Journal of Digital Convergence
Vol. 20. No. 5, pp. 39-46, 2022

ISSN 2713-6434 / eISSN 2713-6442
https://doi.org/10.14400/JDC.2022.20.5.039

앙상블 학습기법을 활용한 보행자 교통사고 심각도 분류: 대전시 사례를 중심으로

강흥식[1], 노명규[2*]
[1]충남대학교 메카트로닉스공학과 박사수료, [2]충남대학교 메카트로닉스공학과 교수

Classifying the severity of pedestrian accidents using ensemble machine learning algorithms: A case study of Daejeon City

Heungsik Kang[1], Myounggyu Noh[2*]
[1]Ph.D. candidate Department of Mechatronics Engineering, Chungnam National University
[2]Professor, Department of Mechatronics Engineering, Chungnam National University

요 약 교통사고와 사회·경제적 손실 간의 연계성이 확인됨에 따라 사고 데이터에 기반을 둔 안전 정책 마련 및 중상·사망 등 그 심각도가 높은 교통사고의 절감 방안의 필요성이 제기되고 있다. 본 연구에서는 인구 대비 교통사고 사망자 비율이 높은 대전시를 대상지역으로 설정하고 보행자 교통사고 데이터를 수집한 후, 기계학습을 통해 최적알고리즘과 심각도 분류의 주요 인자를 도출하였다. 연구의 결과에 따르면, 적용한 9개 알고리즘 중 앙상블 기반의 학습 기법인 AdaBoost (Adaptive Boosting)와 RF (Random Forest)가 최적의 성능을 보여주었다. 이를 기반으로 도출된 대전시 보행자 교통사고 심각도의 주요 인자는 보행자의 연령이 70대 및 20대이거나 사고유형이 횡단사고에 의한 경우로 나타남에 따라 대전시 보행자 사고 저감 대책을 위한 고려요인으로 제안하였다.

주제어 : 보행자 교통사고, 사고 심각도 분류, 앙상블 기계학습, 대전시, 랜덤 포레스트, 아다부스트

Abstract As the link between traffic accidents and social and economic losses has been confirmed, there is a growing interest in developing safety policies based on crash data and a need for countermeasures to reduce severe crash outcomes such as severe injuries and fatalities. In this study, we select Daejeon city where the relative proportion of fatal crashes is high, as a case study region and focus on the severity of pedestrian crashes. After a series of data manipulation process, we run machine learning algorithms for the optimal model selection and variable identification. Of nine algorithms applied, AdaBoost and Random Forest (ensemble based ones) outperform others in terms of performance metrics. Based on the results, we identify major influential factors (i.e., the age of pedestrian as 70s or 20s, pedestrian crossing) on pedestrian crashes in Daejeon, and suggest them as measures for reducing severe outcomes.

Key Words : Pedestrian crashes, Crash severity classification, Ensemble machine learning, Daejeon city, Random Forests, AdaBoost

*Corresponding Author : Myunggyu Noh(mnoh@cnu.ac.kr)
Received March 25, 2022 Revised April 14, 2022
Accepted May 20, 2022 Published May 28, 2022

1. 서론

교통선진국들은 오래전부터 교통사고의 폐해에 대한 심각성을 인식하고 다양한 안전대책 마련을 위해 교통사고 데이터분석 연구를 추진하였다. 특히, OECD는 회원국별 교통사고 데이터를 수집하고, 온라인 통계 데이터베이스를 통해 공유하여 도로안전 연례 보고서(Road Safety Annual Report)를 발간하는 등 교통안전에 대책 수립 연구에 역량을 쏟고 있다[1]. 우리나라는 2001년 이후에 교통사고 사망자 수가 감소세로 돌아섰지만, 2018년 OECD 기준 교통사고 건수는 인구 10만 명당 420.8건으로 OECD 평균인 209.1건보다 2배 이상임을 확인하였다. 교통사고 사망자 수의 경우 전년대비 4.5% 감소율을 보였지만, 65세 이상 노인의 교통사고 사망자 수가 10만 명당 22.8명으로 OECD 평균 7.9명 보다 약 3배 많은 것으로 확인되었다[2-4].

이러한 교통사고에 대한 대책마련의 필요성에 따라 국내·외적으로 다양한 연구가 진행되고 있으며, 그중 교통사고 데이터를 활용한 연구도 매우 활발하게 이루어지고 있다. 교통사고 데이터를 활용한 연구는 현재까지도 전통 통계학적 방법인 포아송 회귀모델 또는 음이항 회귀모델 등을 기반으로 연구가 진행되고 있다[5]. 다만 회귀모델이 수치예측에 탁월한 성능이 있는 반면, 범주형 변수에 대한 분류 예측모델 적용에는 제한적인 부분이 있어 교통사고 연구에 있어서는 빈도분석을 통한 수치예측 모델 개발 연구가 주로 진행되었다[6,7].

하지만 '빅 데이터' 출현에 따른 대용량의 데이터 처리기술개발이 확대되고 통계학의 응용분야인 기계학습 기법 활용이 활성화 되면서 데이터분석 연구가 더욱 확대되었다[8]. 기계학습을 통한 교통사고 분석연구에 있어서도 기존의 제한적인 분류 예측 연구에 적극 활용되었고, 기계학습의 빠른 발전은 초연결 시대의 다양한 네트워크와 연계되어 더욱 빨라지고 있다. 더욱이 개별 기계학습 기법에 국한하지 않고 다양한 알고리즘이 어우러져 더 우수한 성능을 발휘하는 앙상블 기법 연구들이 새롭게 확대되고 있으며, 교통사고 분석연구에서도 적극 활용되고 있다[9,10]. 본 연구에서도 기계학습과 성능이 개선된 앙상블 기법을 활용하여 최적화 분류 모델을 구축하였다.

연구대상의 공간적 범위로는 국내 특·광역시중 인구대비 교통사고 사망자 비율이 높은 대전시를 선정하여 진행하였으며, 선정한 대전시의 교통사고 발생 관측데이터는 한국도로교통공단의 교통사고 분석시스템(TAAS: Traffic Accident Analysis System)으로 부터 취득한 2007년에서부터 2019년까지 교통사고 데이터를 활용하였다. 이 기간 중 발생한 대전시의 총 교통사고는 83,946건이며, 사상자는 130,104명으로 확인 되었다. 교통사고 심각도 분류체계는 사망/중상/경상/부상신고 4단계로 구분되며, 시간 경과에 따라 심각도가 고려된 교통사고발생 추이는 Fig. 1과 같다.

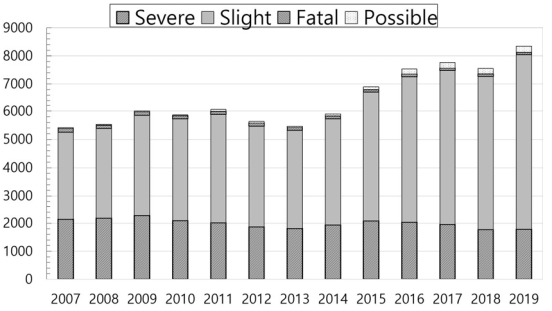

Fig. 1. Trends of crashes in Daejeon (83,946 cases)

Fig. 1을 통해 교통사고 증가폭은 확대되고 있고, 사망 및 중상 사고 건수 감소폭은 완만한 형태를 보이는 것을 확인할 수 있다. 사고 심각도에 미치는 영향을 고려할 때 차대차사고(65,241건, 77.8%)와 차대인사고(18,705건, 22.2%)로 구분할 수 있다. 사고형태별 심각도 발생 비율을 확인하면 Fig. 2와 같이 차대인 사망/중상 사고 비율이 차대차의 사고 비율보다 높게 나타나는 것을 확인하였다.

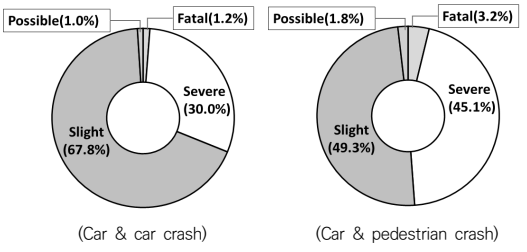

Fig. 2. Ratio by crash type

Fig. 1과 Fig. 2의 데이터 분석을 통해 대전시의 교통사고 감소 대책의 필요성을 확인할 수 있었고, 특히 차대인 사고의 경우 중상이상의 사고발생 비율이 높은 결과를 보이고 있어 보행자 피해 심각도를 감소시키기 위한 방안 마련의 시급함을 확인할 수 있다.

본 연구에서는 대전시의 교통사고 심각도 분류 예측 최적화 모델을 제시하고 모델창출 과정의 기여가 큰 변

수 도출을 통해 보행자 사고 저감 대책 방안 마련 연구를 수행하였다.

2. 관련 이론 및 연구고찰

2.1 교통사고 심각도 분류를 위한 기계학습 알고리즘

2.1.1 개별 모델 알고리즘

기계학습의 예측모델 변수의 특성에 따라 분류(classification)와 수치예측(regression)으로 구분할 수 있다[11,12]. 분류예측 기계학습 알고리즘의 학습방법에 따른 구분으로 모델이 구성되고 기계학습을 진행하는 모델 기반(Model- base) 기계학습과 모델기반 없이 데이터 입력에 따른 인스턴스 기반(Instance-base) 기계학습 방법이 있다[13].

k-최근접이웃(KNN: K-Nearest Neighbor)은 대표적인 인스턴스 기반 기계학습 알고리즘으로 모델 없이 제시된 데이터의 최근접 이웃의 클래스를 분석한다. 제시된 데이터의 클래스를 분류하게 하며, 클래스가 불균형한 데이터와 변수가 적은 다중 클래스 분류에 효율적이다. 그러나 변수의 수가 많아 질 경우 예측 정확도가 심각하게 저하 될 수 있다[14,15]. KNN 알고리즘을 통한 분류모델의 경우, 인접한 학습데이터를 몇 개까지 탐색할 것인가에 대한 식(1)과 일반화된 데이터 간 거리 측정 식(2)을 활용함에 따라 다음과 같이 나타낼 수 있다[16].

$$Misclass Error_k = \frac{1}{k}\sum_{i=1}^{k} I(c_i \neq \hat{c_i}) \quad (1)$$

$$(k = 1,2,3,...,k^*)$$

$$d(X, X^{\mu}) = (X - X^{\mu})^2 = \sum_{j=1}^{N}(x_j - x_j^{\mu})^2 \quad (2)$$

반면 SVM(Support Vector Machine) 및 로지스틱 회귀와 같은 GLM(Generalized Linear Models), LDA(Linear Discriminant Analysis)는 모델기반 기계학습 방법으로 주로 이진 분류를 위해 설계되었다. 특히, SVM은 일부 커널 방법을 통해 원본 데이터의 비선형 입력 및 출력 관계가 선형화될 수 있는 더 높은 n차원 공간으로 매핑 하여 복잡한 비선형 분류 문제를 처리할 수 있으며, n차원 공간에서 SVM은 최적의 n - 1차원 초평면을 찾아 변환된 데이터를 다른 그룹으로 분리할 수 있다[17,18]. 이때 초평면에서 가장 가까운 데이터 포인트까지의 거리가 최대화된다. 초평면은 최대 여백 초평면으로 알려져 있으며 이러한 선형 분류기는 종종 최대 여백 분류기라고도 한다[19].

2.1.2 앙상블(Ensemble) 기법

앙상블(Ensemble)기법은 여러 기계학습 알고리즘 또는 복수의 모델을 활용하여 최적의 예측모델을 생성하는 방법으로, 앙상블 기법을 통한 최적모델은 단독 기계학습 모델보다 더 좋은 성능을 발휘할 수 있다[20-22]. 앙상블 기법은 크게 3가지로 구분할 수 있다. 개별 지도학습을 통해 병렬형 예측분류 결과를 도출하는 배깅(Bagging) 기법[10,23]과 여러 기계학습 모델을 거치며 예측가능성을 높이는 가중 학습데이터 모델을 생성하는 부스팅(Boosting) 기법[24-26], 그리고 개별 기계학습 모델로부터 얻어낸 예측 값을 다시 학습 데이터로 사용하여 최적의 최종모델을 선정하는 방식인 스태킹(Stacking) 기법이 있다[9]. 미국 플로리다 지역 고속도로 교통사고 심각도 분류 기계학습연구 결과에서도 검증에 대한 정확도(Accuracy)와 재현율(Recall) 분류에 있어 기존 기계학습 알고리즘보다 스태킹 기법 적용 더 나은 성능 향상 결과를 보여 주었다[27].

앙상블기법이 활용된 선행연구에서는 수도권 고령 운전자가 보행자의 피해 심각도에 미치는 영향을 분석하는 분류예측 모델 구축으로 배깅 기법의 확장 형태인 RF(Random Forest)알고리즘을 활용한 교통사고 심각도 분류 연구를 수행하였다[10]. 사고심각도 분류 분석 결과로 고령운전자가 차도를 통행 중이거나 횡단 중인 보행자에게 야기하는 사고의 부상 정도가 심각한 것으로 나타났다.

또한, 부스팅 기법의 활용 연구로 XGBoost(eXtreme Gradient Boosting)와 Adaboost(Adaptive Boosting) 등의 알고리즘을 활용하여 10년간의 전국 이륜자동차 교통사고 데이터를 적용하여 사고 심각도 발생에 미치는 영향 요인을 발굴하는 연구를 수행하였다[26]. 해당 연구 결과를 바탕으로 이륜자동차의 심각한 사고유발 방지와 안전관리 강화를 위한 제도 개편방안을 제시하였다. 또한, 기계학습을 통해 교차로 교통사고 심각도에 미치는 요인의 주요 변수를 발굴하였고, 주요 요인을 관리항목으로 선정하여 도심부 교차로 안전성 평가에 접목시키는 연구를 수행하였다[28].

3. 기계학습을 통한 대전시 보행자 교통사고 심각도 분류

3.1 연구 절차 및 방법

기계학습을 통한 대전시 보행자 교통사고 심각도 분류 예측 연구를 위해 Fig. 3 절차에 따라 연구를 진행하였다.

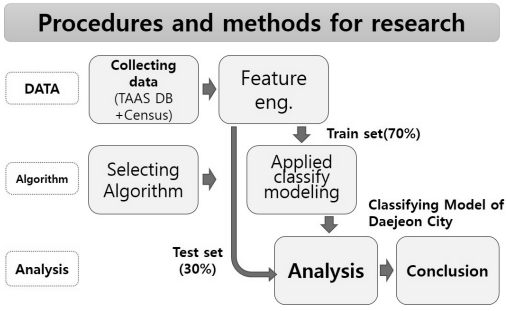

Fig. 3. Study procedure

연구절차에 따른 구체적인 내용을 Table. 1과 같이 요약하였다.

Table 1. Research process in detail

Division	Description
Data collection	· Collecting point: TAAS DB, Census DB(Daejeon) · Period: 2007~2019 year · Accident total(case): 83,963 · Variable(case): 23[num(7), factor(16)]
Feature eng.	· Car & Pedestrian (case): 18,705 · Variable(case): 18(num(3), factor(15))
Selecting algorithm	· Machine Learning Algorithm 〉GLM, LDA, SVM (Model-base) 〉KNN (Instance-base) 〉CART, Bagging, AdaBoost, GBM, RF (Ensemble)
Applied classify modeling	· Comparison of predictive performance measures among algorithms(Training set, 70%)
Analysis	· Performance measure(Test set, 30%) · Collecting affecting factor of importance class
Conclusion	· Optimized classify model · Discovering factors affecting the severity of traffic accidents

3.2 데이터수집 및 전처리

대전시 교통사고 관측데이터 83,946건 중 22.2%인 18,705건을 보행자 교통사고로 구분하여 모델분석에 활용하였다. 4단계(사망/중상/경상/부상신고)로 분류된 심각도에 대하여 범주형 데이터의 불균형에 따른 개선 작업을 위해 이진분류로 재 분류화 하였다[22]. 사고에 대한 정보를 가지고 있는 데이터의 설명력을 적극 활용하기 위하여 대물피해환산계수(EPDO: Equivalent property damage only)를 적용한 심각도 분류로 재분류 하였다[26]. 또한, 기계학습 분류 성능향상의 일환으로 대전시의 지역별 인구 변화 자료를 추가 반영하여 데이터 전처리 과정에 추가 적용하였다. 기계학습 분석을 위한 데이터 전처리 결과는 Table. 2와 같이 정리하였다.

Table 2. Data summary

Variable	Type	Class		Frequency
Severity (Y)	factor	1	Severe(Severe & Fatal)	9,361
		2	Slight(Slight & Possible)	9,344
Year (X1)	numeric	1	2007 ~ 2019 year	18,705
Season (X2)	factor	1	Spring(3~5 month)	4,620
		2	Summer(6~8 month)	4,518
		3	Autumn(9~11month)	5,118
		4	Winter(12~2 month)	4,449
Day of the Week (X3)	factor	1	Monday	2,773
		2	Tuesday	2,706
		3	Wednesday	2,726
		4	Thursday	2,779
		5	Friday	2,956
		6	Saturday	2,707
		7	Sunday	2,058
Hour (X4)	factor	1	Morning(06~12hr)	3,737
		2	Afternoon(12~6hr)	5,291
		3	Night(18~24hr)	7,100
		4	Dawn(24~06hr)	2,577
Offender Driver Age (X5)	factor	1	Teens(~19 age)	355
		2	Early 20(20~24)	1,082
		3	Late 20(25~29)	1,648
		4	Early 30(30~34)	1,564
		5	Late 30(35~39)	1,713
		6	Early 40(40~44)	1,973
		7	Late 40(45~49)	2,254
		8	Early 50(50~54)	2,330
		9	Late 50(55~59세)	1,952
		10	Early 60(60~64세)	1,270
		11	Late 60(65~69)	741
		12	After 70(70~)	530
		13	ETC	1,293
Offender Gender (X6)	factor	1	Male	13,602
		2	Female	3,856
		3	Unidentified	1,247
Census (X7)	numeric	1	Daejeon Census(2007~2019)	18,705
Number of Accidents (X8)	numeric	1	Number of people per accidents	18,705
Accident Type (X9)	factor	1	Crossing on ped	9,177
		2	On the sideroad	1,697
		3	On the sidewalk	1,120
		4	On the road	1,735
		5	ETC	4,976
Violation (X10)	factor	1	Safe driving	13,037
		2	Signal	1,461
		3	Speed	170
		4	ETC	671
		5	Center line	225

		6	Protecting pedestrian	3,141
Road Surface (X11)	factor	1	Dry	16,445
		2	Snow coverd/Icing	194
		3	Wet/humidity	2,066
Climate (X12)	factor	1	Clear	16,050
		2	Cloudy	809
		3	Rain	1,629
		4	Snow/Fog/ETC	217
Position of Accident (X13)	factor	1	Intersection	2,532
		2	Around intersection	2,314
		3	Crosswalk in intersection	762
		4	Around crosswalk(single)	351
		5	Crosswalk(single)	2,201
		6	Etc	9,753
		7	Etc_etc	792
Offender Vehicle (X14)	factor	1	Bicycle	276
		2	Bike	1,091
		3	Sedan	13,515
		4	Truck/ Equipment	2,511
		5	Van	1,312
Number of Accident in Reigion (X15)	factor	1	Daedeok-gu	2,706
		2	Dong-gu	3,805
		3	Joong-gu	3,826
		4	Seo-gu	5,909
		5	Youseong-gu	2,459
Victim Gender (X16)	factor	1	Female	9,026
		2	Male	9,679
Victim Age (X17)	factor	1	~ 9 age	1,269
		2	10 ~ 19 age	2,423
		3	20 ~ 29 age	2,966
		4	30 ~ 39 age	1,921
		5	40 ~ 49 age	2,492
		6	50 ~ 59 age	2,892
		7	60 ~ 69 age	2,230
		8	70 ~ 79 age	1,833
		9	80+	679

3.3 기계학습 알고리즘 선정 및 모델 생성

예측모델 생성을 위해 학습데이터(Training Data Set, 70%)를 활용하는 알고리즘은 모델기반 SVM, GLM, LDA와 인스턴스 기반 KNN 그리고 앙상블기법인 CART, Bagging, AdaBoost, GBM, RF 총9개 알고리즘을 활용하였다. 또한 10-fold cross validation을 적용하여 분류 모델을 생성하였다. 본 연구에서는 기계학습 연구를 위해 R프로그램을 활용하여 분류 예측모델을 생성하였다. 생성된 모델은 성능평가 방법을 활용한 알고리즘 별 비교를 통해 최적화 모델로 선정 된다. 이진분류의 대표적 평가방법으로 분류결과를 혼동행렬(Confusion Matrix)로 분류하고, 4가지 평가 성능지표인 총 정확도(Accuracy), 민감도(Sensitivity), 특이도(Specificity), 정밀도(Precision)를 통해 성능 비교평가를 진행하였다. 추가 평가단계로 ROC (Receiver Operating Characteristic) curve의 AUC 값을 통해 성능의 안정성을 확인하였다[16,29].

학습데이터를 통한 예측모델의 성능 확인을 위해 4가지 성능 평가지표별 알고리즘의 성능을 Fig. 4와 같이 확인하였다. 성능평가 지표의 총 정확도에서는 KNN과 GBM 성능이 60%이하이며, 민감도 성능에서도 KNN이 60%이하의 성능을 보였으며, GBM의 경우 민감도에서 상대적으로 우수한 성능을 보였지만 특이도와 정밀도에서는 낮은 성능을 보여주었다. 성능평가 지표 상대비교를 통해 3개의 알고리즘을 제외한 나머지 6개의 알고리즘을 활용하여 최적화 모델을 선정하기 위해 결과 분석을 실시하였다.

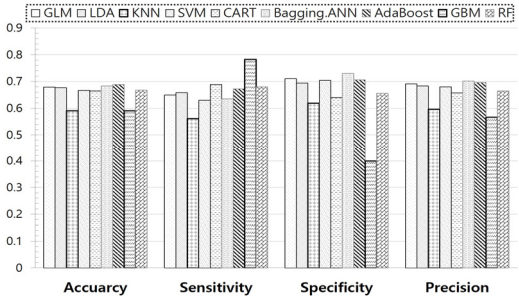

Fig. 4. Comparison of predictive performance measures among machine learning algorithms(Training set)

4. 기계학습 결과분석

4.1 대전시 교통사고 심각도 분류 모델 성능평가

학습데이터 성능 평가를 통해 선정된 6개의 알고리즘으로 시험데이터(Test Data Set, 30%)를 적용하여 모델 성능 검증을 실시하였다. 검증데이터를 적용한 성능평가 지표 별 알고리즘 성능 결과는 Fig. 5와 같다.

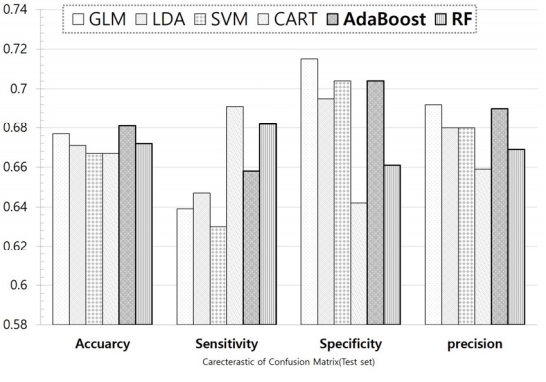

Fig. 5. Comparison of predictive performance measures among machine learning algorithms(Test set)

총 정확도에서는 AdaBoosting(0.681)이 학습데이터에서와 같이 가장 높게 나타났으며, 민감도 성능이상대적으로 낮지만 특이도와 정밀도에서는 상대적으로 높은 성능을 보여주고 있다. GLM, LDA, SVM의 경우 특이도 부분에서는 상대적으로 좋은 성능을 보여주지만, 민감도 부분에서 65% 이하로 낮게 나왔다.

앙상블기법 알고리즘의 경우 민감도 부분에서 상대적으로 우수한 성능이 보여 지지만, 특이도와 정밀도 부분에서 AdaBoost만 상대적으로 높은 성능을 보여 주고 있다. 성능평가지표의 민감도는 참으로 예측된 것 중 실제 참일 확률을 나타내는 지표로서, 교통사고 중상이상의 심각한 사고에 대한 예측 값이 실제 중상이상의 심각한 사고와 일치하는 확률을 보여주는 지표이다. 해당 확률이 높은 것은 중상이상의 심각한 사고에 대한 예측성능이 우수하다고 할 수 있다. 검증데이터를 통한 모형 성능 검증에 있어 앙상블 기법이 상대적으로 우수한 성능을 보여주고 있으며, 특히 민감도 성능지표에서 우수한 성능을 보여 주고 있다. 하지만, Cart의 경우 특이도 성능평가 지표 부분에서 65% 이하의 낮은 성능이 확인되어 Adaboost와 RF를 대전시 교통사고 심각도 분류 예측모델 알고리즘으로 선정하였다.

선정된 알고리즘으로 생성한 모델의 안정화 평가방법으로 ROC 곡선 평가를 실시하였다[29-31]. 유효 임계값 결정에 활용되는 ROC 곡선 AUC(Area Under the roc Curve)의 넓이를 통해 정확도 평가를 실시하였다. 앙상블 기법 분석결과를 통해 AdaBoost(0.753), Cart(0.731), RF(0.743) 모델별 AUC 값을 Fig. 6을 통해 확인 할 수 있다.

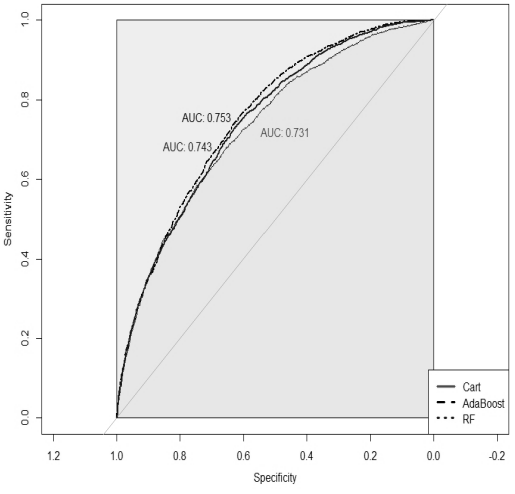

Fig. 6. ROC curve

대전시 보행자 교통사고 심각도에 영향을 미치는 요인을 발굴하기 위해 모델의 분류 성능에 미치는 주요 인자를 선정 모델로부터 도출하였다.

RF 알고리즘의 MDG(Mean Decrease Gini)[12] 도출 값은 독립변수의 분류 기여도를 측정한 값으로, RF 알고리즘의 분석시 분류 기능의 불순도(impurity)를 얼마나 감소시키는지에 대한 의미를 갖는다. 모델 분류에 중요하게 작용할 수록 이 값은 커지게 된다. RF의 MDG로 확인된 주요변수 결과를 Fig. 7과 같이 제시하였으며 Table. 2의 변수(Variable)기호와 Class 번호로 해당 변수를 확인할 수 있다.

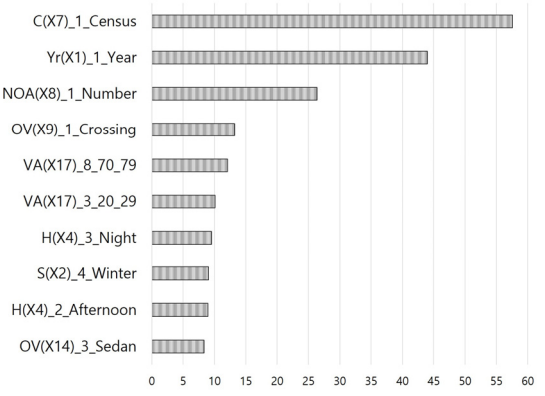

Fig. 7. Variable importance plot (RF)

기계학습 모델을 통해 확인된 대전시 교통사고의 중상이상의 사고 분류에 영향을 미치는 주요인자 10가지로 지역별 인구수 변화, 년도에 따른 변화, 사건별 사고 인원, 차도 횡단시, 70대 보행자, 20대 보행자, 야간, 겨울, 오후, 승용차 순으로 확인되었다. 이를 기반으로 도출된 대전시 보행자 교통사고 심각도의 주요 인자는 보행자의 연령이 70대 및 20대이거나 사고유형이 횡단사고에 의한 경우로 나타남에 따라 대전시 보행자 사고 저감 대책을 위한 고려요인으로 제안하였다.

5. 결론 및 향후 연구방향

본 연구는 대전시 교통사고 데이터를 활용하였으며, 보행자의 교통사고 심각도 사고에 영향을 미치는 요인을 발굴하기 위한 기계학습 분류 예측모델을 개발하였다. 4가지 성능지표(총 정확도, 민간도, 특이도, 정밀도)평가에 따라 기존 기계학습보다 앙상블기법의 AdaBoost와

RF를 최적의 알고리즘으로 선정하였으며, ROC커브의 AUC(AdaBoost: 0.753, RF: 0.743)값의 확인을 통해 선정된 알고리즘의 성능 평가 및 안정성 검증을 확인하였다. 선정된 예측모델은 발생 건수에 대한 예측(Regression)과 구별되는 사고 심각도별 분류(Classification) 모델로 분류 예측모델을 활용하여 피해 심각도에 영향을 미치는 요인을 발굴하였다. 심각도 분류에 영향을 미치는 주요 요인 발굴은 유사한 성능을 가진 AdaBoost와 RF 알고리즘의 중요도 표시방법으로 주요 인자를 발굴하였다.

분류기의 분류 역할에 기여한 공통된 주요변수 중 지역별 인구수 변화(X7), 년도에 다른 변화(X1), 사건별 사고 인원(X8)은 숫자형 변수로서, 결과의 높고 낮음이 어떻게 분류에 영향을 미치는지에 대해 불분명 하므로 저감 대책 인자로서는 배제하였다. 하지만 심각도 분류에 영향을 미치는 보행자 차도 횡단(X9)/ 70대 보행자(X17)/ 20대 보행자(X17) 3가지는 범주형 변수로서 심각도 분류 기준을 통한 변수 역할을 확인할 수 있다. 때문에 이 3가지 변수를 심각한 교통사고의 피해를 저감시킬 수 있는 주요 인자로서 제시하게 되었다.

제시된 주요 인자에 대해 교통사고 DB의 통계량으로 비교해 보면 보행자의 도로횡단시의 13년간 사고건수는 9,177건이며, 중상이상의 심각한 사고는 5,379건으로 59%에 달함을 확인할 수 있었다. 연령대별 중상이상의 심각한 사고는 다른 연령대에서는 감소세이지만 70대에서만 증가세로2007년 45건에서 2019년 61건으로 36% 증가율을 보여주었다. 20대의 경우 경사사고 발생 비율이 가장 높은 추세를 보여주고 있으며, 경사사고 3,783건 중 17%인 651건의 확인을 통해 모델에서 제시된 주요 인자에 대한 신뢰성을 보여 주었다. 더욱이, 대전시의 교통사고 정책기조와 비교해 보면, 고령자의 무단횡단 교통사고가 매우 높기 때문에 교통시설 개선 및 교통운영체제 개선에 집중투자를 추진되고 있다. 때문에 노인보호구역을 설정하여 정비하고, 고령자들의 무단횡단 단속강화와 고령자 기준에 맞는 보행자 중심의 교통신호 운영에 대한 보완 및 강화가 더욱 필요하다. 또한 고령자를 대상으로 경로당 및 노인복지센터를 중심으로 하는 교통안전교육을 강화하는 방안이 모색되어야 할 것이다.

기계학습 앙상블기법을 활용한 보행자 교통사고 심각도 분류 모델 개발로 도출한 교통사고 심각도 저감대책을 위한 주요 인자를 확인할 수 있었으며, 해당 주요인자에 대한 선택과 집중의 대책 마련을 통해 심각한 교통사고 피해를 효율적으로 감소 시킬 수 있음을 제시하였다.

교통사고 저감대책 마련을 위한 세부정책의 실현 가능성 향상을 위해서는 교통사고 심각도 분류 모델의 성능에 대한 연구의 지속성이 요구되며, 성능향상 방법으로 교통사고와 직간접적으로 연계된 설명 변수 발굴 연구가 필요함을 확인하였다. 또한, 대전시와 타 특·광역시의 비교할 수 있은 데이터를 통해 지역별 보행자 교통사고 심각도 분류예측 모델을 구축한다면, 지역별 특성에 맞는 차별화된 사고 감소 정책을 수립함으로써 분류예측 모델은 더욱 더 구체화된 방안으로 활용될 수 있을 것으로 예상된다.

REFERENCES

[1] ITF Author. (2020). Road Safety Annual Report. International Transport Forum. ISSN: 23124571 (online) DOI : 10.1787/23124571

[2] KoROAD. (2020). Comparison of traffic accidents in OECD member countries in 2018. Traffic Accidents Statistical Report. http://taas.koroad.or.kr

[3] B. G. Lee. (2020). Characteristics of Pedestrian Traffic Accidents and Reduction Plans. Daejeon Sejong Institute Basic Research Report. https://www.dsi.re.kr

[4] H. J. Jeon. (2020). Half of the fatalities in road accidents. Daejeon City invested KRW 100 billion. http://www..kmib.co.kr

[5] P. NILSSON & S. NILSSON. (2015). Application of Poisson Regression on Traffic Safety. KTH Royal Institute of Technology. www.kth.se/sci

[6] J. B. Lim, Y. H. Won, S. B. Lee & S. W. Kim. (2012). Bayesian analysis for the bivariate Poisson regression model: Applications to road safety countermeasures. Journal of the Korean Data & Information Science Society, 23(4), 851-858. DOI:10.7465/jkdi.2012.23.4.851

[7] J. P. Jeong & J. H. Choi. (2014). Poisson Regression and Negative Binomial Regression Model Fit for Traffic Accidents. Journal of the Korean Data Analysis Society, 16(1), 165-172

[8] Y. D. Kim & K. H. Cho. (2013). Big data and statistics. Journal of the Korean Data And Imformation Science Society, 24(5), 959-974

[9] S. E. Lee & H. J. Kim. (2020). A New Ensemble Machine Learning Technique with Multiple Stacking. The Journal of Society for e-Business Studies, 25(3), 1-13. DOI : 10.7838/Jsebs.2020.25.3.001

[10] S. H. Kim, Y. B. Lym & K. J. Kim.. (2021). Classifying Severity of Senior Driver Accidents In Capital Regions Based on Machine Learning Algorithms. Journal of Digital Convergence, 19(4), 25-31. DOI: 10.14400/JDC.2021.19.4.025

[11] Hints & Kinks. (2012). Classification and regression trees. International Journal of Public Health, 57, 243-246.

[12] L. Breiman, J. H. Friedman, R. A. Olshen, C.J. stone. (2017). Classification And Regression Trees. DOI:10.1201/9781315139470. Subjects Mathematics & Statistics. Pub. Location New York

[13] Z. Liu, H. Bensmail & M.Tan. (2012). Efficient Feature Selection and Multiclass Classification with Integrated Instance and Model Based Learning. Evol Bioinform Online. 8, 97-205. DOI:10.4137/EBO.S9407

[14] M. Biehl1, B. Hammer & T. Villmann.(2013). Distance measures for prototype based classification. International Workshop on Brain-Inspired Computing. 100-116. DOI:10.1007/978-3-319-12084-3_9

[15] C. W. Ko, H.M. Kim, Y.S. Jeong & J.H. Kim. (2020). A Study on Injury Severity Prediction for Car-to-Car Traffic Accidents. J. Korea Inst. Intell. Transp. Syst. Vol.19 No.4 pp.13~29. DOI : 10.12815/kits.

[16] M. Kuhn & K. Johnson. (2013). Applied predictive Modeling. Springer New York Heidelberg Dordrecht London. DOI: 10.1007/978-1-4614-6849-3

[17] S. R. Gunn.(1998). Support Vector Machines for Classification and Regression. Technical Report. UNIVERSITY OF SOUTHAMPTON

[18] X. Gu, T. Li, Y. Wang, Y., Zhang, L., Wang, Y., & Yao, J. (2018). Traffic fatalities prediction using support vector machine with hybrid particle swarm optimization. Journal of Algorithms and Computational Technology, 12(1), 20-29.
DOI : 10.1177/1748301817729953

[19] N. Cristianini & J. Shawe-Taylor. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press.
DOI: 10.1017/CBO9780511801389

[20] G. Brown. (2010). Ensemble Learning. Encyclopedia of Machine Learning, 312, 15-19.

[21] Z. H. Zhou. (2012).Ensemble methods: Foundations and algorithms. Chapman and Hall/CRC, ISBN 978-1-439-830031

[22] Y. J. Kim. Y. L. Choi, S. L. Kim, K. Y. Park & J. H. Park. (2016). A study on method for user gender prediction using multi-modal smart device log data. The Journal of Society for e-Business Studies, 21(1), 147-163, DOI: 10.7838/ jsebs.2016.21.1.147

[23] L. Breiman. (1996). Bagging predictors. Machine Learning, 24(2), 123-140.

[24] I. Syarif, E. Zaluska, A. Prugel-Bennett and G. Wills. (2012). Application of bagging, boosting and stacking to intrusion detection. International Workshop on Machine Learning and Data Mining in Pattern Recognition, 7376(8), 593-602,
DOI: 10.1007/9783642315374

[25] P. BartlettR, Y. Freund, W. S. Lee, R. Schapire. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. The annals of statistics, 26(5), 1651-1686,
DOI: 10.1214/aos/1024691352

[26] C. W. Kwon & H. H. Chang. (2021). Comparative Analysis of Traffic Accident Severity of Two-Wheeled Vehicles Using XGBoost, J. Korea Inst. Intell Transp Syst, 20(4), 1-12. DOI:10.12815/kits.2021.20.4.1

[27] J. Tang, J. Liang, C. Han, Z. Li, H. Huang. (2019). Crash injury severity analysis using a two-layer Stacking framework. Accident Analysis & Prevention, 122, 226-238. DOI: 10.1016/j.aap.2018.10.016

[28] X. Wen, Y. Xie, L. Jiang, Z. Pu &T. Ge. (2021). Applications of machine learning methods in traffic crash severity modelling: current status and future directions. Transport Reviews. 41(6), 855-879.
DOI : 10.1080/01441647.2021.1954108

[29] D. Altman, J. Bland. (1994). Diagnostic Tests 3: Receiver Operating Characteristic Plots. British Medical Journal, 309(6948), 188.
DOI: 10.1136/bmj.309.6948.188

[30] C. D. Brown & H. T. Davis. (2006). Receiver Operating Characteristics Curves and Related Decision Measures: A Tutorial. Chemometrics and Intelligent Laboratory Systems, 80(1), 24-38.
DOI: 10.1016/j.chemolab.2005.05.004

[31] T. Fawcett. (2006). An Introduction to ROC Analysis. Pattern Recognition Letters, 27(8), 861-874.
DOI: 10.1016/j.patrec.2005.10.010

강 흥 식(Kang Heung Sik) [정회원]

· 2001년 : 숭실대학교 기계공학과 (학사)
· 2003년 : 충남대학교 기계공학과 석사
· 2013년 : 충남대학교 메카트로닉스공학과 박사수료
· 관심분야 : 전자기장, 데이터분석, 교통사고
· E-Mail : linctm@cnu.ac.kr

노 명 규(Noh Myounggyu) [정회원]

· 1986년 : 서울대학교 기계설계공학과 (학사)
· 1988년 : 서울대학교 계설계공학과 석사
· 1996년 : 미국 University of Virginia 박사
· 1996년 ~ 1999년 미국 University of Iowa 대학병원 연구원
· 1999년 9월 ~ 현재 : 충남대학교 메카트로닉스공학과 교수
· 관심분야 : 전자장 해석 및 제어, 데이터분석
· E-Mail : mnoh@cnu.ac.kr