHAP feature importance (b) SHAP summary plot

Fig. 5. Feature importance results using the SHAP package

결 론

국민경제 성장과 함께 자동차의 수요 역시 폭발적으로 증가하였고 자동차의 증가는 도로의 신설 및 확장, 운전면허취득인구의 증가, 교통사고의 증가 등을 유발하였다. 편리한 교통으로부터 얻는 편익이 증가함에 따라 이로 인한 환경파괴, 교통사고 등 각종 사회적 문제를 심각히 고려해야 할 시점이다. 특히 교통 문제는 국가적 차원에서 그 심각성을 무겁게 인식하고 교통안전을 위한 체계적인 계획 수립 및 시행의 다각적인 노력이 필요하다. 이에 따라 본 연구에서는 Random Forest를 활용하여 2019년부터 2021년까지 발생한 10개 노선의 고속도로 교통사고 자료로 사고 심각도에 영향을 주는 요인을 도출하여 각 요인이 미치는 영향력을 분석하였다.

고속도로 사고의 71.7%를 차지하고 있는 10개 노선(경부선, 영동선, 서해안선, 남해선, 수도권제1순환선, 중부선, 중부내륙선, 중앙선, 호남선, 서울양양선)을 대상으로 수집한 교통사고 자료 총 4,475건 중 70%를 학습 데이터로, 30%를 평가 데이터로 활용하였다. 이때, 변수 중 기하구조 자료 중 곡선반경 길이의 평균과 종단경사 데이터의 경우 서로 다른 변수의 범위를 일정하게 맞춰주기 위해 StandardScaler 기법을 활용한 데이터 스케일링 과정을 거쳤으며 개발한 모형의 성능을 평가를 위한 지표로 RMSE를 선정하였다.

파이캐럿(PyCaret)을 활용하여 머신러닝 모델을 자동으로 학습한 결과, 랜덤 포레스트 회귀모형(RandomForestRegressor)이 사고 심각도 분석에 가장 최적의 모델로 채택되었으며 n_estimators, max_depth, bootstrap, min_samples_leaf, min_samples_split, max_features를 통해 최적화 과정을 거쳤다. 모든 최적화 프로세스는 RandomSearchCV를 활용하였다. 최종적으로 하이퍼 파라미터는 ‘n_estimators’는 1000, ‘max_depth’는 None, ‘bootstrap’은 True, ‘min_samples_leaf’는 2, ‘min_samples_split’은 2, ‘max_features’는 auto일 때 RMSE의 성능이 7.167로 가장 최적의 상태인 것으로 나타났다.

최종 구축한 RandomForestRegressor 모형은 SHAP 패키지를 활용해 고속도로 교통사고 심각도에 영향을 주는 변수의 중요도를 해석한 결과, 가해자 연령대가 20세 이상 39세 미만인 경우, 시간대가 주간(06:00~18:00)인 경우, 주말(토~일)인

경우, 계절이 여름과 겨울인 경우, 기상상태가 흐림인 경우, 가해 운전자 차종이 승합차인 경우, 법규위반이 안전운전 불이행인 경우, 도로 형태가 터널인 경우, 기하구조 상 차로 수가 많고 제한속도가 높은 경우로 총 10개의 독립변수에서 고속도로 교통사고 심각도와 양(+)의 상관관계를 나타내었다.

고속도로에서의 교통사고 발생은 매우 다양한 요인들의 복합적인 작용으로 인해 발생하는 만큼 사고의 정확한 예측에는 많은 어려움이 있다. 하지만 도로 종류 중 고속도로에서의 사고 치사율이 최고치를 달하는 만큼, 본 연구로 도출된 결과를 바탕으로 향후 연구에서는 전국 고속도로 노선의 사고 심각도 분석으로 범위를 넓혀 사고 심각도에 영향을 주는 요인을 분석하고 좀 더 효율적이고 합리적인 대응책 수립을 위한 노력이 필요하다. 특히, 관측변수를 더욱 다양화 및 세분화하여 분석에 활용한다면 좀 더 신뢰성 있는 사고 예측모형을 개발할 수 있을 것으로 판단된다.

Acknowledgement

인천대학교 2023년도 자체연구비(국제공동연구비) 지원에 의하여 연구되었음.

References

[1] Almamlook, R.E., Kwayu, K.M., Alkasisbeh, M.R., Frefer, A.A. (2019). “Comparison of machine learning algorithms for predicting traffic accident severity.” IEEE Jordan International Joing Conference on Electrical Engineering and Information Technology(JEEIT), Amman, Jordan, pp. 272-276.

[2] Breiman, L. (2001). “Random forests.” Machine Learning, Vol. 45, pp. 5-32.

[3] Kang, H.-S., Noh, M.-G. (2022). “Classifying the severity of pedestrian accidents using ensemble machine learning algorithms: A case study of Daejeon City.” The Society of Digital Policy & Management, Vol. 20, No. 5, pp. 39-46.

[4] Kim, S.-H., Lym, Y.-B., Kim, K.-J. (2021). “Classifying severity of senior driver accidents in capital regions based on machine learning algorithms.” The Society of Digital Policy & Management, Vol.19, No. 4, pp. 25-31.

[5] Korea National Police Agency (2021). Traffic Accident Statistics. Seoul.

[6] Kwon, C.-W., Chang, H.-H. (2021). “Comparative analysis of traffic accident severity of tow-wheeled vehicles using XGBoost.” Journal of Information Technology Services, Vol. 20, No. 4, pp.1-12.

[7] Lee, G.-H., Rho, J.-H. (2015). “A development of traffic accident model by random parameter: Focus on capital area and Busan 4-legs signalized intersections.” The Journal of The Korea Institute of Intelligent Transport Systems, Vol. 14, No. 6, pp. 91-99.

[8] Lee, J.-E., Kim, Y.-B., Kim, J.-N. (2020). “Hyperparameter optimization for image classification in convolutional neural network.” The Journal of Korea Institute of Convergence Signal Processing, Vol. 21, No. 3, pp. 148-153.

[9] Park, J.-S. (2011). “Severity analysis of the vehicle-pedestrian crashes at signalized intersection.” Regional Policy Review, Vol. 22, No. 1, pp. 1-12.

[10] Yoon, B.-J., Lee, S.-Y., Jung, S.-Y. (2017). “A study on the factors of highway traffic accidents affecting the EPDO.” The Korean Society of Disaster Information, Goyang, pp.251-252.