ISSN 1229-6783(Print)
ISSN 2288-1484(Online)

건설 사고사례 데이터 기반 건설업 사망사고 요인분석

최 지 윤* · 김 시 현* · 이 송 이* · 김 경 훈** · 이 수 동*
*울산대학교 산업경영공학부 · **한국산업안전보건공단

A Data-Driven Causal Analysis on Fatal Accidents in Construction Industry

Jiyoon Choi* · Sihyeon Kim* · Songe Lee* · Kyunghun Kim** · Sudong Lee*
*Department of Industrial Engineering, University of Ulsan
**Korea Occupational Safety and Health Agency

Abstract

The construction industry stands out for its higher incidence of accidents in comparison to other sectors. A causal analysis of the accidents is necessary for effective prevention. In this study, we propose a data-driven causal analysis to find significant factors of fatal construction accidents. We collected 14,318 cases of structured and text data of construction accidents from the Construction Safety Management Integrated Information (CSI). For the variables in the collected dataset, we first analyze their patterns and correlations with fatal construction accidents by statistical analysis. In addition, machine learning algorithms are employed to develop a classification model for fatal accidents. The integration of SHAP (SHapley Additive exPlanations) allows for the identification of root causes driving fatal incidents. As a result, the outcome reveals the significant factors and keywords wielding notable influence over fatal accidents within construction contexts.

Keywords : Occupational safety, Fatal accident, Machine learning, SHAP

1. 서 론

고용노동부 산업재해 현황에 따르면 2020년부터 2022년까지 건설업에서 발생한 사고 재해는 계속해서 증가하고 있다. 근로자 100명당 재해자 수를 나타내는 재해율의 경우, 2022년 건설업 1.25%로 광업(39.32%)을 제외한 운수 · 창고통신업(1.16%), 어업(1.06%), 제조업(0.79%) 등 타 산업에 비해 높은 수준이다. 건설 현장에는 다양한 중장비와 구조물이 섞여 있고, 표준화가 어려운 직접 노동 중심의 작업이 많아 산업재해의 위험이 크고 사고의 예방 및 관리가 어렵다. 건설사고 저감을 위해서는 과거 사례에서 정확한 사고 발생 요인을 분석하고 재발 방지를 위한 적절한 대책을 수립해야 한다. 특히, 중대재해처벌법 시행으로 인한 중대재해 관리의 필요성이 높아짐에 따라 사업장에서의 사망사고 예방을 위한 요인분석 방법론이 필요하다.

빅데이터, 인공지능, 사물인터넷 등 관련 기술의 발전과 안전에 대한 사회적 관심 증대로 인해, 건설 현장에서도 사고 및 안전에 대한 다양한 데이터가 수집되고 있다. 이러한 데이터를 기반으로 건설사고 요인을 분석하기 위한 다양한 선행 연구가 있었다. 선행연구에서 사용된 건설사고 데이터는 텍스트 중심의 비정형데이터와 수치 중심의 정형데이터로 구분된다. 텍스트데이터를 활용한 사례로 Lee[1]는 안전사고 관련 뉴스데이터에 토픽모델링을 적용하여 건설안전사고 동향을 분석하였다. Park 등[2]은 건설업 재해사례 데이터에 텍스트마이닝을 적용하여 건설공사 위험 요소를 도출하고 계절별 중요도를 분석하였다. Zhang 등[3]은 OSHA (Occupational Safety and

†이 논문은 2023년도 한국산업단지공단의 재원으로 울산 스마트제조고급인력양성사업의 지원을 받아 수행된 연구임
†Corresponding Author : Sudong Lee, Department of Industrial Engineering, University of Ulsan, 93 Daehak-ro, Nam-gu, Ulsan, E-mail: sudonglee@ulsan.ac.kr
Received August 22, 2023; Revision September 22, 2023; Accepted September 27, 2023

Health Administration)의 건설 사고사례 데이터에 텍스트마이닝 기법과 support vector machine (SVM), 선형회귀, K-nearest neighbor (KNN), 의사결정나무, Naive Bayes의 앙상블을 활용한 사고원인을 분류 기계학습 모델을 학습했다. 정형데이터를 분석한 사례로 Kim 등[4]은 건설안전사고 사전 예측을 위해 건설사고 사례 데이터에 인공신경망 기법을 적용하였다. Cho[5]는 건설현장 재해유형별 원인 분석 및 예방대책 수립을 위한 통계분석을 수행하였다. Choi 등[6]은 공사 사전정보를 활용하여 건설 현장 사망사고를 사전에 예측하기 위한 기계학습 모델을 개발하였다. Cho 등[7]은 의사결정나무를 활용한 건설 현장 안전사고 유형을 예측 기법을 제안하였다. Xu 등[8]은 설문조사 데이터를 활용하여 건설 사고에 영향을 파악하기 위한 통계적 상관분석을 수행하였다. Hola 등[9]은 사고 발생 요일, 지역, 원인 등 정형데이터를 기반으로 중국의 건설업 사망사고의 특징을 분석하였다. Udawatta 등[10]은 건설 현장의 사망사고와 비사망사고가 가지는 차이를 분석하기 위해 스리랑카에서 수집된 건설 사고사례 데이터에 카이제곱 검정을 적용하였다. Amiri 등[11]은 이란의 건설 사고사례 데이터에 다중 대응분석(multiple-correspondence analysis), 의사결정나무, 연관분석 등 데이터마이닝 기법을 적용하여 사고 유형 별 주요 요인을 도출하였다.

본 연구에서는 국토교통부가 건설공사 안전관리 종합정보망(Construction Safety Management Integrated Information, CSI)을 통해 온라인으로 제공하는 건설 사고사례 데이터에 통계분석 및 기계학습을 적용하여 사망사고의 요인을 파악하는 방법론을 제안한다. 정형데이터 또는 텍스트 비정형데이터 중 한 가지만을 활용했던 대부분의 기존 연구와 달리, 본 연구는 건설 사고의 기본정보를 요약하는 정형데이터와 사고 경위 및 원인을 설명하는 텍스트 비정형데이터를 모두 활용하여 사망사고 요인을 다각적으로 분석한다. Cho 등[12]과 같이 건설 현장 정형 · 비정형 데이터를 모두 활용하여 기계학습 기반의 건설 재해 예측 모델을 제안한 연구도 있으나, 사망사고 발생에 영향을 미치는 주요 요인에 대한 분석은 수행되지 않았다. 본 연구에서는 수집된 사고사례 데이터로 사망사고와 비사망사고를 분류하는 기계학습 모델을 학습한 후, 해석가능한 인공지능 기술인 SHAP(Shapley Additive exPlanations)[13]을 통해 분류모델로부터 각 사고가 발생한 원인을 정량적으로 파악한다. 본 연구가 제안하는 방법론을 통해 건설업 사망사고의 원인을 파악하고 이후 유사 사고의 발생 방지를 위한 대책 마련의 기초 자료를 확보할 수 있다.

2. 연구 방법

2.1 데이터 수집 및 전처리

본 연구에서는 건설업 사망사고 요인분석을 위해 국토교통부 CSI에서 제공하는 2019년 7월부터 2023년 4월까지의 건설 사고사례 데이터를 수집하였다. 이 데이터는 건설공사 참여자와 일반 국민으로부터 건설공사 안전관리 종합정보망으로 신고된 건설 사고의 정형 · 비정형데이터로 구성되어 있다.

정형데이터는 총 14,318건(사망사고 717건, 비사망사고 13,601건)의 건설업 사고사례에 대한 <Table 1>의 34개 변수를 분석 대상으로 한다. 기계학습 기반 분류모델의 학습을 위해 결측치를 가지는 행은 제거하였다. 종속변수인 '사망사고'는 사망자 수가 1명 이상인 경우 사망사고로 간주하여 '1'로, 사망자 수가 0명인 경우 비사망사고로 간주하여 '0'으로 설정하였다. 공사 진행의 시급성을 표현하기 위한 파생 변수 '공사기간'과 '공사비/공사기간' 변수를 추가하였다. 범주형 데이터는 기계학습 모델 적용을 위해 one-hot encoding으로 변환하였다.

비정형데이터는 총 14,602건(사망사고 859건, 비사망사고 13,743건)의 사고사례에 기록된 '구체적사고원인' 항목을 분석에 활용하였다. 내용은 자유 형식으로 기록된 텍스트로 오 · 탈자가 존재하며, 같은 내용을 뜻하는 서로 다른 표현이 혼재한다. 텍스트데이터 기반의 기계학습 모델 학습을 위해 영어, 한국어, 숫자 형태의 건설장비명, 건설자재명 등 1,670개 단어를 사용자 사전에 등록하였다. '

2.2 데이터 분석

본 연구의 데이터 분석 절차는 <Figure 1>과 같다. 첫째, 수집 후 전처리를 거친 2020~2022년 건설업 사고사례 정형 · 비정형데이터를 기반으로 국내 건설 사고 발생 현황을 분석했다. 인적 사고 종류, 사고 객체, 사고 발생 연도 · 월 · 시간, 날씨에 따른 건설 사고와 사망사고의 발생 현황을 비교하였다.

둘째, 통계 기반의 건설업 사망사고 요인분석을 수행하였다. 정형데이터의 각 변수를 그룹화하고 사망사고 발생 여부와의 상관관계를 분석하기 위해 교차분석을 수행하였다. 텍스트로 구성된 비정형데이터의 경우 사망사고와 비사망사고의 '구체적사고원인'에 등장하는 단어들의 빈도를 비교 분석하였다. 사망사고의 사고원인에 특징적으로 높은 빈도로 등장하는 단어를 통해 사망사고의 요인을 유추했다.

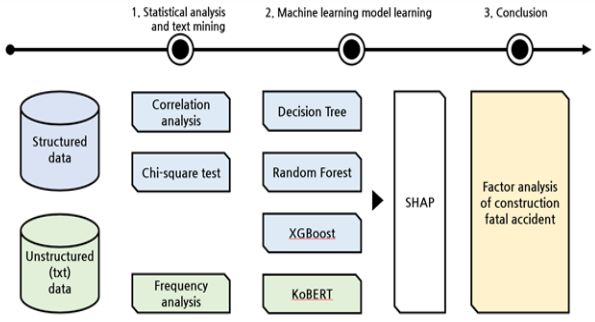
ISSN 1229-6783(Print)
ISSN 2288-1484(Online)

<Table 1> Summary of structured data

Variable	Type (number of categories)	Feature
Fatal accident	Categorical (2)	fatal accident: 1, non-fatal accident: 0
Temperature	Numerical	℃
Humidity	Numerical	%
Construction duration	Numerical	duration between the construction start and end dates
Work duration	Numerical	duration between the work start and end dates
Construction cost/ Construction duration	Numerical	'construction cost' divided by 'construction duration'
Work cost/ Construction duration	Numerical	'work cost' divided by 'work duration'
Year	Categorical (5)	2019~2023 (by year)
Month	Categorical (12)	January~December
Time	Categorical (24)	0~23 (by hour)
Public	Categorical (2)	'public', 'private'
Weather	Categorical (6)	'rainy', 'snowy', etc
Facility category (major)	Categorical (3)	'architecture', 'civil engineering', etc
Facility category (medium)	Categorical (13)	'building', 'water and sewage', etc
Facility category (minor)	Categorical (62)	'community housing', 'educational/research facility', etc
Human accident type (major)	Categorical (15)	'hit by object', 'fall down', etc
Human accident type	Categorical (20)	'electric shock', 'traffic accident', etc
Construction type (major)	Categorical (7)	'mechanical facility', 'electrical facility', etc
Construction type (specific)	Categorical (39)	'temporary work', 'steel concrete work' etc
Accident object (major)	Categorical (9)	'construction equipment', 'construction material' etc
Accident object (specific)	Categorical (118)	'formwork', 'tools' etc.
Work process	Categorical (41)	'painting', 'loading', 'unloading', etc
Cause (major)	Categorical (3)	'construction error', 'design error', etc
Cause (medium)	Categorical (13)	'inadequate construction method', 'safety rule violation, etc
Cause (minor)	Categorical (49)	'inattention', 'inadequate construction method' etc
Damage cost	Categorical (8)	8 intervals between 'below 10 million KRW' and 'over 500 million KRW'
Number of workers	Categorical (6)	6 intervals between '19 or fewer workers' and '500 or fewer workers
Design stability review	Categorical (2)	'yes', 'no'
Accident investigation method	Categorical (2)	'by committee', 'general' etc
Construction cost	Categorical (17)	17 intervals between 'below 100 million KRW' and 'over 100 billion KRW'
Work cost	Categorical (17)	17 intervals between '10 million~20 million KRW' and 'over 100 billion KRW'
Bid rate	Categorical (8)	8 intervals between 'below 60%' and 'over 90%' by 5%
Progress rate	Categorical (10)	10 intervals between 'below 10%'and 'over 90%' by 10%
Safety management plan	Categorical (3)	'site(type 1 and 2)','N/A', etc

셋째, 수집된 데이터로 사망사고를 분류하는 기계학습 모델을 학습하고 SHAP을 활용해 사망사고 요인을 추론하였다. 분류모델의 입력 변수에 대한 SHAP value를 계산하여 각 변수와 사망사고 발생의 인과관계를 확인하고, 큰 절대값의 SHAP value를 가지는 변수를 사망사고 발생의 주요 원인으로 가정하였다. 정형데이터를 위한 분류모델로는 의사결정나무(decision tree), 랜덤포레스트(random forest), XGBoost[14]를 사용하였으며, 텍스

트데이터는 한국어 자연어처리를 위한 딥러닝 모델인 KoBERT[15] 분류모델을 학습했다.

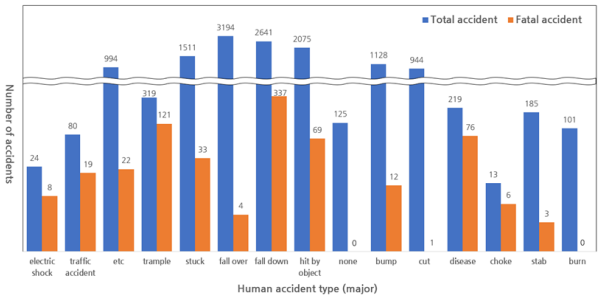


[Figure 1] An overview of the proposed method.

3. 연구 결과

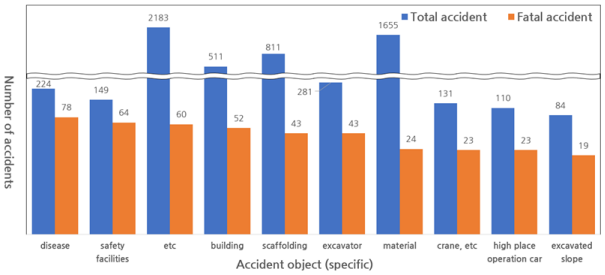
3.1 건설사고 현황 분석

2020부터 2022년 사이에 발생한 국내 건설사고 현황을 파악하기 위해 인적사고 종류, 사고 객체, 사고 발생 연도 · 월 · 시간, 날씨에 따른 전체 건설 사고와 사망사고의 분포를 비교하였다.



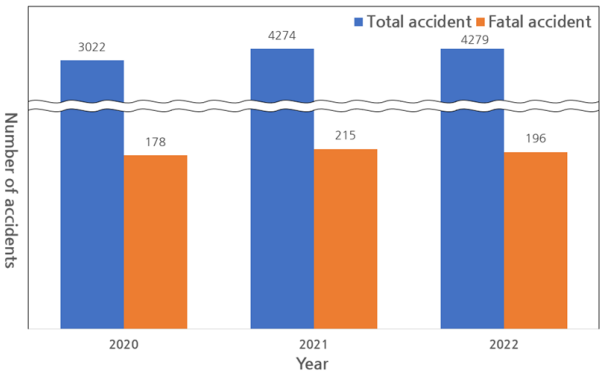
[Figure 2] Construction accidents in Korea between 2020 and 2022 according to 'Human accident type (major)'.

<Figure 2>는 '인적사고종류(대분류)'에 따른 전체사고와 사망사고의 분포를 나타낸다. 사망사고는 '떨어짐(337건)', '깔림(121건)', '질병(76건)', '물체에 맞음(69건)' 등의 순으로 자주 발생하였다. 전체사고 대비 사망사고의 비율은 '질식(32%)', '깔림(28%)', '질병(26%)', '감전(25%)' 등의 순으로 높았다.



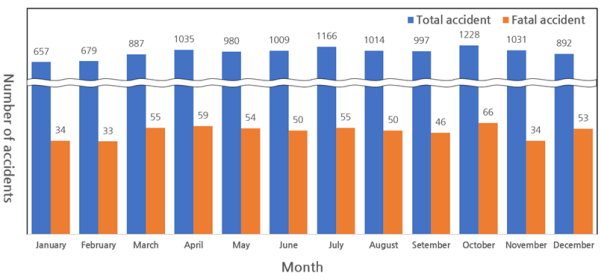
[Figure 3] Construction accidents in Korea between 2020 and 2022 according to 'Accident object (specific)'.

<Figure 3>은 '사고객체(소분류)'에 따른 전체사고와 사망사고의 분포를 나타낸다. 분석 결과 해석의 용이성을 위해 '사고객체(소분류)'의 108개 범주 중 사망사고 발생 건수가 높은 상위 10개 범주만을 분석 대상으로 하였다. 사망사고는 '질병(78건)', '안전시설물(64건)', '건물(52건)', '비계(43건)' 등의 순으로 자주 발생하였다. 전체사고 대비 사망사고의 비율은 '안전시설물(30%)', '질병(28%)', '굴착사면(18%)', '고소작업차(17%)' 등의 순으로 높았다.



[Figure 4] Construction accidents in Korea between 2020 and 2022 according to 'Year'.

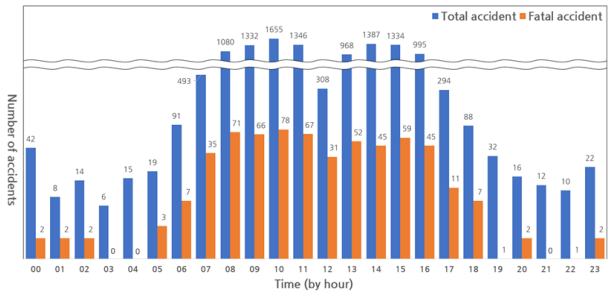
<Figure 4>는 '사고발생 연도'에 따른 전체사고와 사망사고의 분포를 나타낸다. 2020년 전체사고 수는 2021년, 2022년보다 적었으나 전체사고 대비 사망사고의 비율은 6%로 2021년(5%), 2022년(4%)에 비해 높았다.



[Figure 5] Construction accidents in Korea between 2020 and 2022 according to 'Month'.

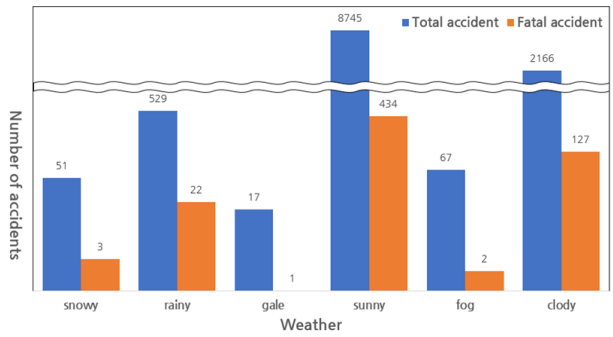
ISSN 1229-6783(Print)
ISSN 2288-1484(Online)

<Figure 5>는 '사고발생 월'에 따른 전체사고와 사망사고의 분포를 나타낸다. 전체사고 수는 다른 월에 비해 '1월(657건)'과 '2월(679건)'에 적었다. 전체사고 대비 사망사고의 비율은 '3월'과 '12월'에 6%로 가장 높았으며, '11월'에 3%로 가장 낮았다.



[Figure 6] Construction accidents in Korea between 2020 and 2022 according to 'Time'.

<Figure 6>은 '사고발생 시간'에 따른 전체사고와 사망사고의 분포를 나타낸다. 12시를 전후로 건설 사고가 자주 발생하는 경향을 보였다. 전체사고 대비 사망사고의 비율은 '01시(20%)'가 가장 높았고, '02시(13%)', '05시(14%)', '20시(11%)', '12시(9%)' 등의 순으로 높았다.

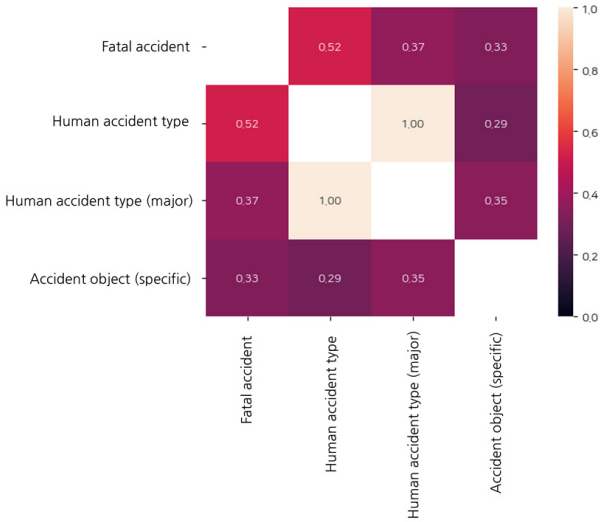


[Figure 7] Construction accidents in Korea between 2020 and 2022 according to 'Weather'.

<Figure 7>은 '날씨'에 따른 전체사고와 사망사고의 분포를 나타낸다. '맑음(8,745건)'과 '흐림(2,166건)'의 경우 전체사고 수가 많았다. 전체사고 대비 사망사고의 비율은 '강설', '강풍', '흐림'이 6%로 높게 나타났으며, '안개'는 3%로 가장 적었다.

3.2 통계 기반 건설업 사망사고 요인분석

정형데이터의 변수들과 사망사고 사이의 상관관계를 분석하기 위해 30개 범주형 독립변수와 종속변수 '사망사고'의 크래머 V 계수(Cramer's V)를 계산하였다. 계산된 크래머 V값이 0.3 이상인 독립변수에 대한 분석 결과는 <Figure 8>과 같다. 분석 결과, 사망사고와의 크래머 V 계수는 '인적사고종류(0.52)', '인적사고종류(대분류)(0.37)', '사고객체(소분류)(0.33)'순서로 나타났다. '온도', '습도' 등의 연속형 독립변수와 사망사고의 관계에 대해서도 피어슨 상관계수 기반의 상관분석을 수행하였으나, '사망자 수' 등 사망자 발생 여부를 직접적으로 나타는 변수 이외에 유의한 상관성을 가지는 변수를 찾을 수 없었다.



[Figure 8] Cramer's V between the categorical input variables and 'Fatal accident'.

<Figure 8>의 범주형 독립변수와 사망사고 간 상관관계의 통계적 유의성을 검정하기 위해 카이제곱 교차검정(Chi-squared test)을 수행하였다. 그 결과, <Table 2>와 같이 유의수준 α=0.05 하에서 '인적사고종류', '인적사고종류(대분류)', '사고객체(소분류)'는 모두 '사망사고'와 유의한 상관관계를 가지는 것으로 나타났다.

<Table 2> Chi-squared test results

Variable	df	F	P-value
Human accident type	21	3634.91	0.0
Human accident type (major)	14	1842.87	0.0
Accident object (major)	117	1563.68	0.0

텍스트데이터로부터 사망사고와 관련 있는 단어를 추출하기 위해 빈도분석을 수행하였다. 2019년 7월부터 2023년 4월까지 발생한 전체 건설사고를 16,778건의 '사망사고'와 859건의 '비사망사고'로 구분하고, '구체적 사고원인'에 자주 등장하는 단어에 차이가 있는지 분석하

였다. 한국어 자연어처리를 위한 오픈소스 라이브러리 KoNLPy[16]의 Mecab 형태소 분석기를 사용하여 명사, 동사, 형용사, 숫자, 영어를 추출하고 '사고', '작업'과 같은 범용어, '사고 장소', '사고 회사'와 같은 사례 고유 정보, 특수문자, 오타는 불용어로 처리 후 제거하였다.

<Table 3> Frequency analysis results of fatal and non-fatal accidents

Fatal Accident		Non-fatal Accident	
Word	Frequency	Word	Frequency
Fall down	259	Material	2,236
Level	142	Installation	1,869
Installation	121	Inattention	1,755
Equipment	79	Rebar	1,736
Construction	72	Formwork	1,481
Harness	72	Floor	1,417
Floor	63	Foot	1,329
Material	63	Scaffold	1,266
Demolition	61	Shortcoming	1,192
Fastening	60	Floor	1,188

<Table 3>은 사망사고 및 비사망사고의 '구체적사고원인' 텍스트 빈도분석 결과 출현 빈도 상위 10개 단어와 빈도를 나타낸 결과이다. '층', '설치', '바닥', '자재'는 사망사고와 비사망사고에서 모두 높은 빈도로 등장하였다. 사망사고에서 특징적으로 자주 등장한 단어는 '추락', '장비', '공사', '고리', '해체', '체결'이 있었다. 반면 비사망사고에서는 '부주의', '철근', '거푸집', '발', '비계', '미흡'이 특징적으로 자주 등장하였다. <Figure 9>와 <Figure 10>은 빈도분석 결과를 워드 클라우드(word cloud)로 시각화한 결과이다. 도출된 빈출 단어는 사망사고와 비사망사고의 요인을 파악하기 위한 기초 자료로 활용할 수 있다.

[Figure 9] A word cloud of 'Cause (minor)' of fatal accidents.

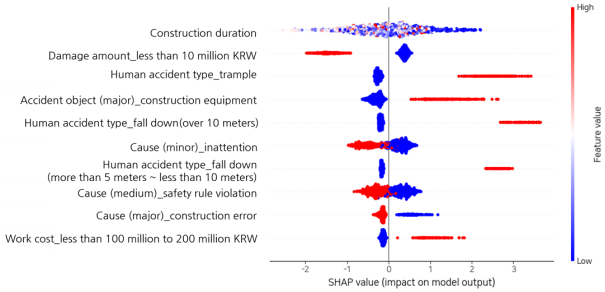
[Figure 10] A word cloud of 'Cuase (minor)' of non-fatal accidents.

3.3 기계학습 기반 건설업 사망사고 요인분석

이번 장에서는 기계학습 기반 건설업 사망사고 요인분석 방법을 제안한다. 통계 기반 요인분석의 경우 각 독립변수와 종속변수 간 상관관계를 파악할 수 있지만, 여러 독립변수가 종속변수에 미치는 상호 복합적인 관계를 설명할 수 없다는 한계가 있다. 본 연구에서는 이 문제를 해결하기 위해 수집된 데이터의 여러 독립변수를 함께 사용하여 사망사고를 예측하는 기계학습 모델을 학습한 후, 각 독립변수가 종속변수에 미치는 인과적 영향을 파악할 수 있는 SHAP을 통해 사망사고의 요인을 분석하고자 한다. SHAP은 기계학습 모델에 입력된 데이터의 종속변수에 대한 각 독립변수의 기여도를 계산하기 위한 방법론으로서, 기계학습 모델의 종류에 관계없이 적용할 수 있다.

수집된 정형데이터의 독립변수 중 중복 정보를 가지는 두 변수인 '인적사고종류'와 '인적사고종류(대분류)' 중 보다 상세한 정보를 나타내는 '인적사고종류'만을 분류모델의 입력변수로 사용하였다. 총 32개(6개 수치형, 26개 범주형)의 입력변수와 종속변수 '사망사고'에 대하여, 의사결정나무, 랜덤포레스트, XGBoost 모델을 통해 분류모델을 학습하였다. 사망사고와 비사망사고의 클래스 불균형을 해결하기 위해 5:5의 비율로 언더샘플링(undersampling)하였다. 성능평가 척도로는 Precision과 Recall의 조화평균인 F1 값을 사용했다. 언더샘플링 후 무작위로 선택된 70%의 학습 데이터로 학습된 모델에 30%의 테스트데이터 예측 결과, F1 score 기준 의사결정나무 0.780, 랜덤포레스트 0.837, XGBoost 0.845로 XGBoost의 성능이 가장 뛰어남을 확인하였다. Optuna[17]에서 제공하는 베이지안 최적화를 적용하여 분류모델의 하이퍼파라미터(hyperparameter)를 최적화했다. 학습된 모델에 SHAP을 적용하기에 앞서 모델 학습 후 각 모델에서의 변수 중요도를 계산하고, 상위 10%의 입력변수를 선택하였다.

ISSN 1229-6783(Print)
ISSN 2288-1484(Online)

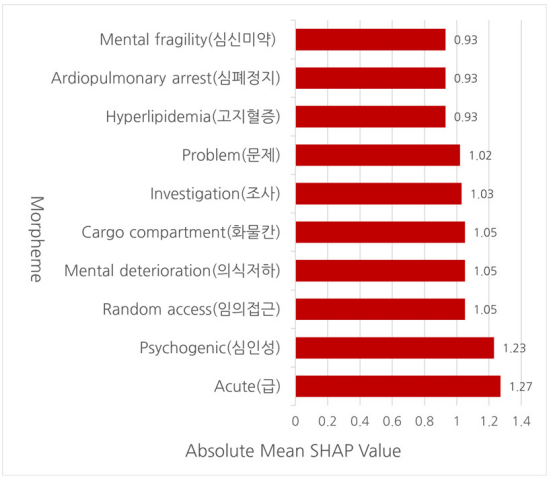


[Figure 11] SHAP analysis results of the fatal accident classifier trained by the structured data.

<Figure 11>은 학습된 XGBoost 모델에 SHAP을 적용한 결과이다. 세로축은 모델의 입력 변수를 의미하며, SHAP value의 절대 평균이 큰 변수가 위쪽에 위치한다. 그래프의 각 점은 샘플, 즉 각 사고를 나타낸다. 점의 색이 붉을수록 해당 행의 변수 입력값이 크고, 푸른 색일수록 작다는 것을 의미한다. 가로축은 입력 데이터의 각 변수가 종속변수 예측값 계산에 미친 영향을 계산한 SHAP value를 의미한다. 예를 들어, one-hot encoding에 의해 입력변수'인적사고종류'로부터 생성된 더미 변수 '인적사고종류_떨어짐(10미터 이상)'의 경우, 해당 값이 클수록 큰 SHAP value를 나타낸다. 이는 해당 사고가 10미터 이상의 높이에서 떨어진 경우에 해당하여 '1'이 되면, '0'인 경우에 비해 사망사고의 예측확률을 증가시킨다는 것을 의미한다. 따라서, '인적사고종류_떨어짐(10미터 이상)'은 사망사고 발생에 양의 영향을 미치며 그 크기는 SHAP value의 절대값으로 파악할 수 있다. '피해금액_1000만원 미만'의 경우 사망사고에 음의 영향을 미치는 것을 알 수 있다.

텍스트데이터 기반 사망사고 분류모델은 한국어 자연어처리에 특화된 KoBERT를 기반으로 학습하였다. 수집된 건설사고 14,602건의 '구체적사고원인'으로부터 클래스 불균형 문제를 해결하기 위해 사망사고와 비사망사고 각각 859건씩 추출하였다. 모델의 입력 데이터의 길이는 문장 길이 200자 이내로 전처리하였다. 총 1,718건의 데이터 중 무작위로 선택된 90%를 학습데이터로, 10%를 테스트데이터로 사용하였다. F1 score 기준 학습데이터 0.737, 테스트데이터 0.663의 성능을 나타냈다.

사망사고 예측에 영향을 미치는 요인을 파악하기 위해 학습된 분류 모델에 SHAP을 적용하였다. <Figure 12>는 학습된 텍스트 기반 사망사고 분류모델에 SHAP을 적용한 후 SHAP value 절대평균 상위 10개 형태소를 나타낸 결과로'급(급성)', '심인성', '임의접근', '의식저하', '화물칸', '조사', '문제','고지혈증', '심폐정지', '심신미약'이 사망사고에 영향을 미친 주요 요인으로 파악되었다.



[Figure 12] SHAP analysis results of the fatal accident classifier trained by the text data.

4. 결론

4.1 요약

본 연구에서는 건설사고 데이터를 기반으로 사망사고의 주요 요인 도출하기 위한 분석 절차 및 결과를 제시하였다. 첫째, 2020~2022년 국내 건설사고 사례 데이터를 통해 건설사고 발생 현황을 분석했다. 인적사고 종류, 사고객체, 사고발생 연도 · 월 · 시간, 날씨에 따른 건설재해 및 사망사고 발생 현황을 파악했다. 둘째, 통계 기반의 건설업 사망사고 요인 분석을 수행하였다. 교차분석을 통해 인적사고종류, 사고객체 등 사망사고 발생 여부와 통계적으로 유의한 상관관계를 가지는 요인을 도출하였다. 또한, 텍스트마이닝 기반 빈도분석을 통해 사망사고와 비사망사고의 '구체적사고원인' 주요 키워드를 비교 분석하였다. '추락', '장비', '공사', '고리', '해체', '체결' 등이 특징적인 사망사고 원인 키워드로 도출되었다. 셋째, 기계학습 기반 사망사고 분류모델을 학습하고 SHAP을 통해 사망사고에 영향을 미치는 요인을 파악하였다. 정형데이터와 비정형데이터를 나누어 학습하였으며, 도출된 결과를 비교 분석하였다. 사고 사전정보에 비해 발생한 사고에 대한 정확한 기록을 담고 있는 사후 정보를 활용하여 데이터 분석 결과의 신뢰성을 높였으며, 정형 및 비정형데이터로부터 도출된 다각적인 정보를 제공하였다. 제안 방법론을 활용하여 건설 현장에서 발생하는 사망사고의 주요 요인을 파악하고 재발 방지책을 수립하기 위한 기초 자료로 활용할 수 있다. 예를 들어, 본 연구의 분석 결과를 기반으로 건설현장 사망사고 방지를 위해 제안할 수 있는 개선안은 다음과 같다. 건설사고 발생 현황 분석 결과 3월과 12월 12시

전후로 건설 근로자들의 사망사고가 많이 발생하였다. 이와 같은 위험시기 및 시간대에 작업 전 보호구 점검 및 안전수칙 준수 등에 대한 캠페인을 실시할 수 있다. 텍스트 마이닝 및 SHAP 분석 결과'추락'이 사망사고에 큰 영향을 주는 것으로 파악됐다. 추락과 관련된 제도 및 안전장치를 우선적으로 보완할 수 있다. 본 연구에서 제안하는 방법론은 타 산업, 데이터, 기계학습 모델에도 적용할 수 있어 확장성이 높다. 이후 본 연구가 산업 현장에서 수집되는 양질의 안전 데이터에 적용되어, 건설업을 비롯한 다양한 산업 현장에서의 사고를 방지하는 데에 활용되기를 기대한다.

4.2 연구의 한계 및 향후 연구과제

본 연구를 발전시키기 위해서는 다음에 대한 후속 연구가 필요하다. 첫째, 정형데이터와 비정형데이터를 함께 활용할 수 있는 분석 기법이 필요하다. 조민건 등[12]과 같이 정형데이터와 비정형데이터를 하나의 모델로 학습하는 연구도 있었으나, 각 데이터가 가지는 이질적인 정보를 상보적으로 활용하지 못했다. 최근 기계학습 분야에서 연구되고 있는 멀티모달(multi-modal) 기법을 활용한다면 보다 유용한 시사점을 도출할 수 있을 것이다. 둘째, 텍스트 데이터의 표준화가 필요하다. 현재 사고사례에 포함된 텍스트 데이터는 자유 양식으로 기록되고 있어 오 · 탈자 및 부정확한 용어를 다수 포함하고 있다. 관련 기관에서 텍스트 데이터 수집 양식 및 용어를 표준화할 수 있다면 보다 나은 품질의 기계학습용 학습 데이터를 확보할 수 있을 것이다. 셋째, 사망사고 데이터 부족으로 분류모델의 예측 성능을 높이는 데에 한계가 있었다. 분류분석에서의 클래스 불균형을 해결하기 위한 방법론을 함께 적용한다면 보다 정확한 모델을 학습할 수 있을 것이다.

5. References

[1] S. G. Lee(2018), "A study on the trends of construction safety accident in unstructured text using topic modeling." Journal of the Korea Academia-Industrial Cooperation Society, 19(10): 176-182.

[2] K. Park, H. Kim(2021), "Analysis of seasonal importance of construction hazards using text mining." Journal of the Korean Society of Civil Engineers, 41(3):305-316.

[3] F. Zhang, H. Fleyeh, X. Wang, M. Lu(2019), "Construction site accident analysis using text mining and natural language processing techniques." Automation in Construction, 99:238-248.

[4] Y. C. Kim, W. S. Yoo, Y. S. Shin(2017), "Application of artificial neural networks to prediction of construction safety accidents." Journal of the Korean Society of Hazard Mitigation, 17(1):7-14.

[5] J. H. Jo(2012), "A study on the causes analysis and preventive measures by disaster types in construction fields." Journal of the Korea Safety Management & Science, 14(1):7-13.

[6] S. J. Choi, J. H. Kim, K. Jung,(2021), "Development of prediction models for fatal accidents using proactive information in construction sites." Journal of the Korean Society of Safety, 36(3):31-39.

[7] Y. Cho, Y. C. Kim, Y. Shin(2017), "Prediction model of construction safety accidents using decision tree technique." Journal of the Korea Institute of Building Construction, 17(3):295-303.

[8] Q. Xu, K. Xu(2021), "Analysis of the characteristics of fatal accidents in the construction industry in China based on statistical data." International Journal of Environmental Research and Public Health, 18(4): 2162.

[9] B. Hola, T. Nowobilski, I. Szer, J. Szer(2017), "Identification of factors affecting the accident rate in the construction industry." Procedia Engineering, 208:35-42.

[10] N. Udawatta, R. Rameezdeen(2012), Fatalities and non-fatalities in construction accidents. Deakin University. https://hdl.handle.net/10536/DRO/DU:30088163

[11] M. Amiri, A. Ardeshir, M. H. Fazel Zarandi, E. Soltanaghaei(2016), "Pattern extraction for high-risk accidents in the construction industry: A data-mining approach." International Journal of Injury Control and Safety Promotion, 23(3):264-276.

[12] M. Cho, D. Lee, J. Park, S. Park(2022), "Development of machine learning-based construction accident prediction model using structured and unstructured data of construction sites." KSCE Journal of Civil and Environmental Engineering Research, 42(1): 127-134.

[13] S. M. Lundberg, S. I. Lee(2017), "A unified approach to interpreting model predictions." Advances in

Neural Information Processing Systems, 30.

[14] T. Chen, C. Guestrin(2016, August), "Xgboost: A scalable tree boosting system." Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785-794.

[15] S. K. TBrain(2019), Korean BERT pre-trained cased (KoBERT). https://github.com/SKTBrain/KoBERT

[16] E. L. Park, S. Cho(2014, October), "KoNLPy: Korean natural language processing in Python." Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology, Chuncheon, Korea, 6:133-136.

[17] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama(2019, July), "Optuna: A next-generation hyperparameter optimization framework." Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2623-2631.

저자 소개

최 지 윤
울산대학교 산업경영공학부 학사.
관심분야 : 산업인공지능, 데이터마이닝 등

김 시 현
현 울산대학교 산업경영공학부 재학 중.
관심분야 : 산업안전, 산업인공지능 등

이 송 이
울산대학교 산업경영공학부 학사.
관심분야 : 산업안전보건, 산업안전, 품질관리, 생산관리 등

김 경 훈
현 산업안전보건공단 재직 중.
연세대학교 경영전문대학원 석사.
울산대학교 안전보건전문학과 박사.
관심분야 : 산업안전보건, 기계안전

이 수 동
현 울산대학교 산업경영공학부 조교수.
포항공대 산업경영공학과 박사.
관심분야 : 산업인공지능, 데이터분석