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

Journal of the Society of Disaster Information
Vol. 17, No. 3, pp. 427-441, September 2021
https://doi.org/10.15683/kosdi.2021.9.30.427

Original Article

# 안전사고 예측모형 개발 방안에 관한 연구(군 교통사고 사례를 중심으로)

# A Study of Safety Accident Prediction Model (Focusing on Military Traffic Accident Cases)

기재석[1] · 홍명기[2]*
Jae-Sug Ki[1], Myeong-Gi Hong[2]*

[1]Professor, Department of Sports ICT Convergence, Sangmyung University. Seoul, Republic of Korea
[2]PhD, Department of Sports ICT Convergence, Sangmyung University. Seoul, Republic of Korea

*Corresponding author: Myeong-Gi Hong, hong8623@nate.com

## ABSTRACT

**Purpose:** This study proposes a method for developing a model that predicts the probability of traffic accidents in advance to prevent the most frequent traffic accidents in the military. **Method:** For this purpose, CRISP-DM (Cross Industry Standard Process for Data Mining) was applied in this study. The CRISP-DM process consists of 6 stages, and each stage is not unidirectional like the Waterfall Model, but improves the level of completeness through feedback between stages. **Results:** As a result of modeling the same data set as the previously constructed accident investigation data for the entire group, when the classification criterion was 0.5, Significant results were derived from the accuracy, specificity, sensitivity, and AUC of the model for predicting traffic accidents. **Conclusion:** In the process of designing the prediction model, it was confirmed that it was difficult to obtain a meaningful prediction value due to the lack of data. The methodology for designing a predictive model using the data set was proposed by reorganizing and expanding a data set capable of rational inference to solve the data shortage.

**Keywords:** Accident Cause Information, Predictive Model, Exploratory Factor Analysis, Machine Learning, Traffic Accident

## 요 약

**연구목적:** 본 연구는 군에서 가장 많이 발생하는 교통사고의 예방을 위해 부대별로 교통사고가 발생할 확률을 사전에 예측하는 모형의 개발 방안을 제시하는 것이다. **연구방법:** 이를 위해 CRISP-DM(Cross Industry Standard Process for Data Mining) 방법론을 적용하였다. CRISP-DM 프로세스는 6단계로 구성되어 있고, 각 단계는 Waterfall Model처럼 일방향으로 구성되어 있지 않고 단계 간 피드백을 통하여 단계별 완성도를 높이게 되어 있다. **연구결과:** 전체 집단을 대상으로 기 구축된 사고조사 데이터와 동일한 데이터 세트(data set)를 구축하여 모델링한 결과 분류기준 0.5로 했을 때, 교통사고예측을 위한 모형의 정확도, 특이도, 민감도, AUC에서 의미있는 결과치를 도출하였다. **결론:** 예측모형을 설계하는 과정에서 데이터의 부족으로 인해 의미 있는 예측값을 얻기 어려운 문제점이 확인되었다. 이를 해결하기 위해 합리적 추론이 가능한 데이터 세트(data set)를 재구성 및 확대하여 데이터 부족을 해소하고, 이를 활용한 예측모형을 설계할 수 있는 방법론을 제시하였다.

**핵심용어:** 사고요인정보, 예측모형, 탐색적 요인분석, 머신러닝, 교통사고

Received | 3 May, 2021
Revised | 17 September, 2021
Accepted | 24 September, 2021

# 서 론

## 연구의 배경 및 목적

본 연구는 군에서 가장 많이 발생하는 교통사고의 예방을 위해 부대별로 교통사고가 발생할 확률을 사전에 예측하는 모형의 개발 방안에 대한 것이다. 과거 군에서 발생한 교통사고에 관련된 요인 데이터를 확보하여 예측모형을 설계하는 과정에서 데이터의 부족으로 인해 의미 있는 예측값을 얻기 어려운 문제점이 도출되었다. 그래서 이러한 문제점을 해결하기 위해 모형 개발에 충분한 데이터 세트(data set)을 어떻게 구축하고, 신뢰성 있는 예측모형을 어떻게 설계할 것인가에 대한 방안을 제시한 연구이다.

그동안 많은 분야에서 축적된 데이터를 활용하여 미래를 예측하는 연구는 꾸준히 진행되어왔다. 통계적 기법을 활용하여 요인 변수를 찾고 다양한 알고리즘을 적용하여 가장 합리적인 예측치를 찾아내는 방법부터, 최근에는 빅데이터를 활용하여 기계학습과 AI 알고리즘을 적용하여 보다 다양하고 세분화된 예측정보를 얻기 위한 많은 연구가 시도되고 있다. 미래에 일어날 일을 예측하는 것은 쉬운 일은 아니다. 그럼에 도 불구하고 과거에 발생한 일에 대한 조사기록을 근거로 사건의 패턴 등을 찾아내 미래의 사건을 예측하는 연구가 활발히 진행되고 있다. 사례로 인류가 "암" 질병을 정복하기 위해 요인이 되는 환자 발생 정보, 개인의 생체 및 유전자 정보, 생활환경 정보, 음식물 및 물질 정보 등을 활용하여 암의 발생 요인과 발생할 확률(예측치)를 연구하고 있다. 이와 같은 방식으로 사건 발생 정보, 개인 신체 및 성격 정보, 기인물 정보, 사고 발생 환경 정보 등을 활용하여 사고 발생 요인을 찾고 사고가 발생할 확률(예측치)를 구하는 모형을 연구하는 것이다. 사람을 대상으로 개별적인 예측이 제한되는 경우는 시설이나 물건을 대상으로 하거나, 특정 집단의 성향을 고려한 집단별, 지역별 사고 발생확률도 예측이 가능할 것이다.

## 연구의 범위 및 방법

사고의 발생을 예측하기 위해서는 어떠한 정보가 필요한지, 그리고 정보를 어디서 획득할 수 있는지, 획득된 정보를 분석하여 유용한 예측값을 도출하기 위해서는 어떠한 논리적, 물리적, 기술적 절차가 필요한지를 군의 사고유형과 관련하여 연구를 진행하였다. 현재의 군의 교통사고 예측모형을 구상하는 과정에서 기존 연구 내용과 군의 관련된 자료와 비교하여 문제점을 식별하고 개선 방향을 제시하였으며, 현 데이터 확보 수준에서 개발이 가능한 교통사고 예측 모형을 제시하였다.

금 번 교통사고 예측 모형을 개발하는 데 있어 가장 먼저 고려되어야 하는 사항은 수집 가능한 데이터의 양과 품질이다. 본 연구를 수행하기 위해 수집한 데이터는 전 군의 과거 사고 조사데이터와 일부 부대의 차량운행 기록과 관련된 데이터이다. 이는 각종 안전사고에 대한 사고를 조사하여 종합한 데이터로 EDA 분석 결과를 제시하였다. 연구에서는 데이터 가용성(품질, 양)과 모형 개발 가능성 및 결과의 합목적성을 고려하여 교통사고 위험예측에 대한 모형은 다음 방안으로 수행하였다. 교통사고를 발생시키는 요인(속성)에 대한 데이터 값 중 첫째, 사고조사데이터에 포함되어 있지 않은 데이터를(결측 속성) 식별하여 기 구축된 차량운행 데이터를 기준으로 결측 속성을 통계적 기법으로 생성하여 데이터 프레임을 구축하고, 둘째, 머신러닝 알고리즘을 적용한 교통사고 예측모형을 개발하는 것이다.

## 이론적 고찰

### 하인리히의 사고발생 이론[1)]

일상생활에서 도로를 걷다가 또는 차를 타고 다니면서 또는 사무실 안에서 수많은 크고 작은 사고를 경험하게 된다. 하지만 이과 같은 사고가 큰 사고의 잠재적 요인이라고 연관시키기 보다는 단순한 실수로 간과해 버린다. 이에 관하여 50,000건의 사고 통계를 분석한 하인리히(Heinrich HW)의 연구결과에 따르면, 상해가 없거나 극히 경미한 사고가 중상해를 합친 사고보다 10배나 더 많다는 것이다. 이를 '사고의 피라미드 모형'이라 하며 적어도 300번 이상의 불안전 행동을 반복하던 사람은 경상해를 입을 경우가 평균 29회, 중상해를 입을 경우가 1회 이상 발생할 수 있을 것이다. 사고 발생의 요인을 볼 때, 한 가지만의 요인은 거의 없으며, 일반적으로 여러 가지 요인이 복합적으로 작용하게 된다. 이를 크게 구분하여 인적 요인(연령, 성별, 태도 및 심리적 상태, 행동특성), 물적 요인(기계의 결함, 설비 부족), 환경적 요인(날씨, 도로환경, 작업장 환경)으로 구분할 수 있다.

그중 사고의 88%는 사람에 기인하고 나머지 10%가 불안전한 물적 요인에 의한 것이며, 불가항력으로 인한 환경적 요인에 의한 것은 단지 2%에 불과하다. 이에 따라 행위자의 사고 예방 교육을 통하여 사고 발생의 인적 요인인 불안전한 행동을 수정하면 88%의 사고는 예방될 수 있다.

### 재해원인 구조 이론[2)]

재해 원인은 직접 원인과 간접원인으로 구분되며, 간접원인은 기초원인과 2차 원인으로 나뉘어져 세 가지재해 원인이 존재한다. 직접 원인인 1차 원인의 하나인 인적 원인은 인간은 주의력이 부족한 유동적 존재로서 안전교육의 실시와 안전화된 기계설비의 물적 원인, 기타 불가항력적 원인에 대한 대처가 이루어지지 않게 되어 재해가 발생한다. 간접원인인 2차원인 중 교육적 원인은 안전에 관한 경험과 지식 부족으로 인하여 재해유발의 원인이 되고 있고, 간접원인인 기초원인 중 교육적 원인은 학교 등의 교육기관에서 안전교육이 제대로 이루어지지 않게 되면 재해를 발생시킬 수 있다는 이론이다(Fig. 1. 참조).

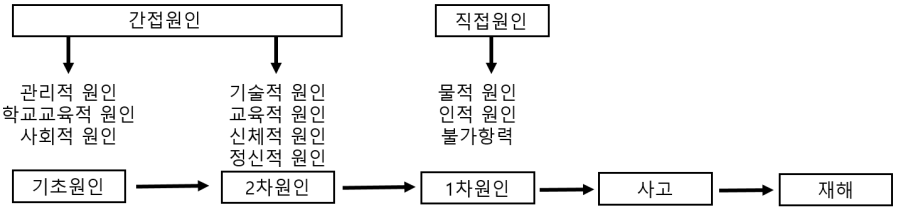



Fig. 1. Structural theory of causes of disaster

### 사고이론에 관한 고찰

사고발생 이론에 관한 고전 법칙인 하인리히 법칙에서 주목할 만한 것은 잠재적 상해와 경상해 및 중상해의 구별을 전제로 한 법칙이란 점이다. 잠재적 상해는 상해가 발생하지 않은 상황적 개념이기 때문에 위험 또는 사고가 현실화 된 경상해 및 중

1) 허버트 윌리엄 하인리히(Herbert William Heinrich, 1885년 10월 6일 ~ 1962년 6월 22일)는 1930년대부터 산업안전 연구, 1931년 "Industrial Accident Prevention, A Scientific Approach"이라는 제목의 책에서 주장한 이론

2) 최기봉(2001) "안전관리론 입문", 구미서관

상해와는 구별되는 개념이다.

결국 사고 이론에서 추론해 낼 수 있는 것은 위험을 잠재적 위험과 표출된 위험으로 구분하는 것이 이론상 가능하다는 점이다. 위험의 잠재적 위험과 표출된 위험의 구분은 안전 및 재난정책에 있어서 매우 중요한 역할을 한다. 다시 말해 잠재적 위험은 구조적·내재적 위험으로서 평상시에 관리하여야 하는 위험요인으로서 안전관리정책에 포함 시킬 수 있으나, 표출된 위험은 언제든지 재난으로 전이될 가능성이 농후하기 때문에 위기의식을 가지고 대응하여야 할 필요성이 제기된다.

위의 논리에 의한다면 사고 예방을 위해 인적 요인을 수정하기 위해서는 사람 또는 사람의 집단에 대해서 행동을 관찰하고 데이터를 실시간 수집하여 분석하는 시스템이 필요하다. 과거에는 이러한 시스템이 부재하여 사고가 일어난 사건에 대해서만 조사를 통해 자료 즉, 데이터를 수집하고 이를 분석하여 사고 발생을 예측할 수 밖에 없었으므로 사고의 예측을 통한 예방보다는 사후 사고 발생에 대한 통계와 요인 분석 수준에서 이루어졌다.

데이터의 처리기술 (데이터 마이닝 등)이 비약적으로 발전되었음에도 불구하고 현재는 사고에 대한 데이터 즉 미상해 데이터의 부족으로 상해가 예상되는 사고의 예측(대상, 시간, 장소, 요인물 등)에 대해 어려움을 느끼고 있거나 사고 발생 결과만을 가지고 예측하다 보니 정확도가 떨어지는 결과를 보고 있다. 하인리히의 법칙에 의한다면 사고예측을 위해 미상해 자료 수집 및 분석과 상해 데이터의 분석 및 비교를 통해 미래 사고를 예측해야 하는 것이 논리적인 방향일 것이다.

## CRISP-DM 연구 방법론

본 연구에서는 CRISP-DM(Cross Industry Standard Process for Data Mining)을 적용하였다.

CRISP-DM는 1996년 유럽연합의 ESPRIT에서 있었던 프로젝트에서 시작되었으며, CRISP-DM 프로세스는 6단계로 구성되어 있고, 각 단계는 Waterfall Model처럼 일방향으로 구성되어 있지 않고 단계 간 피드백을 통하여 단계별 완성도를 높이게 되어 있다. 각 단계는 아래와 같으며 Fig. 2를 참조한다.

1단계 : 업무 이해(Business Understanding) : 초기 프로젝트 계획을 수립하는 단계.

2단계 : 데이터 이해(Data Understanding) : 분석을 위한 데이터를 수집, 데이터 속성 이해, 데이터 품질 및 문제점 식별

3단계 : 데이터 준비(Data Preparation) : 데이터 세트(data set) 선택, 정제, 편성, 통합, 포맷팅 과정

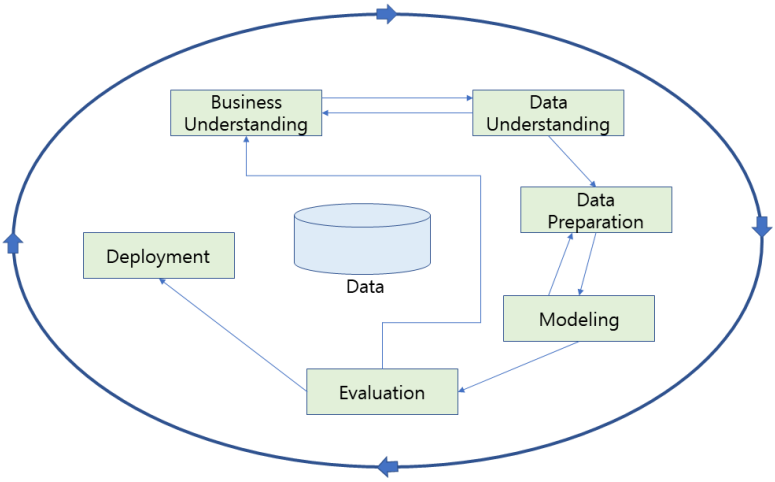



Fig. 2. CRISP-DM diagram

4단계 : 모델링(Modeling) : 다양한 모델링 기법과 알고리즘을 선택하고 모델링 과정에서 파라메타를 최적화하는 단계, 모델을 테스트용 프로세스와 데이터 세트(data set)로 평가

5단계 : 평가(Evaluation) : 데이터 마이닝 결과를 수용할 것인지 결정

6단계 : 전개(Deployment) : 완성된 모델을 실 업무에 적용하기 위한 계획 수립, 모델링과 모델의 유지보수 계획 수립

## CRISP-DM 방법론을 적용한 예측모형 개발

### 업무의 이해

CRISP-DM 방법론에서 1단계 업무 이해(Business Understanding) 단계이다. CRISP-DM 프로세스는 6단계로 중 에서 첫 단추로 생각되는 부분으로 첫 단추가 잘 정리되어야 결론에서 의도하는 결과를 얻을 수 있다. 그런 이유로 업무에 대한 이해가 중요하며 각 단계의 시행착오를 최소화 할 수 있다.

#### 교통사고의 요인 정리

교통사고의 요인(causes)은 충돌·손상·피해 등 사고의 직접적인 결과를 양산하지만 요인(factor)은 사고의 잠재적인 가능성이 있을 뿐, 반드시 사고의 결과를 발생시킨다고 말할 수는 없다. 따라서 교통사고의 발생 요인과 발생 가능성이 있는 요인과는 구분할 필요가 있다. 사고 다발 지역인 교차로의 기하 구조와 안전시설을 개선하여 교통사고의 발생 건수를 크게 감소시켰다고 가정하자. 이때 불량한 교차로의 기하 구조와 안전시설은 사고의 주요 요인이 될 수 있으나 이것이 교차로에서 발생하는 모든 사고의 요인이라고 단정할 수는 없다. 아마도 교차로의 기하 구조나 안전시설을 완전하게 개선하더라도 운전자의 부주의 등에 의해 여전히 사고는 발생하게 될 것이다. 마찬가지로 음주운전이나 졸음운전도 사고의 잠재적인 요인이 될 수 있으나 반드시 사고의 요인이 되는 것은 아니다. 따라서 교통사고의 요인은 하면 사고 조사과정에서 밝혀진 교통사고를 유발한 요인으로 해석하는 것이 타당하고, 교통사고를 예측 하는데는 교통사고의 요인되는 데이터를 변수로 사용하는 것이 타당하다. 교통사고는 사람과 차량, 도로환경의 3요소로 구성되기 때문에 교통사고의 요인도 인간 요인과 도로환경요인, 그리고 차량 요인이 개별적 또는 유기적으로 결합되어 발생하게 된다.

- 교통사고의 도로환경요인

도로의 설계 속도에 비해 커브의 곡선반경이 작으면 차량은 원심력에 의해 곡선부의 바깥쪽으로 도로를 이탈할 위험성이 높아지고, 동시에 운전자는 곡선반경이 작은 커브구간에서 속도를 줄이지 않고 도로를 가로질러 큰 곡선반경을 그리며 주행하려는 습성이 강해 중앙선 침범의 위험성도 높아지게 된다. 또한 도로상에 운전에 필요한 안전표지와 시설이 적절하게 설치되어 있지 않으면 운전자의 주의력이 저하되고 판단 착오의 비율이 높아져 결과적으로 사고의 잠재적 위험성을 증가시키게 된다. 이러한 도로환경요인은 크게 도로 구조적 설치 및 관리 하자와 교통안전시설의 설치 및 관리 하자로 구분할 수 있다. 도로 구조적 요소는 도로이용자가 쾌적하고 안전하게 이용할 수 있도록 도로 자체가 갖추어야 할 최소한의 기준으로 차도, 중앙분리대, 길 어깨(갓길), 주·정차대, 보도 등 도로의 횡단 구성 결함과 곡선반경, 편경사, 완화구간, 시거, 선형조합, 교차로 설계 등 도로의 선형요소 결함이 있다. 교통안전시설은 도로의 구조상태를 보완하여 사고에 의한 차량 및 운전자의 피해를 최소화시키고, 운전자에게 전방의 도로상태나 운전에 필요한 정보를 사전에 정확하게 전달하여 미리 주의환기 시키거나 적

절한 행동을 유도할 수 있도록 한 시설로써 방호 울타리, 시선 유도시설, 미끄럼방지시설, 과속방지시설, 조명시설, 도로반사경, 충격흡수시설, 안전표지, 신호기(신호설계) 등의 설치 및 관리하자가 있다. 본 연구에서는 도로상의 요인 데이터 분석은 생략하였으며 국토교통부에서 기 구축된 데이터를 향후 활용하는 것으로 하였다.

- 교통사고의 차량 요인

교통사고의 차량 요인은 전체 사고에 비해 극히 미미하지만 차량의 구조적 결함이나 정비 불량에 기인된 사고는 결과물인 충돌사고에 겹쳐지면서 묻혀 버리는 경우가 많고 대부분 승차자의 진술에 의존하여 조사가 이루어지기 때문에 실체적 요인을 규명 하는데는 어려움이 있다. 차량 요인에 있어 큰 비중을 차지하는 것은 타이어 및 브레이크의 결함, 정비 불량, 차량 화재 등이 있다.

- 교통사고의 인간 요인

교통사고에 있어 인간은 운전자 또는 탑승자 그리고 보행자 등으로 참여하게 된다. 운전자는 운행의 주체로서 상황을 인지하고 입력된 정보를 이해하고 판단하여 적절한 운전조작을 실행하게 되므로 교통 환경의 인지 지연이나 판단 착오, 운전조작의 부주의 등은 곧바로 사고로 이어질 가능성이 매우 높고, 실제로도 운전자의 부주의나 조작 잘못 등에 의해 대부분의 교통사고가 발생하고 있다. 이러한 통계는 인적 요인, 즉 운전자 요인 중에서도 전방 주시 태만 등의 안전운전 불이행이 주요한 요인으로 작용하고 있음을 보여주고 있다.

위와 같이 사고 요인에 대한 통찰을 통해 얻고자 하는 결과물의 성격과 결과물을 산출하기 위한 과정에서 필요한 핵심적인 데이터를 식별하고 “Table 1. 교통사고의 주요 요인”에서 정리된 것과 같이 테이터의 범주를 정하는 중요한 단계이다.

**Table 1.** Main factors in traffic accidents

| division | | target variable |
|---|---|---|
| human factor | | Gender, age, driving experience, aptitude, driving habits, personality, etc. |
| vehicle factor | | Vehicle specifications, performance, vehicle equipment status, inspection results, etc. |
| environmental factors | meteorological environment | Temperature, precipitation, wind speed, etc. |
| | road environment | Linearity of the road, number of lanes, road surface, road width, radius of curvature, existing accident location, etc. |

## 군 안전관련 데이터 이해

### 데이터 준비

- 현 군의 사고조사 항목 평가

사고 관련 데이터는 사고조사 활동으로 부터 수집된다. 따라서 사고조사 항목이 사고 관련 데이터를 획득하는 출발점인 것이다. 현 육군의 사고조사 항목은 Table 2에서 보듯이 13개 항목으로 사고부대, 사고일시 및 장소, 사고자, 사고유형, 사고요인, 사고 피해자, 범행도구(사고기재), 피해 정도, 근무형태, 사고처리 결과, 기상 등이며, 타 기관의 사고조사 항목에 비해 체

계적이지 못하고 발생 요인에 대한 조사정보가 부족하다. 또한, 사고 유형별 사고조사 항목과 내용이 보다 세분화하여 조사될 필요가 있다. 예를 들면 교통사고 조사항목, 화재사고 조사항목 등이며 이와 관련하여 타 기관의 조사항목을 참조할 필요가 있다. 사고분석 및 예측모형 개발을 위해서는 사고 요인에 중점을 두고 이와 관련된 자료를 수집하는 것이 바람직하다.

- 군이 구축 중인 교통사고 관련 데이터 항목

| 구분 | | DB 공통 입력항목(40개) | | | 비고 |
|---|---|---|---|---|---|
| 1 | 소속 | 군 | 군단 | 사/여단 | 최초 입력 값 자동 분류 |
| | | 연대 | 발생 부대 | | |
| 2 | 발생일시 | 발생 연도 | 발생 월 | 발생 일 | 최초 입력 값 자동 분류 |
| | | 발생 시간 | 발생 요일 | | |
| 3 | 사고개요 | 8하 원칙(6하 원칙 + 누구와 + 누구에게) | | | 직접 입력 |
| 4 | 사고자 | 사고자 신분 | 사고자 계급 | 사고자 병과 | 입력 값 별도 설정 |
| | | 사고자 직책 | 사고자 복무 개월 | | |
| 5 | 사고유형 | 사고유형 대 | 사고유형 중 | 사고유형 소 | 입력 값 별도 설정 |
| | | 사고유형 세부구분 | | | |
| 6 | 사고원인 | 사고원인 개인 | 사고원인 부대 | 사고원인 환경 | 입력 값 별도 설정 |
| | | 음주여부 | | | |
| 7 | 사고장소 | 영내/영외 | 사고지점(건물) | 사고지점(도로) | 입력 값 별도 설정 |
| 8 | 사고대상(피해자) | 사고대상 대(군인/민간인) | 사고대상 중(남/여) | 사고대상 소(세부계급 등) | 입력 값 별도 설정 |
| 9 | 범행도구 | 차량(종류) | 흉기 | 총기/탄약 | 입력 값 별도 설정 |
| | | 물품 등 | | | |
| 10 | 피해정도 | 군인 사망/부상 | 민간인 사망/부상 | 재물피해(금액) | 직접입력 |
| 11 | 근무형태 | 출타여부 | 세부 근무형태 | | 입력 값 별도 설정 |
| 12 | 사고처리결과 | 사고자 처벌(형사처벌,징계), 보험 처리 등 | | | 입력 값 별도 설정 |
| 13 | 기상 | 사고 일 기상 | | | 입력 값 별도 설정 |

Fig. 3. Traffic accident data items

### 현 군이 보유한 데이터 분석

- 데이터 기초 분석 진행 방법 : 탐색적 요인 분석(Fig. 4. 참조)

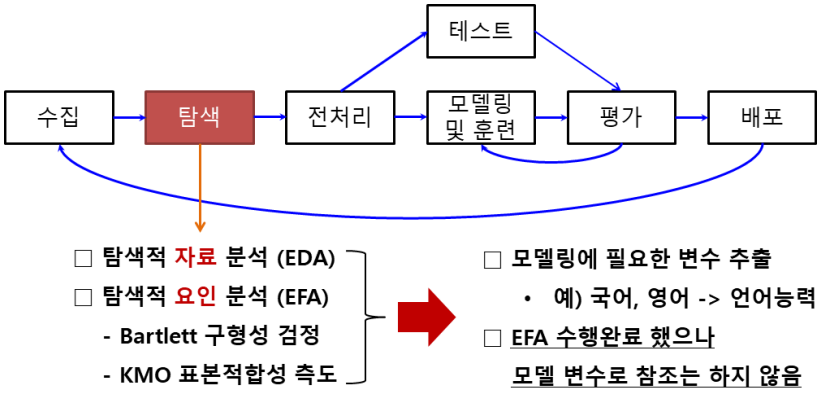


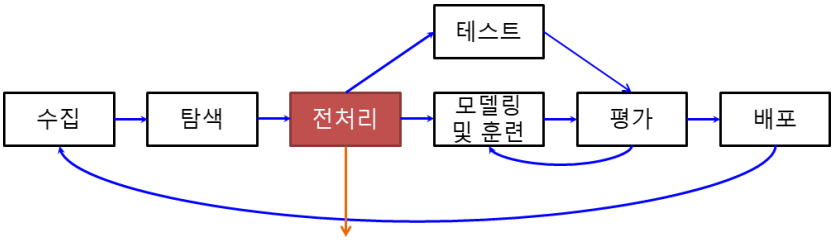

Fig. 4. Traffic accident exploratory factor analysis process

· 전제조건으로 사용되는 변수들이 모두 등간 척도나 비율 척도로 측정한 양적 변수이고, 관찰 치들은 서로 독립적이며 정규분포를 이루며, 변수별 분산은 모두 동일하다는 가정에 만족해야 한다. 입력되는 변수들 간에는 어느 정도 수준 이상의 상관관계가 존재해야 한다.

· 그러나 분석 대상 데이터가 대부분의 변수가 명목척도[3]로 분석 시 모두 등간 척도라는 가정하에 소스 스크립 트를 작성하였다.

- 바틀렛 검정 / KMO 검정 / 상관행렬 판별식 검정 모두 통과함
- Factor 개수는 8개

• 데이터 전처리(Fig. 5. 참조)




□ **Feature Engineering 공학 기법 적용 (일반적인 방법론)**

**낮은 변수 제거 → 결측값 대체 → 중심화 → 척도조정 → 상관변수 추출**

Fig. 5. Traffic accident pretreatment process

· Caret(Classification and Regression Training) 분석 패키지를 사용하여 복잡한 회귀와 분류 문제에 대한 모형 훈련과 조절 과정을 간소화하였다. 이는 훈련 데이터의 전처리, 변수의 중요성 계산 및 모형 시각화 방법 포함되어 있으며, 병렬처리하여 모델 훈련 시간을 단축할 수 있었다.

· Caret 패키지 주요함수 소개

- createDataPartition() 함수 : 훈련 및 테스트 데이터로 분류
- trainControl() 함수 : 일관된 비교방법을 각 모델 후보에게 동시 적용
- train() 함수 : 학습을 위한 표준화된 인터페이스
- confusionMatrix() 함수 : 테스트데이터에 적용한 정확도 확인

• 최종 데이터 구조 개요 (모델링 데이터 449 obs. 20 vars)

· 변수 정의 (Table 2 참조)

- 기존 EDA 데이터에서 중복되는 변수 제거 (예, 군단, 사단 -> 사단 통일)
- 결측치 있는 행은 모두 제거 (사고조사 데이터에 결측치는 없다는 전제)
- accidentType와 fatal_level 변수 추가(선행연구에 근거)

3) 예) 학년, 성별 / 사단, 부대 종류 (수송, 보병 등) 등

- fatal_level 참고 시, '산재 요양기준 교통사고 인적피해의 구분' 확인(Table 3 참조)
- 사망, 중상, 경상, 부상신고, 대형사고 등 10개 기준안 제시

  (예) 중상: 교통사고로 인하여 3주 이상의 치료 요하는 부상 입음

**Table 2.** Definition of traffic accident variables

| 컬럼명 | 의미 | 변수타입 | 레벨개수 | 샘플변수 |
|---|---|---|---|---|
| [1]season | 사계절 | Factor | 4 | "가을","겨울", |
| [2]month | 사고 발생월 | Factor | 12 | "1","2","3", |
| [3]weekdays | 사고 발생요일 | Factor | 7 | "금요일","목요일" |
| [4]time1 | 사고 발생 시간 구분 | Factor | 4 | "오전","오후", |
| [5]weather | 사고 발생 기상 | Factor | 9 | "강설","기타", |
| [6]troopCategory | 사고 관련 부대 종류 | Factor | 5 | "GOP 부대","격오지부대", |
| [7]division | 사고 관련 사단 | Factor | 75 | "11 사단","11 사단 포병여단", |
| [8]department | 사고 관련 특기 | Factor | 17 | "감찰","공병", |
| [9]rank | 운전 당사자 계급 | Factor | 23 | "6 급","7 급","8 급" |
| [10]accidentPlace1 | 사고 발생 위치 1 | Factor | 2 | "영내","영외", |
| [11]accidentPlace2 | 사고 발생 위치 2 | Factor | 28 | "D.M.Z","계곡", |
| [12]crimeTool | 사고 당시 차량 | Factor | 16 | "1/4 톤","2½톤", |
| [13]workType | 사고 당시 업무 | Factor | 21 | "개인용무","교육중", |
| [14]accidentReason | 사고 주요 요인 | Factor | 25 | "과속","과실", |
| [15]alcohol | 운전자 음주 여부 | Factor | 2 | "미음주","음주" |
| [16]case | 사고 결과 처리 법 | Factor | 13 | "과실군용물손괴" |
| [17]deadCnt | 사고 결과 사망자 수 | Int | - | 사망사고 발생건 수 30 회 |
| [18]injuredCnt | 사고 결과 부상자 수 | Int | - | 부상사고 발생건 수 300 회 |
| [19]accidentType | 사고 유형 | Factor | 4 | "car_car","car_human" |
| [20]fatal_level | 피해자 상해 심각도 구분 | Factor | 4 | "상해없음", "경상해" |

**Table 3.** Traffic accident injury level

| | | | |
|---|---|---|---|
| 상해 없음 | 단순 추돌사고 | 경상해 / 물적피해 | 인적피해 : 4 주 미만 치료 , 2 명 미만<br>물적피해 : 1 억 미만 |
| 치명적 상해 | 사망 사고 건 | 중상해 / 물적피해 | 인적피해 : 4 주 이상 치료 , 골절 , 2 명 이상<br>물적피해 : 1 억 이상 |

- 탐색적 요인 분석 결과 문제점 및 해결 방안

예측모형을 설계(모델링)하는 과정에서 데이터를 수집하고, 수집된 데이터의 이해하기 위해 탐색적 요인 분석을 실시해 본 결과 인적/물적 피해 규모를 기준으로 단순 추돌사고와 치명적 사고로 구분되는 인과관계의 의미있는 결과를 식별하였다. 이는 사고조사가 이루어진 587개 사건에 관한 데이터 범주 내에서 얻어진 결과이다. 사고는 발생했지만 경미한 사고와 중대한 사고로 구분되어 요인을 분석하고 사고의 경·중을 예측하는 모형을 설계할 수 있다는 의미이지 사고 발생과 미발생을 예

측하는 모형을 설계할 수 있다는 의미는 아니다.

앞의 2장에서 사고발생 관련 이론을 살펴본 바와 같이 50,000건의 사고 통계를 분석한 하인리히(Heinrich HW)의 연구결과에 따르면, 상해가 없거나 극히 경미한 사고가 중상해를 합친 사고보다 10배나 더 많다는 것이다. 이를 '사고의 피라미드 모형'이라 하며 적어도 300번 이상의 불안전 행동을 반복하던 집단은 경상해를 입을 경우가 평균 29회, 중상해를 입을 경우가 1회 이상 발생할 수 있을 것이라는 연구결과를 참고하였다. 사고조사에 포함되지 않았지만 경미한 사고(아차사고)[4]를 수없이 반복하고 있는 전체 집단을 대상으로 기 구축된 사고조사 데이터와 동일한 데이터 세트(data set)을 구축하여 모델링 할 수 있다는 가정하에, 전체 집단에 대한 데이터 세트(data set)의 구축을 위해 기 구축된 데이터 프레임에 추가하여 기 연구된 자료를 분석하여 얻어진 추가 항목을 포함한 새로운 데이터 프레임을 설계하였다.

### 데이터 세트(data set) 부족 문제 개선

모형개발을 위한 데이터는 기 구축된 교통사고 조사 데이터 및 대상 부대 배차데이터 그리고 새로 설계한 데이터 스키마의 데이터 항목을 융합하여 통계적 기법을 활용 새로운 데이터 세트(data set)을 생성하였다. 모형개발을 위한 데이터 프레임 형식은 아래 Table 4와 같다.

**Table 4.** Traffic accident prediction model input data frame

| 구분 | 속성 | 타입 | 레벨 | 변수 |
|---|---|---|---|---|
| 개인신상 정보 | 부대유형 | 명목형 | 11 | A-K |
| | 나이 구분 | 범주형 | 4 | 20-23세 |
| | 입대전입일 | 숫자형 | - | 1-640 |
| | 계급 | 범주형 | 4 | 이병, 일병, 상병, 병장 |
| | 혼인여부 | 범주형 | 2 | 미혼, 기혼 |
| 운전기량 정보 | 면허종류 | 범주형 | 5 | 대형, 1종, 2종, 소형, 무면허 |
| | 면허경과년수 | 범주형 | 5 | 나이구분 23세에 맞추어 계산 |
| | 운전지역숙지정도 | 범주형 | 3 | 상, 중, 하 |
| | 운전차량숙달정도 | 범주형 | 5 | 최우수, 우수, 보통, 미흡, 저조 |
| | 교통사고경험횟수 | 범주형 | 3 | 0회, 1회, 2회 |
| | 법규위반횟수 | 범주형 | 3 | 0회, 1회, 2회 |
| 차량정보 | 선탑자계급 | 범주형 | 9 | 하사-중령 |
| | 운행시간 | 범주형 | 4 | 오전, 오후, 전반야, 후반야 |
| | 운행도로 | 명목형 | 10 | A-J |
| | 주운행위치 | 명목형 | 25 | 개인용무, 훈련중 등 |
| | 차량종류 | 명목형 | 10 | 2 1/2톤, 5톤, 승용 등 |
| | 차량생산연월일 | Date | - | 1990-2018 |
| | 누적운행거리 | 숫자형 | | 연도별 누적 운행거리 추정 |
| | 최근정비기록 | Date | | 2018년 1월 1일 ~ 12월 31일 |
| 운행환경 | 운행월일 | Date | | 2018년 1월 1일 ~ 7월 31일 |
| | 날씨 | 범주형 | 4 | 우천, 강설, 맑음, 흐림 |
| | 강수량 | 숫자형 | - | 강우 또는 강설량 |
| | 도로위험예측정보 | 범주형 | 4 | 안전, 주의, 위험, 심각 |
| 종속변수 | 사고유무 | 범주형 | 2 | 사고/미사고 |

4) 아차사고 : 작업자의 부주의나 현장 설비 결함 등으로 사고가 일어날 뻔하였으나 직접적인 사고로는 이어지지 않은 상황을 말한다. 이러한 아차사고는 대형 사고의 전조증상이라고도 할 수 있다.

총 데이터 세트(data set) 10587개 중, 10000개의 데이터(비사고 데이터)[5]는 기존정보를 재구성하는 방법을 토대로 데이터 세트(data set)을 만들었다. 레이블이 되는 587개 데이터는 실제 사고조사 데이터를 입력하였다.

### 교통사고 예측모형 설계

CRISP-DM의 모델링(Modeling) 단계로 R을 활용한 머신러닝 모형을 개발하여 탑재하여 시연될 수 있도록 진행하였고, 프로그래밍 언어는 R과 R Shiny을 활용하였다. 예측모형 설계를 위한 개발환경은 Table 5를 참조한다.

**Table 5.** Traffic accident model development environment

| | |
|---|---|
| server environment | Google Cloud Compute Engine |
| OS environment | Ubuntu 18.04 LTS |
| model development programming language | R |
| web development framework | Shiny |
| R main package | tidyverse, caret, shiny |

적용 알고리즘은 로지스틱 회귀분석, 랜덤포레스트, 서포트백터머신을 3가지를 사용하였고, 총 데이터 10587개 중 6353개를 이용하여 모형을 훈련시켰으며, 모형의 검증은 4234개의 데이터를 활용하였다.

분류기준을 0.5로 했을 때, 적용 알고리즘별 오분류표는 아래 Table 6, 7, 8과 같다.

**Table 6.** Logistic regression misclassification table

| Prediction | Reference | |
|---|---|---|
| | non-accident | accident |
| non-accident | 3999 | 46 |
| accident | 1 | 188 |

**Table 7.** Random forest misclassification table

| Prediction | Reference | |
|---|---|---|
| | non-accident | accident |
| non-accident | 3986 | 45 |
| accident | 14 | 189 |

5) 사고가 실제 발생하지 않은 사건 데이타로 최초 수집되지 않았으나 데이터 준비시 군 운행기록 정보체계에서 수집하여 동일한 데이터 프레임으로 재구성하였으며, 기 수집된 사고데이터 587건과 함께 비사고 데이터 10.000건을 구축하여 데이터 셋를 만들었다.

**Table 8.** Support vector machine misclassification table

| Prediction | Reference | |
|---|---|---|
| | non-accident | accident |
| non-accident | 4000 | 47 |
| accident | 0 | 187 |

분류기준 0.5로 했을 때, 각 모형별 정확도, 특이도, 민감도, AUC는 아래 Table 9와 같이 나타났다. 오분류 표에서 미사고 사건을 사고로 오분류 하는 비율과 실제 사고 사건을 미사고로 오분류 하는 비율의 차이가 발생한다. 이는 최초 모형을 설계할 때 실제 사고조사 데이터 587개의 탐색적 요인분석 결과, 인적/물적 피해 규모를 기준으로 단순 추돌사고와 치명적 사고로 구분되는 사고요인과 사고결과의 인과관계를 반영하였고, 단순 추돌사고는 미사고로 레벨을 낮추어 데이터 세트(data set)을 구성하는 기준으로 설계하였기 때문이다. 그리고 하인리히 이론을 살펴보면 사고와 미사고로 단순히 구분하는 것보다 피해 수준에 따른 구분이 더 이론에 부합하다. 따라서 사고조사 데이터이지만 피해 기준을 어떻게 설정하였는가에 따라 미 사고로 분류될 수 있다. 개발모형에서는 사고조사 데이터와 미사고 데이터를 통합하여 분류한 결과치를 사용하였다.

**Table 9.** Accuracy, specificity, sensitivity and ACU by predictive model

| 구분 | Logistic Regression | Random Forest | Support vector machine |
|---|---|---|---|
| .Accuracy | 0.986 | 0.988 | 0.988 |
| specificity | 0.807 | 0.803 | 0.799 |
| sensitivity | 0.996 | 0.999 | 1.000 |
| AUC | 0.972 | 0.966 | 0.915 |

정확도에는 큰 차이가 보이지는 않지만, AUC가 상대적으로 높은 로지스틱회귀모형을 선택하고, R 프로그래밍 언어에서 제공하는 Web Framework를 통해 가상의 서버에 실제 Shiny 웹개발 및 모형을 구축하였다.

## 예측모형 평가

예측모형을 평가는 10 Fold Cross Validation을 이용하였다. 예측모형의 과적합 문제에 대해서는 머신러닝의 학습 데이터 및 검증을 위한 데이터가 재구성된 데이터 세트(data set)이 사용되어 연구에서 제외하였다. 차이는 미미하지만 오분류율, AUC 등에 있어서 로지스틱 회귀분석이 랜덤 포레스트나 서포트벡터머신에 비해 상대적으로 안정적임을 확인할 수 있었다.

## 향후 데이터 확보 및 예측모형 개선 방안

완성된 모델을 실 업무에 적용하기 위해서는 평시 사고 발생 데이터와 동등한 수준의 비사고(아차 사고[6]) 데이터를 동시에 획득하는 것이 중요하며, 이를 위해 관련 정보수집 수단 및 정보체계를 통합하거나 데이터를 유통시키는 시스템이 추가로

6) 하인리히의 '사고의 피라미드 모형'에서 적어도 300번 이상의 불안전 행동을 반복하던 사람은 경상해를 입을 경우가 평균 29회, 중상해를 입을 경우가 1회 이상 발생하는데 최초 300번 이상의 미상해 행동을 군에서는 아차 사고로 규정함.

개발되어야 할 것이다.

향후 발전시킬 사항으로 운행 기록 정보(배차데이터)에 교통사고를 발생시키는 요인(속성)에 대한 데이터 값 즉, 실제 운행과 관련된 데이터를 기준으로 사고조사 데이터에 포함되어 있는 데이터 항목을(결측 속성) 추가 포함한 데이터 프레임을 생성하고 이를 데이터 세트(data set)로 준비한다. 그리고 이를 다양한 모델링 기법과 알고리즘을 선택하고 모델링 과정에서 파라메타를 최적화하는 단계와 모델을 테스트용 프로세스와 데이터 세트(data set)로 평가하는 방법을 적용하여 교통사고 예측모형으로 개발할 수 있다. 이를 가칭 "사고조사데이터 및 배차데이터가 결합된 머신러닝 형태의 교통사고예측 모형" 개발 방안으로 Fig. 6와 같이 제시하였다.

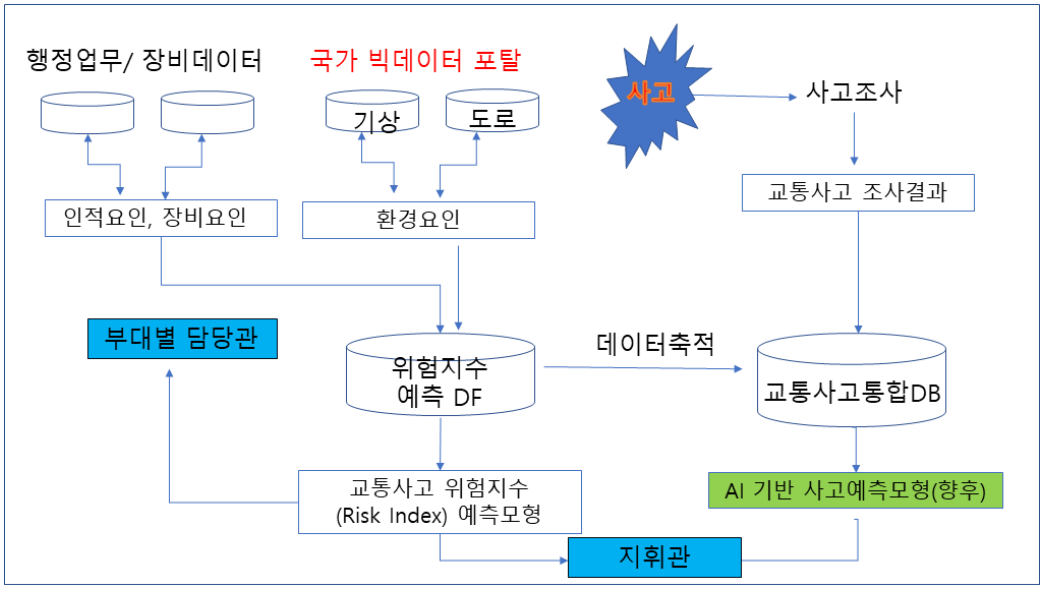



Fig. 6. Risk index prediction model (draft)

## 결 론

머신러닝에 의한 문제해결의 가장 큰 문제는 사용 가능한 형태로 데이터를 수집하는 일이다. 데이터 과학자들은 문제해결 시간의 80%를 데이터 준비에 소모되는 것으로 조사된바[7] 있다. 오늘날 데이터를 활용한 연구는 훨씬 더 방대한 양의 데이터와 다양한 형태의 데이터가 필요하나, 실제로는 보유하고 있는 데이터가 연구에 필요하도록 설계되어 있지 않다는 점이다. 특히 예측하려고 하는 문제가 복잡하다면 훨씬 더 많은 데이터와 복잡한 알고리즘이 필요하며 데이터의 양적 충분성과 품질이 보장되지 않으면 예측 결과에 대한 정확도를 담보하기 어려운 것이 현실이다. 또 한 학습된 모형이 훈련 데이터에는 완벽하게 적합하지만 실제 데이터에는 정확도가 떨어지는 과적합(Overfitting) 문제 역시 해결해야 하는 과제들이다.

사고예측을 위해서 10여 년의 군 사고데이터(헌병의 사고조사 데이터)와 일부 부대의 차량운행과 관련된 운행기록 데이터 및 보험가입 데이터를 수집하였다. 센서로부터 장시간 축적된 데이터와는 달리 정보처리 시스템에 의해서 처리/가공된 데이터의 대부분은 데이터의 작성 용도가 특정되어 있으며, 용도의 처리에 필요한 데이터의 속성 이외의 속성들은 포함하고 있지

7) 스티브로어(Steve Lohr) "For Big-Data Scientists, Janitor work is key hurdle to insight" (New York Times, 2014, 8.17)

않은 것이 일반적이다. 예를 들어 군에서 제공한 교통사고 데이터로부터 사고와 관련된 의미 있는 속성을 발견하기 위한 변수 간의 상관관계나 다변량 분석 등을 수행해서 유의성 있는 변수 집합을 찾아내 는 것이 쉽지 않았으며, 따라서 이러한 데이터로 머신러닝으로 예측하는 모형을 개발하여도 정확도가 현저하게 떨어지는 현상이 나타났다. 이러한 점들을 보완하기 위하여 교통사고예측과 관련된 민간 연구들을 벤치마킹하여 사고에 영향을 미치는 요인들을 찾아내고 그 요인에 대한 값을 추정하여 데이터를 보완하는 작업을 수행하였다.

기존의 연구들을 통해 보면 교통사고의 3대 요인인 도로환경요인, 차량 요인, 인간 요인 중 인간 요인이 가장 큰 영향을 미치는 것으로 조사되었으며, 특히 운전경력과 운전자 성격이 가장 영향을 많이 미치는 것으로 확인되었다. 그러나 군의 사고조사 데이터에는 운전자에 대한 정보가 부족하였고, 차량의 주요한 속성이나 차량의 정비기록과 관련된 정보는 교통사고 조사 데이터에 포함되어 있지 않았다. 예측모형에 필요한 데이터는 일반적인 사고조사 자료와 단순 차량운행 기록만으로는 획득이 어렵고, 운전자의 인성검사 결과와 차량 정비기록 등을 입체적으로 확인해야 획득할 수 있다. 또 한 사고 순간에 필요한 모든 데이터 들이 실시간 종합적으로 조사되고 기록되지 않으면 사후에는 시간을 역행해서 해당 데이터를 확보하는 것이 제한된다. 본 연구에서는 위와 같은 문제점을 해결하기 위해 사고 요인에 대한 추가 변수를 식별하여 결측 속성에 대한 합리적인 가정 값을 통계적 기법으로 추론하였으며, 재구성된 데이터 세트(data set)을 활용하여 예측모형을 통해 의미 있는 결과치를 도출하였다.

## References


[1] Choi, H.-Y., Min, Y.-H. (2015). “Intelligent information systems; Introduction to deep learning and major issues.” Korea Information Processing Society, Vol. 22, No. 1, pp. 7-15.

[2] Choi, W., Kim, Y., Jang, D., Kim, G., Jeong, Y. (2017). “A study on the development of a fire risk prediction model for manufacturing facilities using artificial neural networks.” Magazine of The Korean Society of Hazard Mitigation, Vol. 17, No. 1, pp.161-167.

[3] Jo, I.-H. (2016). Data analysis-based Accident Prediction Service for Reducing Traffic Accidents. National Information Society Agency, NIA Ⅱ-RER-D-16017. Seoul.

[4] Jeong, J.-U., Lee, S.-J. (2020). “Analysis of characteristics of hazardous chemical transport vehicle accidents in Korea.” Journal of the Society of Disaster Information, Vol. 16, No. 2, pp. 310-317.

[5] Kim, D., Kim, D.-K., Lee, C. (2013). Safety Performance Functions Reflecting Categorical Impact of Exposure Variables for Freeways. Transportation Research Board Annual Meeting, Korea.

[6] Lee, G.-H., Roh, J.-H. (2015). “Development of a traffic accident prediction model using probability parameter - Targeting the 4 intersections of the metropolitan area and Busan Metropolitan City.” Journal of the Korean ITS Society, Vol. 14, No. 6, pp. 91-99.

[7] Lee, S. (2012). “Comparison of initial center selection methods in K-means clustering.” Internet Information Society, Vol. 13, No. 6, pp. 1-8.

[8] Pan, G., Fu, L., Thakali, L. (2017). “Development of a global road safety performance function using deep neural networks.” International Journal of Transportation Science and Technology, Vol. 6, No. 3, pp. 159-173.

[9] Park, J.-B. (2015). Development of a Prediction Model for Traffic Accidents Using Probability Parameters. Ph.D. Dissertation, University of Seoul Graduate School.

[10] Simon, W.P., Mattew, G.K., Fred L.M, (2010). Statistical and Econometric Method ForTransportation Data

Analysis. CRC Press, New York.

[11] Yoo, J.-D. (2018). Development of a Prediction Model for Highway Traffic Accidents Using Deep Learning. Ph.D. Dissertation, Graduate School of Ajou University.

[12] Yun, Y., Lee, J., Kim, J., Kim, Y. (2020). “Detection scheme of heart and respiration signals for a driver of car with a doppler radar.” Journal of the Society of Disaster Information, Vol. 16, No. 1, pp. 87-95.