차대차 교통사고에 대한 상해 심각도 예측 연구

A Study on Injury Severity Prediction for Car-to-Car Traffic Accidents

고 창 완* · 김 현 민** · 정 영 선*** · 김 재 희****

* 주저자 : 전남대학교 산업공학과 석사과정
** 공저자 : 전남대학교 산업공학과 공학사
*** 교신저자 : 전남대학교 산업공학과 부교수
**** 공저자 : 전북대학교 경영학과 교수

Changwan Ko* · Hyeonmin Kim* · Young-Seon Jeong* · Jaehee Kim**

* Dept. of Industrial Eng., Chonnam Nat'l. Univ.
** Dept. of Business Admin., Jeonbuk Nat'l. Univ.
† Corresponding author : Young-Seon Jeong, young.jeong@jnu.ac.kr

Vol.19 No.4(2020)
August, 2020
pp.13~29

pISSN 1738-0774
eISSN 2384-1729
https://doi.org/10.12815/kits.2020.19.4.13

Received 21 April 2020
Revised 20 May 2020
Accepted 27 July 2020

요 약

자동차는 우리의 일상에 필수재가 된 지 오래지만 자동차 교통사고로 인한 사회적 비용이 국가 예산의 9%를 넘을 정도로 심각하여 이에 대한 국가적인 예방 및 대응 체계 구축이 매우 필요한 실정이다. 이에 본 연구에서는 빅데이터 분석 기법을 활용하여 차대차 교통사고의 상해 심각도를 정확히 예측할 수 있는 모형을 제시하고자 하였다. 이를 위해 과거 3년간의 전국 교통사고 발생 데이터를 토대로, K-최근접 이웃, 로지스틱 회귀분석, 나이브베이즈, 의사결정나무, 앙상블 알고리즘을 적용하여 각 모델의 상해 심각도 분류의 성능을 비교 분석하였다. 특히 이 과정에서 각 상해 심각도 수준 간의 데이터 수에 차이가 있음에 주목하여 표본수가 많은 그룹에 대해서는 과소표본추출을 시행하는 등의 방법을 통해 분류 예측의 정확도를 높일 수 있었고, 분산 분석을 통해 모델의 유의성을 검증하였다.

핵심어 : 교통사고, 상해 심각도, 과소표본추출, 예측모델

ABSTRACT

Automobiles have long been an essential part of daily life, but the social costs of car traffic accidents exceed 9% of the national budget of Korea. Hence, it is necessary to establish prevention and response system for car traffic accidents. In order to present a model that can classify and predict the degree of injury in car traffic accidents, we used big data analysis techniques of K-nearest neighbor, logistic regression analysis, naive bayes classifier, decision tree, and ensemble algorithm. The performances of the models were analyzed by using the data on the nationwide traffic accidents over the past three years. In particular, considering the difference in the number of data among the respective injury severity levels, we used down-sampling methods for the group with a large number of samples to enhance the accuracy of the classification of the models and then verified the statistical significance of the models using ANOVA.

Key words : Traffic incident, Injury severity, Undersampling, Prediction model

Ⅰ. 서 론

1. 연구의 배경 및 목적

우리나라의 비약적인 경제발전과 더불어 도로교통은 국가 경제의 중추적인 역할을 해오고 있다. 2019년 기준 2,300만대가 넘는 자동차가 등록되어 있으며 연간 자동차 등록대수는 매년 높아지고 있다. 차량 없이는 일상생활이 어려울 정도로 자동차는 우리 생활 속에서 중요한 역할을 하고 있지만 교통사고로 인해 국가의 소중한 자원인 인명과 재화가 손실되고 도로의 효율성이 저하되는 등 피해가 날로 증가하고 있다(Korea Road and Traffic Authority, 2014).

2017년 기준 대한민국 인구 10만 명당 교통사고 사망자 수는 8.1명으로 OECD 회원국 중 최상위권에 속해 있으며 OECD 회원국 평균인 5.1명에 비해서는 약 1.6배 많은 수치이다(Korea Road and Traffic Authority, 2019). 교통사고로 인해 소비되는 사회적 비용은 총 23조 6,805억 원으로 국가 전체 예산의 9.7%에 이르는 규모이다. 사회적 비용은 크게 인적피해 비용, 물적피해 비용, 사회기관 비용으로 구분되며 그 중 인적피해 비용이 12조 553억 원으로 가장 많은 부분을 차지하였다. 또한 인적 피해 비용은 심각도에 따라 달리 책정되는데 사고로 인한 1인당 평균 인적 피해 비용은 사망 4억 4,517만원, 중상 6,292만원, 경상 424만원, 부상신고 204만원으로 보고되어 있다(Korea Road and Traffic Authority, 2018). 이에 정부는 2022년까지 교통사고 사망자 수를 절반으로 감축하는 '교통안전 종합대책'을 발표했으며 경찰청과 국토교통부는 GPS 좌표 정보를 활용한 데이터 분석을 주요과제로 선정하고 교통사고 감소를 위한 노력을 기울이고 있다. 즉, 도로 기하구조 특성, 차량 특성, 도로 이용자 특성과 같은 교통사고에 영향을 미칠 수 있는 요인 데이터를 체계적으로 관리하여, 각 요인들이 교통사고에 어떤 영향을 미치는지에 대한 과학적인 분석을 검토하고 있다.

이러한 배경에서 본 연구에서는 교통사고에 영향을 미칠 수 있는 각종 요인들을 파악하고, 이들을 통해 교통사고의 심각도를 예측할 수 있는 모형을 연구하고자 하였다. 이를 위해 도로교통공단(Korea Road Traffic Authority)과 교통안전정보관리시스템(Traffic Safety Information Management Complex System; TMACS)에서 제공받은 2015년부터 2017년 사이에 발생한 전국의 교통사고 데이터를 활용하여, 데이터마이닝(Data mining) 기법 중 분류(Classification) 문제에 주로 사용되는 K-최근접 이웃(K-Nearest Neighbor; KNN), 로지스틱 회귀(Logistic regression; LR), 나이브베이즈(Naive Bayes; NB), 의사결정나무(Decision Tree; DT), 앙상블(Ensemble) 알고리즘을 이용하여 상해 심각도 예측 모델을 학습하고 각 알고리즘별 성능을 비교 분석하였다. 특히 각 상해 심각도 수준 간의 데이터 수에 차이가 있음에 주목하여 표본수가 많은 그룹에 대해서는 과소표본추출을 시행하는 등의 방법을 통해 분류 예측의 정확도를 높이고자 하였다. 또한 제시된 방법을 통해 각종 요인 변수를 토대로 사고 후에 예상되는 피해 정도를 추정하고, 나아가 응급조치 수행 시 사고피해자의 2차 상해를 최소화하기 위한 도구로 활용될 수 있도록 하였다.

본 연구의 나머지 구성은 다음과 같다. 제 2장에서는 교통사고 상해 심각도 예측모델에 관한 관련 연구 고찰 및 모델 학습에 사용된 데이터마이닝 기법의 이론적인 배경을 설명한다. 제 3장에서는 본 논문에서 사용된 데이터와 데이터 집합의 구성 과정을 설명한다. 제 4장에서는 상해 심각도 예측 모델을 구축하고 모델별 민감도를 비교한다. 마지막으로 5장에서는 결론과 함께 본 연구의 한계점 및 향후 연구방향을 제시한다.

Ⅱ. 선행 연구 및 관련 연구 고찰

1. 선행 연구 고찰

국외 교통사고 상해심각도 예측 모델에 대한 연구는 교통사고에 영향을 주는 특정 요인을 고려하거나 통계적 모델 또는 기계학습 기법을 이용하는 등 다양한 방면으로 연구가 진행되고 있다. Uddin and Huynh (2020)은 미국 오하이오 주에서 발생한 2011년부터 2015년 교통사고 데이터를 이용하였으며, 날씨 상태에 따른 트럭 운전자의 상해심각도 예측 연구를 수행하였다. 위의 연구에서는 5단계(Fatal, Disabling, Evident, Possible, No) 상해수준을 3단계(Major, Minor, No)로 통합하고, 날씨 상태(Normal, Rainy, Snow) 별 트럭 운전자의 상해정도 예측을 위해 혼합 로지스틱 회귀모형을 이용하였다.

Jeong et al.(2018)은 2016년부터 2017년 사이에 발생한 미국 미시건 주 교통사고 데이터를 이용하여 운전자의 상해 심각도를 예측하였다. 데이터 불균형 문제를 해소하기 위해 과소·과대표본추출 방법과 앙상블 방법을 사용하였으며, 기하평균을 이용해 분류모델의 성능을 평가하였다. 분류모델로는 로지스틱 회귀분석, 의사결정 나무, 그래디언트 부스팅, 신경망, 나이브베이즈 방법이 사용되었다.

국내 교통사고 상해 심각도 예측 모델 개발에 대한 연구의 경우 국외 연구에 비해 활발한 편은 아니지만 꾸준히 이루어져 왔다. Sohn and Shin(1998)은 1996년 서울에서 발생한 교통사고 데이터를 활용하여 교통사고 심각도 분류모형을 추정하였다. 모형 추정에 사용되는 중요변수를 찾기 위해 χ^2 검정과 의사결정 나무를 이용하여 주요 변수를 추출했다. 종속변수인 사고 심각도 항목의 경우 5개 범주(사망, 중상, 경상 부상신고, 물적피해)로 이루어져 있으나 연구에서는 3개의 범주(치명적 상해, 경미한 상해, 물적피해)로 줄인 후 인공신경망, 의사결정나무, 로지스틱 회귀분석 모델을 사용해 교통사고 심각도 분류 예측모형 구축 및 분류 정확도를 비교하였다.

Lee and Lee(2009)는 미국 NASS(National Automotive Sampling System)의 GES(General Estimates System) 데이터 집합 중 2008년 교통사고 데이터를 이용하였으며 6단계로 구분되는 상해 심각도 수준을 하나씩 예측하고 단계별로 분류하는 의사결정나무 기반의 앙상블 모델을 개발하였다. 심각도 수준을 예측하기 위해 단계별로 다른 의사결정나무를 생성하고 적중률 향상을 위해 ROC(Receiver Operating Characteristic) 곡선을 이용하여 최적 임계값(Threshold)을 단계별로 적용시켰다.

Lee and Heo(2011)는 미국 NASS의 GES 데이터 집합 중 2008년 교통사고 데이터를 이용하여 총 5개의 상해 심각도 수준을 예측하는 하이브리드 모델을 구축하였다. 모델은 4단계로 이루어져 있고 각 단계별로 하나의 상해 심각도를 순차적으로 예측하였으며, 인공 신경망, 의사결정나무, 로지스틱 회귀분석, 사례기반 추론 기법 중 가장 좋은 성능을 내는 모델을 선택함으로써 모델의 적중률을 높였다.

Hong et al.(2015)은 TMACS로부터 제공받은 2011년부터 2013년에 발생한 24,285건의 강원도 교통사고 데이터를 이용하여 교통사고 상해 심각도 예측 모델을 구축하였다. 데이터 불균형 문제를 해소하기 위해 과대샘플링(Upsampling)기법 기반의 One-vs-All(OVA)방식을 고려한 분류기법으로 CART(Classification And Regression Trees) 의사결정나무 알고리즘과 랜덤포레스트 알고리즘이 결합된 하이브리드 기법을 활용하였다.

국내 대다수의 연구들은 대체로 교통사고 데이터 확보의 어려움으로 인해 미국의 데이터를 사용하거나 국내 일부 지역에 국한된 데이터를 이용해 왔다. 그러나 교통사고 피해자들의 상해 심각도를 예측하기 위해 고려되는 변수들은 국가 간 교통 문화와 환경 등의 차이에 큰 영향을 받기 때문에 외국의 자료나 국내 일부 지역만을 대상으로 한 연구 결과를 그대로 활용하는 것은 적합하지 않다. 기존의 연구 중 Lee and Lee(2009)와 Lee and Heo(2011)의 경우 미국 NASS의 GES 2008년 데이터를 사용했는데, 한국과 미국 사이에는 지형적

특성이나 도로의 형태와 환경에서 큰 차이가 있고, 교통법규와 교통의식에 따른 문화적 차이도 상당해, 이를 고려하지 않은 연구 결과를 국내에 그대로 적용하는 것은 적합하지 않다. 또한, Sohn and Shin(1998)과 Hong et al.(2015)의 연구는 국내 교통사고 데이터를 사용하였으나, 각각 서울과 강원도에 국한되어 있다. 국내에서도 지역별 교통 환경이 서로 달라 운전자의 행동과 사고유형의 차이가 존재하는데, 특정 지역의 교통사고 데이터만을 이용해 나온 연구 결과를 통해 국내 교통사고 현실을 파악하는 것은 적합하지 않은 측면이 있다(Hahn et al., 2002). 다만, Hong et al.(2015)의 연구는 다른 연구들과 비교할 때 향상된 상해 심각도 적중률을 보여주었다는 점에서 의미가 있다. 하지만 모델 학습 시 상해 심각도에 매우 직접적으로 연관되는 '사망자수', '부상자수'를 이용하여 '사망', '중상', '경상' 의 적중률을 대폭 상승시켰으며, 이는 객관적 평가 측면에서 적합하지 않다고 판단된다.

이상에서 살펴본 바와 같이 기존의 연구들은 국내 자동차 사고 전반을 반영할 수 있는 데이터의 부재 또는 모형의 한계로 인해 현 시점에서 국내 자동차 교통사고 심각도 예측에 활용하기에는 한계가 있다. 이에 본 연구에서는 국내 교통사고에 초점을 맞추고 국내 전역에서 발생한 교통사고 데이터를 대상으로 하여 우리의 교통현실을 충분히 반영하고, 상해 심각도 수준과 사실상 대동소이한 변수들을 제외하여 예측 모델로서의 가치를 확보하고자 하였다. 그리고 다양한 데이터마이닝 기법을 적용하여 각각의 성능을 비교 분석함으로써 실제 교통사고 환경에서의 활용 가능성을 높인 상해 심각도 예측 모델을 개발하고자 하였다.

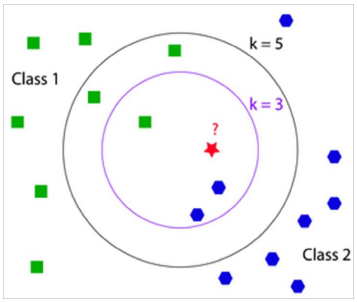
2. 데이터 마이닝 이론 고찰

1) K-최근접 이웃

K-최근접 이웃 분류 모형은 1968년 Cover and Hart(1967)에 의해 제안된 알고리즘으로 새로운 한 개체에 대하여 훈련 데이터 집합 안에 있는 K개의 가장 가까운 개체와 유사도를 비교하고, 가장 높은 빈도의 그룹으로 개체를 분류하는 방법이다. 유사도는 일반적으로 거리 개념을 이용하며 본 연구에서는 거리 개념 중 대표적으로 사용되는 유클리드 거리(Euclidean distance)를 이용하였다. 유클리드 거리는 식 (1)에서 제시하였으며 X와 Y는 두 점을 나타내고 x_i와 y_i는 각 점에 속한 좌표를 나타낸다.

$$D(X, Y) = \sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} \quad \cdots\cdots (1)$$

<Fig. 1>은 기존의 데이터 집합에 새로운 데이터 '★'가 들어왔을 때 분류되는 과정을 보여주며 입력 값과 거리가 가장 가까운 K개의 데이터('■', '⬣')를 이용한다. 새로 들어온 데이터(★)는 K=3인 경우 Class 2(⬣)



<Fig. 1> Concept of K-nearest neighbor classification

로 분류가 되지만, K=5일 경우에는 Class 1(■)로 분류 되는 것을 확인할 수 있다. 이와 같이 K-최근접 이웃은 모델자체가 단순하고 효율적이기 때문에 이해하기 쉽고 결과에 대한 해석이 용이하다는 장점을 가지고 있지만 K값 및 유사도 측정방법에 따라 결과 값이 달라지는 단점이 있다.

2) 로지스틱 회귀

로지스틱 회귀는 종속변수가 이진 또는 다범주인 경우 사용할 수 있는 분석방법으로 독립변수들의 선형결합을 통해 각 범주의 확률값을 구하고, 임계값(Cutoff; C)을 이용하여 n개의 범주로 분류하게 된다. 로지스틱 회귀의 확률적 선형 분류기를 구하는 과정으로 초기 매개변수를 초기화 시킨 후, 적절한 결정 경계를 찾기 위해 $h_\theta(x)$ 함수를 이용하여 확률값을 계산한다. $h_\theta(x)$는 식 (2)과 같이 각 독립변수들의 곱으로 표현할 수 있으며, 각 범주의 확률값은 식 (3)에 제시된 시그모이드(Sigmoid) 함수의 입력 값으로 사용된다(Isaac and Harikumar, 2016).

$$h_\theta(x) = \theta^T x = \theta_0 x_0 + \theta_1 x_1 + ... + \theta_n x_n \quad \cdots\cdots (2)$$

$$g(\theta^T x) = \frac{1}{(1+e^{-\theta^T x})} \quad \cdots\cdots (3)$$

식 (3)을 통해 나온 결과는 0과 1 사이에 값을 갖게 되는데 지정된 임계값을 기준으로 식 (4)와 같이 두 개의 범주로 분류할 수 있다.

$$h_\theta(x) = \begin{cases} 0, & \text{if } g(\theta^T x) \leqq C \\ 1, & elsewhere \end{cases} \quad \cdots\cdots (4)$$

로지스틱 회귀는 구현이 간단하며 범주형 반응변수의 확률을 예측할 수 있을 뿐만 아니라 분류문제에도 적용이 가능하다는 장점이 있으며 입력변수 x_i가 종속변수에 미치는 영향을 파악하는 데에도 사용할 수 있다.

3) 나이브베이즈

나이브베이즈는 조건부 확률을 이용하여 범주형 종속변수를 갖는 데이터에 적용 가능한 분류기이다. 기본 원리로는 조건부 확률에 베이즈 정리(Bayes's Theorem)를 적용하고, 입력 데이터 집합의 모든 특징들에 대해 독립성을 가정하여 입력 벡터의 확률을 계산한 후 분류하는 과정을 거친다. 식 (5)는 베이즈 정리를 이용하여 $x_{1,...,}x_n$로 구성된 집합 X가 C_k에 속할 확률을 수식으로 나타낸 것이다.

$$P(C_k|X) = \frac{P(X|C_k)P(C_k)}{P(X)} \quad \cdots\cdots (5)$$

식 (5)을 계산하는 과정에서 독립성을 가정하지 않고 계산을 수행할 경우 식 (6)과 같은 복잡한 계산을 수행해야한다. 하지만 나이브베이즈는 독립성을 가정하기 때문에 각 특성의 확률에 대한 곱으로 표현이 가능하며, 식 (7)과 같이 간단하게 계산할 수 있다(Jeong, 2017).

$$
\begin{aligned}
P(C_k,x_1,...,x_n) &= P(C_k)P(x_1,\cdots,x_n|C_k) \\
&\quad P(C_k)P(x_1|C_k)P(x_2,\cdots,x_n|C_k,x_1) \\
&\quad P(C_k)P(x_1|C_k)P(x_2|C_k,x_1)P(x_3,\cdots,x_n|C_k,x_1,x_2) \\
&\quad P(C_k)P(x_1|C_k)P(x_2|C_k,x_1)\cdots P(x_n|C_k,x_1,x_2,x_3,\cdots,x_{n-1})
\end{aligned} \quad \cdots\cdots (6)
$$

$$
\begin{aligned}
P(C_k|x_1,\cdots,x_n) &\propto P(C_k,\cdots,x_n) \\
&\quad P(C_k)P(x_1|C_k)P(x_2|C_k)P(x_3|C_k)\cdots \\
&\quad P(C_k)\prod_{i=1}^{n}P(x_k|C_k)
\end{aligned} \quad \cdots\cdots (7)
$$

나이브베이즈는 개념이 단순하고 계산이 효율적이며 좋은 성능을 보이는 장점이 있으나 좋은 성능을 위해 많은 수의 데이터를 필요로 하며 종속변수의 범주가 학습에 사용되지 않은 경우 0의 확률을 가지게 되는 단점이 있다. 이를 해결하기 위해 일반적으로 Laplace 평활화(smoothing)를 적용하여 0의 확률을 갖지 않도록 0보다 큰 값을 설정하여 계산한다.

4) 의사결정나무

의사결정나무란 관심대상이 되는 집단을 나뭇가지처럼 몇 개의 소집단으로 분류하거나 예측하는 데이터마이닝 기법으로 입력 변수 $X_{1,..}X_{j,..}X_d$와 $K=\{c_1,..,c_{k,..}c_K\}$개의 클래스로 이루어진 종속변수 Y를 이용해 나무 T를 성장시킨다. 나무 성장 시 노드의 최적 분리기준으로 고려되는 입력 변수(j)와 기준점(s)을 찾기 위해 상쇄 관계(Trade-off)에 있는 나무 크기와 최종 노드의 불순도를 고려한 목적함수를 설정하고 목적함수 값을 최소로 하는 나무를 생성한다(Hastie et al., 2009). 사용되는 대표적인 분리기준으로 CHAID(Kass, 1980), CART(Breiman et al., 1984), C4.5(Quinlan, 1993) 알고리즘 등이 있다. CHAID(Chi-squared Automatic Interaction Detection)은 카이제곱 검정(Chi-Squared Test)이나 F-검정(F test)을 이용하고 CART(Classification and Regression Tree)와 C4.5 알고리즘은 지니 지수(Gini Index) 또는 엔트로피 지수(Entropy Index)를 사용한다. 식 (8)은 여러 분리기준 알고리즘 중 엔트로피 지수에 사용되는 목적함수를 보여준다.

$$
\min_{T=\{R_1,R_2,...,R_{|T|}\}}\sum_{t=1}^{|T|}N_t\left\{-\sum_{k=1}^{K}\hat{p}_{tk}\log_2\widehat{p_{tk}}\right\}+\alpha|T| \quad \cdots\cdots (8)
$$

식 (8)에서 $\hat{p}_{tk}$는 최종 노드 t에 포함된 클래스 k의 비율을 나타내고 R_t는 최종 노드 t의 영역을 나타낸다. $|T|$와 N_t는 최종 노드의 개수와 노드 t에 포함된 관측치의 개수를 의미한다. 식 (8)의 관점에서 j와 s를 찾기 위해 욕심 알고리즘(Greedy Algorithm)을 사용하는 것은 계산적인 측면에서 불가능하다. 따라서 j와 s를 고려하는 이진 분할 조건식을 이용하며, 식 (9)에 제시하였다.

$$
R_{l(j,s)}=\{X|X_j\le s\} \text{ and } R_{r(j,s)}=\{X|X_j>s\} \quad \cdots\cdots (9)
$$

각 노드의 엔트로피 합이 최소가 되도록 하는 j와 s를 갖는 엔트로피 지수는 식 (10)과 같으며 |.|은 해당 노드의 크기를 의미한다.

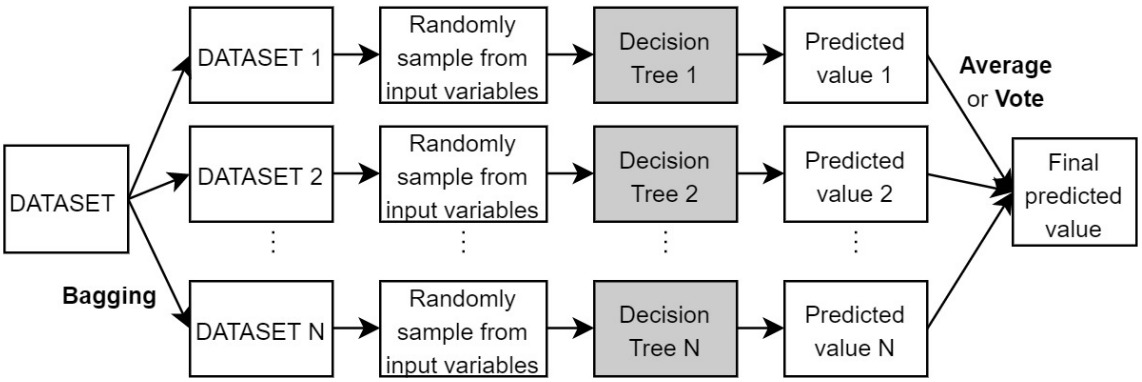
$$
Entropy(j,s)=|l(j,s)|\left(-\sum_{k=1}^{K}\hat{p}_{l(j,s)k}\log_2\hat{p}_{l(j,s)k}\right)+|r(j,s)|\left(-\sum_{k=1}^{K}\hat{p}_{r(j,s)k}\log_2\hat{p}_{r(j,s)k}\right) \quad \cdots\cdots (10)
$$

의사결정나무의 장점은 파라미터가 없어 파라미터 최적화가 필요 없고, 구축 과정의 단순성과 빠른 수행 능력 그리고 부모와 자식마디로 표현되는 직관적인 나무구조를 시각적으로 제시할 수 있기 때문에 결과에 대한 이해와 해석이 비교적 수월하다는 장점을 가지고 있다(Lee and Lee, 2018).

5) 앙상블 기법

앙상블 기법은 주어진 통계적 학습이나 모델 적합 기술의 예측성능을 향상시키기 위한 방법으로 단일 분류기를 사용하는 대신 여러 분류기의 선형 조합을 통해 결과를 도출한다(Gentle and Hadle, 2012). 앙상블 방식은 단일 분류기를 이용하는 것보다 성능이 더 좋다고 알려져 있으며 크게 배깅(Bagging : Bootstrap aggregating)과 부스팅(Boosting)으로 나눌 수 있다(Dietterich, 1997). 배깅은 학습 집합을 복원 추출하여 여러 개의 부분집합을 만들고 이를 이용해 서로 다른 분류기를 만들어 입력 데이터에 대한 결과를 산출한다. 모델의 성능은 산출된 결과의 평균 값을 이용해 계산한다. 부스팅은 학습된 이전 분류기를 통해 각 변수들의 가중치를 변화시키며 결과를 출력하는데, 이전 분류기에서 정확히 예측한 데이터는 분류가 가능하다고 판단하여 가중치를 낮추지만 예측에 실패한 데이터는 가중치를 높여 다음 분류기에서는 맞출 수 있도록 학습을 진행한다. 부스팅의 경우 분류하기 힘든 데이터에 대해 높은 가중치를 적용시켜 학습을 진행하므로 배깅과 비교하였을 때 오차를 효과적으로 감소시킨다고 알려져 있다(Jung et al., 2010).

여러 앙상블 모델 중 대표적으로 사용되는 랜덤포레스트(Random Forest; RF) 알고리즘은 Breiman(2001)에 의해 제안되었으며 의사결정 나무의 단점을 보완하기 위해 배깅의 기본 원리에 임의성을 더한 앙상블 알고리즘이다(Yoo, 2015). 랜덤포레스트의 기본 개념은 <Fig. 2>와 같이 전체 데이터 집합으로부터 N개의 부분집합 $\Theta_{1,\ldots,}\Theta_N$을 생성하며 독립이면서 동일한 분포(Independent and Identically Distributed; i.i.d.)를 따른다고 가정한다. 이후 무작위로 m개의 독립변수를 뽑고 $h(x,\Theta_N)$로 정의된 분류기를 이용해 나무를 성장시킨다. 이를 식 (11)에 제시하였으며 x는 입력 데이터를 의미한다. 생성된 N개의 나무를 이용해 목적변수가 연속형인 경우는 평균값을 계산하고 명목형인 경우 다수의 클래스로 투표(Voting)하는 과정을 거쳐 결과 값을 출력한다.



<Fig. 2> Process of random forest analysis

$$\{h(x,\Theta_N),\ N=1,\ldots\}\ where\ \Theta_N \sim i.i.d. \quad \cdots\cdots (11)$$

생성된 N개의 분류기를 통해 마진(Margin)을 계산하며 마진은 식 (12)을 통해 구할 수 있다. 여기서 $I(\bullet)$은 표시 함수(Indicator function)를 가리키며 분류기를 거친 X의 클래스가 Y와 같은 경우 1, 아닌 경우는 0으

로 처리한다. 분류문제에서 마진의 값이 클수록 큰 신뢰를 주는 척도로 작용한다(Breiman, 2001). 랜덤포레스트는 하나의 나무가 아닌 여러 개의 나무로 확장시키기 때문에 과적합(Overfitting)을 방지할 수 있으며 설명변수 개수가 많은 고차원 자료에 적용하더라도 예측력이 높다는 장점이 있다. 본 연구에서는 배깅 계열의 랜덤포레스트와 부스팅 계열의 그래디언트 부스팅(Gradient boosting; GB)을 상해 심각도 예측에 활용하였다.

$$mg(X,Y) = av_k I(h_N(X) = Y) - \max_{j \neq Y} av_k I(h_k(X) = j) \quad \cdots\cdots (12)$$

Ⅲ. 데이터 수집 및 데이터 집합 구성

1. 데이터 수집

본 연구에서는 도로교통공단과 교통안전정보관리시스템에서 제공받은 2015년부터 2017년 사이에 발생한 669,287건의 전국 교통사고 데이터를 이용하였다. 두 기관에서 받은 데이터는 동일한 사건의 교통사고를 다루고 있으나 기관별로 취급하는 속성이 다르다. 도로교통공단의 경우 34개의 속성을 취급하고 있었으며 교통안전정보관리시스템은 19개의 속성을 다루고 있다. 모델 학습에 있어 다양한 변수를 고려하기 위해 동일한 교통사고를 기준으로 통합하는 과정을 거쳐 53개의 변수를 가진 원시 데이터 집합을 구성하였다.

2. 변수 선정 및 데이터 전처리

변수 선정 과정에서 동일한 의미를 가지는 중복 변수는 하나로 축소하고 의미가 불분명한 사고번호(ACC_NO), 위도(latitude)와 경도(longtitude) 등과 같이 모델 분석과 무관하다고 판단되는 변수는 제거하였다. 가해자연령(Attacker age)과 피해자연령(Victim age) 변수는 10대 단위로 구간을 나누어 처리하고, 기존의 발생일(Date) 변수는 '연/월/일'과 같이 단일 변수로 구성되어 있어 이를 '연도(Year)', '월(Month)', '일(Day)'로 분리했다. 사고유형_대분류(Acc_type)는 '차대차', '차대사람', '차량단독', '철길건널목'이 있었으나 본 연구에서는 '차대차'의 경우만 고려하여 데이터를 추출하였다. 그 이외에 '불명/알수없음' 등으로 표시된 데이터와 값이 없는(Null) 경우는 모두 이상치로 판단해 제거하였으며, 사고내용(ACC_contents)과 가해자신체상해정도(Attacker_injury), 사망자수(num_death), 부상자수(num_injury)와 같은 변수의 경우 사고 심각도 수준과 대동소이한 변수로서 모델에 대한 객관적 평가를 왜곡시킬 가능성이 있다고 판단해 제거하였다. 종속변수인 피해자신체상해정도(Victim injury)는 '상해없음(No injury)', '부상신고(Report an injury)', '경상(Minor injury)', '중상(Serious injury)', '사망(Death)', '알수없음(No info)'으로 구분되어 있었으나 '알수없음'을 제외한 5가지 상해정도만 고려하였다. 최종적으로 22개 변수와 368,681개의 데이터 집합을 구성하였다. <Table 1>을 통해 본 연구에서 사용하는 데이터 집합의 변수를 확인할 수 있으며, <Table 2>에는 상해 심각도 별 관측치의 개수와 비율을 나타냈다.

<Table 2>에서 확인할 수 있듯이 '사망'의 비율은 전체 데이터 중 0.27%에 불과하며 한 클래스의 데이터 수가 매우 작은 경우 데이터 불균형(Data imbalanced) 문제를 발생시킨다. 데이터 불균형 문제는 기계학습 알고리즘의 성능을 저하시키는 주된 원인 중 하나이다. 예를 들어, 2개의 클래스 분류 문제 중 하나의 클래스는 전체의 99%를 차지하고 있고 나머지 클래스는 1%를 차지한다고 가정해보자. 모델의 단순 성능만을 고려할 경우 대부분의 데이터를 다수의 클래스로 분류하는 모델이 생성된다. 그 결과로 99%에 가까운 적중률을

보이며 이는 좋은 모델로 생각될 수 있지만 소수의 데이터에 중요한 정보가 포함되어 있는 경우 해당 정보를 고려하지 않는 문제가 발생하게 된다(Kang and Cho, 2006). 본 연구는 심각도가 높은 '사망'과 '중상'을 예측하는 것이 중요하다고 판단하여 상해 심각도별 비율을 맞춰주기 위해 '사망'은 993개 모두를 사용하고 '상해없음', '부상신고', '경상', '중상'에 해당하는 관측치를 각각 993개로 과소표본추출(Undersampling)하여 클래스에 속한 자료의 비율을 동일하게 맞춰주었다.

<Table 1> Data and their description

Variable	Description
Year	2015 to 2017 (every year)
Month	1 to 12 (every month)
Day	1 to 31 (every day)
Time	0 to 24 (every 2 hours)
City	City name(seoul/incheon/gyeonggi/daegu/busan/gwangu/jeonllanam-do etc.)
Local_name	Local government name(gangnam/bukgu/)
Acc_type	Accident type(Head-on collision/Collision_parking/Collison_driving etc.)
Violation	Content of violation (overspeed/signal_violation/center_line etc..)
Weather	sunny/rain/cloud/fog/snow
Road_type1	Single lane road/Intersection road
Road_type2	on_bridge/on_intersection/in_tunnel/on_crosswalk etc..
Attacker_driving	Duration of holding driver's license(under_5/under_10/under_15/over_15)
Attacker_vehicle	Vehicle types(passenger car/truck/special vehicle/van/motorcycle etc..)
Attacker_gender	Male / Female
Attacker_age	0 to 90 (every 10 ages)
Attacker_drunk	Drunk driving(Yes or No)
Victim_driving	Duration of holding driver's license(under_5/under_10/under_15/over_15)
Victim_vehicle	Vehicle types
Victim_gender	Male / Female
Victim_age	0 to 90 (every 10 ages)
Victim_drunk	Drunk driving(Yes or No)
Victim_injury	Injury level(No injury/Report an injury/Minor injury/Serious injury/Death)

<Table 2> The number of data and ratio by injury severity

Injury Severity	# of data	ratio
No injury	52,660	14.28%
Report an injury	9,044	2.45%
Minor injury	237,081	64.31%
Serious injury	68,903	18.96%
Death	993	0.27%
Total	368,681	100%

Ⅳ. 상해 심각도 예측 모델 구축

1. 모델 성능 평가

상해 심각도를 예측하기 위해 2장에서 소개한 6가지 데이터 마이닝 기법을 활용하여 모델의 성능을 비교하였다. 본 연구에서는 모델의 성능 평가를 위해 혼동 행렬(Confusion matrix)을 이용한 민감도(Sensitivity)와 특이도(Specificity)를 계산하였다. 혼동 행렬 구조는 <Table 3>에 제시되어 있다.

<Table 3> Concept of confusion matrix

		True Condition	
		True	False
Predicted Condition	True	True Positive(TP)	False Positive(FP)
	False	False Negative(FN)	True Negative(TN)

민감도는 실제 값이 참(True)인 경우에 대해서 참이라고 예측한 경우의 비율을 나타내며 특이도는 실제 값이 거짓(False)인 경우에 대해 거짓으로 예측한 경우의 비율을 의미한다. 본 연구에서는 범주의 중요도가 다르다는 점에 집중하여 모델성능의 비교분석 시 민감도를 기준으로 평가하였다. 민감도는 식 (13), 특이도는 식 (14)을 통해 계산된다.

$$Sensitivity = \frac{TP}{TP+FN} \quad \cdots\cdots (13)$$

$$Specificity = \frac{TN}{TN+FP} \quad \cdots\cdots (14)$$

예측 모델을 학습시키기 위해 난수를 발생시켜 7대 3의 비율로 학습용 데이터와 평가용 데이터로 나누고 각각을 이용해서 모델을 학습하고 각 상해 심각도 수준을 예측하였다. 객관적이고 평균적인 성능 비교 및 평가를 위해 시드(Seed) 값을 달리하여 10개의 서로 다른 데이터 집합을 모델 학습에 사용하였으며 학습 시 표본 추출로 야기될 수 있는 편의(Bias)를 최소화하고 모델의 신뢰성을 높이기 위해 5-겹 교차 검증(5-Fold Cross Validation) 방식을 적용하였다.

<Table 4>는 2장에서 설명한 데이터마이닝 기법을 이용하여 계산된 5가지 상해수준에 대한 모델 별 민감도와 특이도의 평균을 제시하였으며 소괄호 안에 값은 표준편차를 의미한다. 각 모델의 경우 서로 다른 매개변수를 가지고 있으며 모델의 성능은 매개변수를 어떻게 조정하느냐에 따라 달라진다. 본 연구에서는 각 모델이 상해 심각도를 예측하는데 최적의 성능을 낼 수 있도록 최적 매개변수 탐색(Hyperparameter tuning) 과정을 거쳐 선정된 매개변수를 적용하고, 이를 통해 모델별 성과를 평가할 수 있었다.

<Table 4>은 모델별 민감도와 특이도의 결과를 보여주며, 6가지 모델 중 전체적인 성능은 로지스틱 회귀가 0.445로 가장 높고, 다음으로 랜덤포레스트, 나이브베이즈, 그레디언트 부스팅, 의사결정나무, K-최근접 이웃 순으로 나타났다. 상해 심각도가 큰 '사망'의 경우 랜덤포레스트 모델이 0.732, 중상의 경우 그래디언트 부스팅 모델이 0.286으로 다른 모델과 비교하였을 때 높은 민감도를 보여주었으나 각 상해수준에 대해 전체적으로 낮은 수치를 보이고 있다. 중상, 경상, 부상신고의 경우 민감도가 50%도 되지 않는데 이는 상해정도

간 구분이 모호하기 때문으로 생각되며 5가지 상해 심각도를 예측하기에는 무리가 있다고 판단하였다. 실제 교통사고로 인한 응급상황 발생 시 상해 심각도를 세분화하여 낮은 정확도로 예측하는 것보다 소수의 범주를 높은 정확도로 예측하고 그에 따른 신속한 조치를 취하는 것이 바람직하다. 이에, 기존의 5가지 상해심각도 수준을 3가지로 분류하기 위해 기존의 데이터 집합을 재구성하고 보다 현실에 적용 가능한 상해 심각도 예측 모델을 구축하고자 하였다.

<Table 4> Performance comparison for test models

Model / Injury severity	KNN		LR		NB		DT		RF		GB	
	*sen.	+spec.	sen.	spec.	sen.	spec.	sen.	spec.	sen.	spec.	sen.	spec.
No injury	0.468 (0.052)	0.894 (0.018)	0.488 (0.029)	0.936 (0.011)	0.507 (0.031)	0.907 (0.008)	0.529 (0.051)	0.908 (0.028)	0.529 (0.027)	0.885 (0.014)	0.543 (0.023)	0.895 (0.004)
Report an injury	0.399 (0.049)	0.761 (0.032)	0.425 (0.076)	0.827 (0.037)	0.477 (0.031)	0.795 (0.014)	0.309 (0.09)	0.863 (0.046)	0.428 (0.028)	0.831 (0.019)	0.374 (0.032)	0.840 (0.014)
Minor injury	0.335 (0.042)	0.802 (0.033)	0.438 (0.091)	0.763 (0.058)	0.337 (0.029)	0.84 (0.019)	0.404 (0.053)	0.789 (0.039)	0.317 (0.047)	0.848 (0.017)	0.326 (0.022)	0.824 (0.016)
Serious injury	0.212 (0.029)	0.881 (0.023)	0.164 (0.056)	0.92 (0.022)	0.261 (0.027)	0.862 (0.008)	0.173 (0.033)	0.911 (0.021)	0.197 (0.015)	0.892 (0.014)	0.286 (0.03)	0.839 (0.009)
Death	0.528 (0.029)	0.908 (0.014)	0.711 (0.036)	0.86 (0.008)	0.606 (0.026)	0.895 (0.013)	0.729 (0.034)	0.816 (0.026)	0.732 (0.025)	0.845 (0.007)	0.638 (0.022)	0.893 (0.01)
Average sensitivity	0.388	0.849	0.445	0.861	0.438	0.86	0.429	0.857	0.441	0.86	0.433	0.858

*sen.: sensitivity; +spec.: specificity

2. 문제 재정의

앞서 기술한 바와 같이, 당초 5개의 상해 심각도 수준을 3개로 재구성했다. 즉, '상해없음'과 '부상신고'를 '경미한 상해(Minor injury)'로, '경상'과 '중상'은 '중간 상해(Intermediate injury)', '사망'은 '심각한상해(Serious injury)'로 분류하여 세 가지 심각도 수준에 대한 분류 문제로 재정의 하고 이에 의거하여 데이터를 가공하였다. 수정된 데이터의 상해 심각도별 관측치 개수와 비율은 <Table 5>를 통해 확인할 수 있다. 재정의 된 문제에 대한 모델 구축과정은 5가지 상해심각도 예측 모델을 구축하는 것과 동일하게 진행하였으며 모델의 안정적인 성능을 위해 시드를 바꿔가며 반복횟수를 100회로 늘려서 진행하였다. <Table 6>은 그 결과로서, 각 모델별 민감도와 특이도를 확인할 수 있다.

<Table 5> The number of data and ratio by reclassified injury severity level

Injury severity	# of data	ratio
Minor injury	61,704	16.74%
Intermediate injury	305,984	82.99%
Serious Injury	993	0.27%
Total	368,681	100%

<Table 6> Performance comparison for test models

Model / Injury severity	KNN		LR		NB		DT		RF		GB	
	*sen.	+spec.	sen.	spec.	sen.	spec.	sen.	spec.	sen.	spec.	sen.	spec.
Minor injury	0.561 (0.045)	0.807 (0.03)	0.488 (0.047)	0.924 (0.02)	0.60 (0.035)	0.836 (0.015)	0.412 (0.071)	0.851 (0.035)	0.586 (0.036)	0.846 (0.018)	0.611 (0.035)	0.825 (0.017)
Intermediate injury	0.583 (0.043)	0.722 (0.027)	0.687 (0.045)	0.702 (0.027)	0.570 (0.03)	0.747 (0.021)	0.665 (0.05)	0.765 (0.033)	0.551 (0.034)	0.785 (0.02)	0.539 (0.033)	0.773 (0.018)
Serious Injury	0.666 (0.025)	0.836 (0.016)	0.788 (0.026)	0.855 (0.018)	0.726 (0.028)	0.864 (0.015)	0.799 (0.036)	0.836 (0.023)	0.794 (0.025)	0.835 (0.017)	0.773 (0.025)	0.835 (0.017)
Average sensitivity	0.603	0.788	0.655	0.827	0.632	0.816	0.625	0.817	0.644	0.822	0.641	0.811

*sen.: sensitivity; +spec.: specificity

<Table 6>에 제시된 결과를 보면 전체적인 성능 측면에서 로지스틱 회귀분석의 민감도가 0.655로 가장 우수하고, 개별 상해 수준에 대해서는 의사결정나무 모델이 '심각한 상해'를 80%에 가까운 민감도로 예측하여 가장 나은 성능을 보였다. 그리고 '중간 상해' 예측에는 로지스틱 회귀분석이, '경미한 상해'에는 그래디언트 부스팅 방법이 가장 좋은 성능을 보여주었다. 이는 기존의 5가지 상해 수준을 이용하여 예측한 것보다 전체적으로 크게 향상된 것으로 앞서 소개한 상해수준 통합과정이 각 상해별 경계에 위치한 데이터를 효과적으로 분리하고 결과적으로 모델의 예측 성능을 향상시킬 수 있음을 확인하였다.

학습에 영향을 미치는 주요 변수를 알아보기 위해 LR, DT, RF 모델을 이용하여 상위 10개의 주요 변수를 <Table 7>에 나타내었다. 피해차량이 승용차(victim_vehicle_3)인 경우 모델에 상관없이 가장 큰 영향을 주는 것으로 나타났고, 피해차량 또는 가해차량이 이륜(vehicle_6)일 경우에도 상해 심각도에 큰 영향을 주는 것으로 나타났다. 모델에 무관하게 전반적으로 차량특성과 관련된 변수들이 교통사고 상해 심각도에 주요한 영향을 미치며, 이외에도 중앙선침범(violation_8), 피해자나이(victim_age_70)와 같은 도로 이용자 특성 변수와 지역특성에 해당하는 서울(city_9) 등이 상해심각도 예측모델을 학습하는데 중요한 역할을 함을 알 수 있다. 전체적으로, 상위 주요 변수들의 경우 모델과 무관한 영향력을 확인할 수 있으나 일부 변수의 경우 모델별로 그 영향력이 엇갈려, 상이한 알고리즘 특성이 변수 중요도에 영향을 준 것으로 보인다.

<Table 7> Top 10 variables for Injury severity prediction models

Var imp	LR	DT	RF
1	victim_vehicle_3	victim_vehicle_3	victim_vehicle_3
2	victim_vehicle_6	victim_vehicle_4	victim_vehicle_6
3	attacker_vehicle_6	victim_vehicle_9	attacker_vehicle_6
4	victim_vehicle_4	attacker_vehicle_6	attacker_vehicle_9
5	attacker_vehicle_3	victim_vehicle_6	victim_age_70
6	attacker_vehicle_9	attacker_vehicle_5	attacker_vehicle_3
7	victim_age_70	viloation_8	victim_vehicle_4
8	victim_vehicle_5	victim_vehicle_5	victim_vehicle_5
9	violation_8	attacker_vehicle_9	attacker_vehicle_5
10	attacker_vehicle_5	violation_1	city_9

<Table 6>에 제시된, 각 모델 별 성능의 차이를 확인하기 위해 분산분석(Analysis of variance; ANOVA)을 진행하였다. 분산분석의 경우 세 개 이상의 집단 평균을 비교하기 위해 사용되는 통계적 기법으로 회귀분석에서 회귀계수의 유의성을 검정하거나 모델 간 성능 차이가 존재하는지 확인하기 위해 사용된다. 따라서 분산분석을 진행하기 위해 집단의 평균이 동일하다는 가설을 세우고 유의수준과 유의확률(probability value; p-value)을 비교하여 집단 간 평균의 차이 여부를 판단한다. 본 연구에서는 가설검정 시 유의수준을 0.05로 설정하였다.

경미한 상해에 대해 분산분석 결과를 <Table 8>에 정리하였다. 이 결과를 보면, p-value가 0에 가까워 유의수준 5%하에서 모델 별 경미한 상해 예측에 대한 성능 차이가 있다는 결론을 얻을 수 있다. 추가적으로, 각 모델별 상세한 비교를 위해 터키(Tukey) 방법을 이용하였다. <Table 9>는 그 결과로서, 성능 비교의 기준이 되는 model(i)와 비교 대상 model(j)의 평균의 차이(diff), 95% 신뢰구간에 대한 상한(upper)과 하한(lower), 유의확률(p-value)을 보여준다. 이 결과에서도 유의확률은 모두 0에 가까워 모델 간 성능차이와 우위 관계를 확인할 수 있다. 즉, GB 모델이 경미한 상해를 가장 잘 예측하고 RF 모델이 다음으로 좋은 성능을 보여주었다.

<Table 8> ANOVA table for the case of minor injury

Source of variation	SS	df	MS	F	p-value
Treatment	3.016	5	0.6031	280	2e-16
Residual	1.280	594	0.0022		
Total	4.296	599			

· SS : Total sum of square
· df : degree of freedom
· MS : Mean square
· F : F statistic

<Table 9> Tukey's honestly significant difference test for the case of minor injury

model(i)	model(j)	diff	95% confidence interval		p-value
			lower	upper	
KNN	LR	0.072	0.054	0.091	4.79E-10
	NB	-0.039	-0.058	-0.020	3.00E-01
	DT	0.148	0.13	0.167	6.08E-08
	RF	-0.026	-0.044	-0.007	4.79E-10
	GB	-0.051	-0.069	-0.032	0.503
LR	NB	-0.112	-0.130	-0.093	4.79E-10
	DT	0.076	0.057	0.095	4.79E-10
	RF	-0.098	-0.117	-0.079	4.79E-10
	GB	-0.123	-0.142	-0.104	4.79E-10
NB	DT	0.188	0.169	0.206	1.53E-3
	RF	0.014	-0.005	0.032	4.79E-10
	GB	-0.011	-0.030	0.007	0.002
DT	RF	-0.174	-0.193	-0.155	4.79E-10
	GB	-0.199	-0.218	-0.18	4.8E-10
RF	GB	-0.025	-0.044	-0.006	4.79E-10

<Table 10>는 중간 상해에 대한 분산분석 결과로서 모델 간 성능의 차이가 있다는 결론을 유의확률을 통해 얻을 수 있다. <Table 11>은 사후분석을 진행하여 얻은 결과로 중간상해를 가장 잘 예측하는 모델을 확인한 결과 LR모델이 가장 우수하였으며 DT모델이 다음으로 우수한 성능을 보여주었다.

<Table 10> Results of ANOVA table for the case of Intermediate injury

Source of variation	SS	df	MS	F	p-value
Treatment	1.9058	5	0.3812	241.8	2e-16
Residual	0.9363	594	0.0016		
Total	2.8421	599			

· SS : Total sum of square
· df : degree of freedom
· MS : Mean square
· F : F statistic

<Table 11> Tukey's honestly significant difference test for the case of Intermediate injury

model(i)	model(j)	diff	95% confidence interval		p-value
			lower	upper	
KNN	LR	-0.104	-0.120	-0.088	4.79E-10
	NB	0.013	-0.003	0.029	1.63E-01
	DT	-0.082	-0.098	-0.065	4.79E-10
	RF	0.032	0.016	0.048	2.98E-07
	GB	0.044	0.028	0.060	4.80E-10
LR	NB	0.118	0.102	0.134	4.79E-10
	DT	0.023	0.007	0.039	8.07E-04
	RF	0.136	0.120	0.152	4.79E-10
	GB	0.148	0.132	0.164	4.79E-10
NB	DT	-0.095	-0.111	-0.079	4.79E-10
	RF	0.019	0.003	0.035	1.28E-02
	GB	0.030	0.014	0.046	1.62E-06
DT	RF	0.113	0.097	0.130	4.79E-10
	GB	0.125	0.109	0.141	4.79E-10
RF	GB	0.012	-0.004	0.028	3.04E-01

<Table 12>는 심각한 상해에 대한 분산분석 결과로, 이 경우에도 모델 간 성능 차이가 있음을 확인할 수 있다. 또한, <Table 13>의 사후분석 결과를 보면 심각한 상해의 경우 LR모델과 DT모델의 차이는 0.01로 매우 작고 유의확률이 0.082로 유의수준 0.05 하에서 통계적으로는 성능의 차이가 있다고 할 수 없다. 하지만 그 값의 차이 0.01과 <Table 6>의 성능차이를 고려하였을 때 심각한 상해를 예측하는데 DT모델이 LR모델 보다 다소 우위에 있다고 판단된다.

<Table 12> ANOVA table for the case of serious injury

Source of variation	SS	df	MS	F	p-value
Treatment	1.3521	5	0.27042	351.6	2e-16
Residual	0.4568	594	0.00077		
Total	0.3442	599			

· SS : Total sum of square
· df : degree of freedom
· MS : Mean square
· F : F statistic

<Table 13> Tukey's honestly significant difference test for models of serious injury

model(i)	model(j)	diff	95% confidence interval		p-value
			lower	upper	
KNN	LR	-0.122	-0.133	-0.111	4.79E-10
	NB	-0.060	-0.071	-0.049	4.79E-10
	DT	-0.132	-0.144	-0.121	4.79E-10
	RF	-0.127	-0.139	-0.116	4.79E-10
	GB	-0.107	-0.119	-0.096	4.79E-10
LR	NB	0.062	0.051	0.073	4.79E-10
	DT	-0.010	-0.022	0.001	0.082
	RF	-0.005	-0.017	0.006	0.740
	GB	0.015	0.003	0.026	0.003
NB	DT	-0.073	-0.084	-0.061	4.79E-10
	RF	-0.068	-0.079	-0.056	4.79E-10
	GB	-0.048	-0.059	-0.036	4.79E-10
DT	RF	0.005	-0.006	0.016	0.789
	GB	0.025	0.014	0.036	5.06E-09
RF	GB	0.020	0.009	0.031	6.37E-06

Ⅴ. 결 론

본 연구에서는 최근 공공데이터 개방 정책의 일환으로 확보된 도로교통공단과 교통안전정보관리시스템의 방대한 자료를 활용하여 국내 교통사고, 그 중 차량과 차량 간의 사고 데이터를 이용해 상해 심각도를 예측할 수 있는 방법을 제시하였다. 각 상해 심각도 수준 간의 데이터 수에 차이가 있음에 주목하여 표본수가 많은 그룹에 대해서는 과소표본추출(Under-sampling)을 시행한 가운데, 교통사고 현장에서의 활용도를 제고하기 위해 5가지 상해심각도 수준으로 분류된 데이터를 3가지 분류 체계로 변경하여 문제를 재정의한 결과, 예측 정확도의 뚜렷한 향상을 확인할 수 있었다.

나아가, 세 가지 상해 심각도에 대한 분산분석과 사후분석을 진행한 결과, GB 모델이 경미한 상해를 가장 잘 예측하였으며 중간상해는 LR 모델이, 심각한 상해는 DT 모델이 우수한 성능을 보여주었다. 본 연구에서는 중증상해 예측에 초점을 맞추어 진행하였고, 여러 연구결과와 상해 심각도의 중요도를 고려할 때 6가지 예측모델 중 DT 모델이 상해 심각도 예측에 가장 적합함을 알 수 있었다. 또한, 모델에 영향을 주는 주요 변

수를 파악한 결과 피해차량이 승용차인지 여부가 가장 큰 영향을 미쳤으며, 차량과 관련된 차량특성 변수뿐만 아니라 교통법규, 피해자 나이 등과 같은 관련된 운전자 특성, 지역 특성 변수들의 중요도를 확인할 수 있었다. 또한, 연구에 사용된 변수들의 특성상 변수들의 중요도 보다는 모델 별 알고리즘에 따라 심각도 예측 성능에 차이가 있음을 알 수 있었다.

본 연구는 국가 기관으로부터 제공받은 2015년부터 2017년 동안 발생한 약 67만 건의 전국 교통사고 데이터를 이용해 국내 교통사고 현실을 반영한 상해 심각도 예측 모델 개발을 시도했다는 점에서 의미를 가진다. 그럼에도 몇 가지 한계점이 있는데, 첫째, 상해 심각도 간의 구분이 모호해 다섯 가지 상해 심각도 분류 문제를 세 가지로 바꾸어 진행하였기 때문에 보다 세심한 분류 예측을 못했다는 점이다. 둘째, '심각한 상해'에 대해서는 79.9%라는 높은 민감도를 보였으나, '경미한 상해', '중간 상해' 클래스의 민감도는 60%대로 다소 낮은 성능을 보여 실제 교통사고 현장에 활용하는데 다소 아쉬움이 있다.

향후 연구 과제로 차량의 전복여부, 차량 속도, 안전벨트 착용여부 등과 같은 상해 심각도를 구분 짓는 데에 있어 결정적인 역할을 할 수 있는 변수들이 추가적으로 수집된다면 전체 민감도를 향상시키는 모델 개발이 가능할 것으로 기대된다. 또한 각 지역의 교통 및 지리적인 특성과 관련된 추가적 자료를 이용할 경우 지역별 상해 심각도 예측 모델 개발이 가능할 것이다.

ACKNOWLEDGEMENTS

본 논문은 2018년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2018S1A5A8026857). 또한 본 논문의 일부는 2019년도 전북대학교 연구기반 조성비 지원에 의하여 연구되었음.

REFERENCES

Breiman L.(2001), "Random forest," *Machine Learning*, vol. 45, pp.5-32.

Breiman L., Friedman J. H., Olshen R. A. and Stone C. G.(1984), *Classification and Regression Trees*, Chapman & Hall, pp.3-4.

Cover T. M. and Hart P.(1967), "The nearest neighbor decision rule," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp.21-27.

Dietterich T. G.(1997), "Machine learning research: four current directions," *AI Magazine*, vol. 18, no. 4, pp.97-136.

Gentle J. E. and Hadle W.(2012), *Handbook of Computational Statistics: Concepts and Methods*, pp.985-1022.

Hahn D. W., Park K. S. and Shin Y. K.(2002), "A Research on Regional Differences in Traffic environments and Driver's Behaviors in Korea," *The Korean Journal of Psychological Association*, vol. 8, no. 1, pp.17-40.

Hastie T., Tibshirani R. and Friedman J.(2009), *The Elements of Statistical Learning*, Springer, pp.307-310.

Hong S. E., Lee G. Y. and Kim H. J.(2015), "A Study on Traffic Accident Injury severity Prediction Model Based on Public Data," *Journal of Advanced Information Technology and Convergence*,

vol. 13, no. 5, pp.109-118.

Isaac J. and Harikumar S.(2016), "Logistic regression within DBMS," *2nd International Conference on Contemporary Computing and Informatics (IC3I)*, pp.661-666.

Jeong H. J., Jang Y. C., Bowman P. J. and Masoud N.(2018), "Classification or motor vehicle crash injury severity: A hybrid approach for imbalanced data," *Accident Analysis and Prevention*, vol. 120, pp.250-261.

Jeong H. R., Kim H. H., Park S. M., Han E., Kim K. H. and Yun I. S.(2017), "Prediction of Severities of Rental Car Traffic Accidents using Naive Bayes Big Data Classifier," *The Journal of The Korea Institute of Intelligent Transport System*, vol. 16, no. 4, pp.1-12.

Jung Y. H., Eo S. H., Moon H. S. and Cho H. J.(2010), "A Study for Improving the Performance of Data Mining Using Ensemble Techniques," *Communications for Statistical Applications and Methods*, vol. 17, no. 4, pp.561-574.

Kang P. and Cho S.(2006), "EUS SVMs: Ensemble of Under sampled SVMs for Data Imbalance Problems," *Lecture Notes in Computer Science*, vol. 4232, pp.837-846.

Kass G.(1980), "An exploratory technique for investigating large quantities of categorical data," *Applied Statistics*, vol. 29 no. 2, pp.119-127.

Korea Road and Traffic Authority(2014), *Estimation of Traffic Accident Costs by region*.

Korea Road and Traffic Authority(2018), *Estimation and Evaluation of Traffic Accident Costs*.

Korea Road and Traffic Authority(2019), *Comparison of Traffic Accident of OECD Members States*.

Lee J. S. and Heo G.(2011), "Injury Severity Prediction of Traffic Accident using Data Mining," *Proceedings of the 2011 Fall Conference of Korean Intelligent Information Systems Society*, pp.199-206.

Lee J. S. and Lee E. J.(2009), "Analysis of Traffic Accidents using Decision Tree Ensemble Model," *Proceedings of the 2009 Fall Conference of Korean Intelligent Information Systems Society*, pp.211-218.

Lee J. Y. and Lee Y. J.(2018), "Exploration of the Factors Determining the Lecture Education of Liberal Arts Courses Utilizing the Decision Tree Analysis," *Korean Journal of General Education*, vol. 12, no. 6, pp.67-93.

Quinlan J. R.(1993), *C4.5 : Programs for machine learning*, Morgan Kaufmann, San Mateo.

Sohn S. Y. and Shin H. W.(1998), "Data Mining for Road Traffic Accident Type Classification," *Journal of the Korean Institute of Industrial Engineers*, pp.542-549.

Uddin M. and Huynh N.(2020), "Injury severity analysis of truck-involved crashes under different weather conditions," *Accident Analysis and Prevention*, vol. 141.

Yoo J. E.(2015), "Random forests, an alternative data mining technique to decision tree," *Journal of Educational Evaluation*, vol. 28, no. 2, pp.427-448.