「정보시스템연구」 제33권 제3호
한국정보시스템학회
2024년 9월, pp. 187～205

http://dx.doi.org/10.5859/KAIS.2024.33.3.187

산재보험 빅데이터를 활용한 장해등급 예측 모델 개발*

최근호** · 김민정*** · 이정화****

〈목 차〉

Ⅰ. 서 론

산업 안전 · 보건 분야를 경험적으로 연구한 결과들의 공통적인 견해는, 업무상의 사고나 상해 · 질병이 무작위적인(random) 것이 아니며 기저에는 특정한 패턴과 경향이 깔려 있다는 것이다(Hallowell, et al., 2017; Kakhki et al., 2019; Koklonis, et al., 2021; Sarkar, et al., 2020). 여기에는 사업장 환경이나 산업 · 업무의 특성, 개인의 특성을 비롯한 다양한 요인들에 의해서 결정되는 성질이 내재한다는 연유에서이다. 때문에 사고(injury) 내지는 위험(risk)의 심각성(severity)을 예측하는 연구는 그 원인을 파헤치기 위함이나 예방의 측면에서도 중요성을 갖는다(Sarkar, et al., 2020).

예측 모델은 환경의 위험성을 경험적으로 평가할 뿐만 아니라, 위험의 파장이 크지만 간과하기 쉬운 상황(near-miss situations of high impact)을 명료하게 드러낼 수 있다(Hallowell, et al., 2017). 이에 의학 · 보건(박종호, 강성홍, 2019; Kijowski, et al., 2020; Koklonis, et al., 2021), COVID-19(Sayed, et al., 2021; Zoabi, et

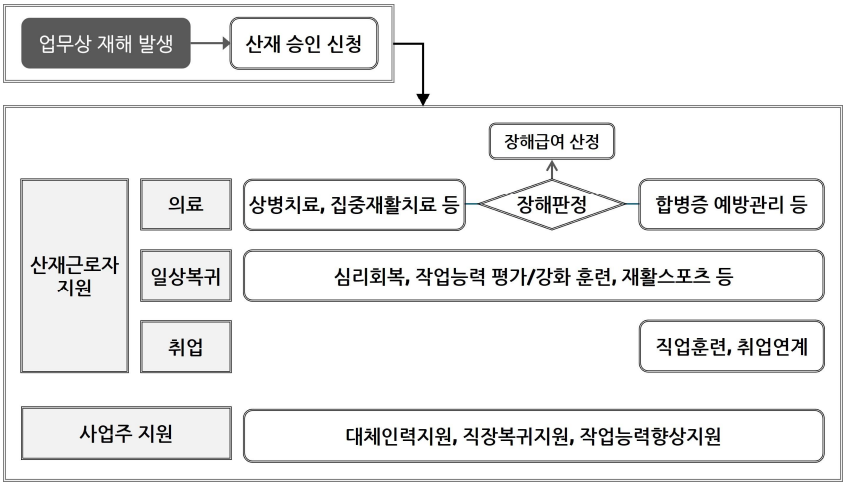
* 이 논문은 고용노동부 정책연구용역사업 「산재보험 재활시스템 선진화 방안 연구」(2021)의 일부를 발췌 및 보완한 연구임.
** 국립한밭대학교 융합경영학과 부교수, keunho@hanbat.ac.kr(주저자)
*** 한국소비자원 정책연구실 책임연구원, upmjok@kca.go.kr
**** 근로복지공단 근로복지연구원 책임연구원, jeong0112@gmail.com(교신저자)

al., 2021), 각종 사고(고창완 등, 2020; 이용범 등, 2019; 전민성 등, 2023; Sarkar, et al., 2020; Yu, et al., 2020), 스포츠 의학(Ayala, et al., 2019) 등 광범위한 분야에서 체계적인 데이터 분석 기법을 활용해 사고의 결과 내지는 심각성에 대한 예측 모델을 개발하는 시도가 지속적으로 이루어지고 있다.

마찬가지로 산재보험 분야에서도 업무상 재해로 인한 산재근로자의 장해 심각성에 대한 예측은 정책 운영상의 여러 측면에서 중요하다. 우리나라 산재보험제도의 장해등급은 장해 심각성을 1~14급 및 '무장해'의 15개 등급으로 정의 및 분류한다. 우리나라의 산재보험 적용사업장 수는 298만여 개소이며 전체 근로자 수는 2,017만여 명이고(2022년 기준; 근로복지공단, 2023), 재해율(산재보험적용근로자 수 100명당 발생하는 재해자 수의 비율)은 0.66%이다 (2023년 기준; 고용노동부, 2024).

업무상 재해의 종류(사고・질병 및 출퇴근 재해)에 따라 차이가 있지만, 대게는 업무상 재해가 발생하여 산재 승인을 받으면 치료 등의 의료서비스와 더불어 사회복귀 및 직업복귀를 지원하는 정책이 제공된다. 산재보험이 의료 및 재활 서비스를 제공하며 보상급여를 지급하는 일련의 과정에서, 장해등급은 산재보험 정책 대상의 적격성을 결정하고 이들에게 투입되는 의료서비스 및 재활프로그램 자원의 수준, 산재보험급여 및 기타 관련된 현금 급여를 산정하는 절차 등에서 주요한 기준으로 활용된다. 개괄적인 업무상 재해 처리 과정은 <그림 1>과 같다.

업무상 재해의 승인을 받은 후 요양 종결 시에 장해등급이 판정되며 산재근로자의 상당수는 이 기간이 6개월을 경과하는 것으로 파악된다[1]. 현재 근로복지공단에서는 '중증도 지수



<그림 1> 업무상 재해 발생 후 요양・보상・재활서비스 절차 개괄
(출처: 근로복지공단 홈페이지 www.comwel.or.kr, 저자 재구성)

1) 2022년 통계자료에 따르면 산재요양환자 4만9천여 명의 절반을 조금 넘는 52.4%만이 6개월 내에 요양을 종결하였고, 이 기간이 6개월~1년 미만으로 소요된 요양환자의 비율은 18.2%로 집계되었다. 요양기간이 1년~2년 내인 경우는 10.4%를 차지하는 것으로 집계되었다. 송선영과 오종은(2023)은 2018년 1월 사업주 확인 제도의 폐지가 요양 기간을 감소시키는 효과를 가져온 것으로 보았다.

(Work Ability Recovery Score; WARS)'(최근호, 이승욱, 2016)를 개발하여 요양 초기에 산재근로자의 장해 심각성을 예측하여 여러 서비스 제공 시에 대상자를 선별하는 참고자료로 활용하고 있다. 중증도 지수는 상병코드, 재해당시 연령, 상해부위, 재해유형의 4개 변수를 기반으로 산재근로자의 중증도를 극도, 고도, 중등도, 경도, 경미, 무장해의 6개 등급으로 분류하여 중요한 정보를 제공하고 있으나, 14개의 세부 장해등급을 보다 정밀하게 예측하는 측면에서 있어서 개선의 여지가 있다고 할 수 있다. 또한, scoring 방식을 사용하고 있는 중증도 지수의 특성으로 인해 4개의 변수 값 관점에서 과거에 한 번도 발생하지 않은 조합의 산재근로자가 발생할 경우, 정확한 중증도 지수 산출에 어려움을 겪을 수 있다.

이러한 배경으로 본 연구는 고용·산재 빅데이터를 활용해서 머신러닝 기법을 기반으로 업무상 재해의 심각성, 즉 장해등급 예측 모델을 개발하고 성능을 비교·평가함으로써 중증도 지수가 지니는 한계점을 보완하고 개선된 예측 도구를 제시하고자 한다. 이를 위해서 산재가 발생한 후 초기 단계인 요양 신청 시에 수집하는 다양한 정보를 모델 개발에 활용하였고, 머신러닝 기법으로 Decision Tree, DNN, XGBoost, LightGBM 4종의 예측 모델을 개발하여 성능을 비교하고 최적의 모델을 제시하였다. 예측 모델의 가장 큰 장점이자 기능이 체계적이며 과학적인 의사결정과 정책 관리·운영인 바, 산재보험 정책 실무자 및 산재근로자 등의 이해관계 집단에게 향상된 정보를 제공함으로써 정책 관리 및 운영에 유용한 방안을 제공하고자 한다.

본 연구에서 개발한 장해등급 예측 모델은 다음과 같은 특징과 강점을 갖는다. 첫째, 예측 모델을 통해 산재보험 정책 수요를 예측함으로써 효율적인 관리 및 대응이 용이해진다. 산재보험 정책은 '예측된 심각성(predicted severity)'을 통해서 더욱 효율적으로 필요한 수단과 자원을 파악하고 배분할 수 있다. 의료기관은 장해 수준을 고려한 의료 행위를 제공하도록 대비할 수 있게 되고, 행정 절차적으로 재활서비스 기관의 적합한 프로그램을 모색할 수 있다. 다른 한편으로 이해관계 당사자인 사업주와 산재근로자가 업무상 재해의 결과를 가늠하고 요양 후의 절차를 예상할 수 있다. 이처럼 전반적인 산재보험 정책의 대응 과정과 소요되는 자원을 예상할 수 있게 되면서 정책 담당자와 산재근로자 및 고용주 측이 더욱 구체화 된 정보를 공유하게 되고, 그에 따른 정책적 대응이 효율적이고 용이해진다.

둘째, 장해등급의 조기 식별(early identification)에 중점을 둠으로써 산재보험 정책의 적시성(timeliness)을 보완하여 정책의 효과성을 높인다. 노동시장 및 사회의 급속한 변화 속에서 산업구조가 세분화되고 일자리 형태나 일하는 방식, 근로관계 등도 다변화되고 있다. 업무상 재해가 발생하는 양상이 다양해졌으며 산재보험의 제도·정책 개선과정을 통해서 적용 범위도 넓어졌다. 또한 산재근로자의 치료, 재활, 직업복귀 및 생활보장의 문제에는 다양한 이해관계자와 노동시장·사회 구조적 이슈가 얽혀있다. 복잡한 역학관계 속에서 정책 문제와 정책 대응 간의 시간적 격차를 줄일 수 있는 정책 역량은 산재보험을 포함한 제반의 사회정책이 당면하는 과제이다. 본 연구의 예측

모델은 산재 발생 후 초기에 장해 심각성에 대한 정보를 생성함으로써 정책의 선제적 대응을 도울 수 있다.

셋째, 본 연구는 고용・산재 빅데이터를 활용해서 머신러닝 기법을 적용하여 모델을 개발하였다. 공공정책 분야에서 정보에 입각해(informed) 의사결정을 하기 위한 행정데이터의 역할은 이견이 없는 주제이다. 본 연구는 데이터 품질의 제고를 위해 실시한 데이터 보정 작업이 완료된 2018년부터 2020년 사이에 발생한 업무상 재해 29만여 건을 활용하였다. 해당 데이터셋은 기존에 구축된 행정데이터와 산재 발생 후 수집되는 정보를 결합한 것이다. 이러한 두 종류의 데이터를 결합함으로써 효율적인 데이터 기반 접근을 수행할 수 있다(Sarkar, et al., 2020).

넷째, 분석 데이터의 특성을 고려한 4종의 머신러닝 기법을 적용하여 예측 모델을 개발하고 예측력과 정확성을 평가하였다. 다수의 선행연구에서 머신러닝 기법은 높은 정확도와 예측력으로 우수한 성능을 증명하여 빅데이터를 활용한 예측모델 개발에 적합한 방법으로 검증되었다(Kakhki et al., 2019; Sarkar et al., 2020; Zhang, et al., 2018). 이는 광범위한 문제를 다루는데 유용하며, 특히 예측이라는 과업을 수행하기에 매우 유용한 잠재력을 갖고 있다(Sarkar et al., 2020). 이에 빅데이터와 정밀한 분석 기법을 결합하였을 때 예측에 기반한 예방 행정(preventive administration)을 도모할 뿐만 아니라 기존의 사전적 예방규제 방식을 개선하는 효과도 기대할 수 있다(안준모 등, 2022).

이상의 논의에 따라서, 본 연구는 산재 요양・보상・재활 전반에서 초기 단계에 초점을 맞추어 장해등급 예측 모델을 개발하였다. 빅데이터를 활용해서 경향이나 원인을 분석하는 탐색적・설명적 접근을 취한 선행연구와 달리, 본 연구는 더욱 정밀한 예측 기능을 통한 정책적 응용에 중점을 두어 논의를 확장하였다. 예측 모델의 정확도를 높이면서도 실용적인 모델을 찾는 것이 이 연구의 주요한 목적이자 장점이며, 장해등급에 대한 세분화된 모델을 개발함으로써 증거 기반의 정책 운영과 선제적 정책 대응을 위한 기초자료를 제공하고자 하였다.

Ⅱ. 정책 및 선행연구 검토

2.1 산재보험의 장해등급 개요 및 관련 정책

우리나라 산재보험에서의 장해는 부상 또는 질병을 치유하였음에도 노동력이 손실 또는 감소되고 그 증상이 고정된 상태를 의미한다. 산재보험은 업무상 재해로 인한 장해등급을 '1~14급' 및 '무장해'의 15개로 분류하여 진단하며 장해 1급이 가장 심각한 장해의 수준을 뜻한다[2]. 장해등급은 업무상 재해로 인한 요양이 종결되고 증상이 고정된 상태가 되었을 때 판정한다(단, 요양 종결 시 증상이 고정되지 않은 경우에는 의학적으로 6개월 내 증상이 고정될 것으로 예상될 때 판정한다). 장해등급은 의학적 진단과 의사소견서 등을 주요 자료로 참고하고, 필요한 경우 심의를 진행한다. 업무상 질

2) 산재 장해등급에 관한 기준은 「산업재해보상보험법」 제57조제2항 및 「산업재해보상보험법 시행령」 제53조제1항 및 별표 6)에서 규정하고 있다.

병은 업무상질병판정원회에서 심의를 거친다(일부 상병 제외). 2022년 기준 장해등급별 산재근로자 규모는 <표 1>과 같다.

산재보험은 장해등급에 따라 차등화된 장해급여를 지급하며, 산재근로자의 특성과 수요를 고려하는 재활서비스에서도 장해등급이 주요한 자격 요건으로 작용하는 프로그램을 운영하고 있다(<표 2>).

<표 1> 2022년 장해등급별 산재근로자 현황

장해등급	(명)	(%)	장해등급	(명)	(%)
1급	3,826	3.4	8급	1,405	1.3
2급	3,730	3.3	9급	2,388	2.1
3급	4,904	4.4	10급	3,775	3.4
4급	2,868	2.6	11급	9,631	8.6
5급	9,266	8.3	12급	7,267	6.5
6급	15,644	14.0	13급	7,582	6.8
7급	18,431	16.5	14급	22,400	20.0

출처: 근로복지공단(2023)

<표 2> 장해등급과 관련된 산재보험 정책

<table>
<tr><th>분야</th><th>정책</th><th>내용</th></tr>
<tr><td>의료</td><td>합병증 등 예방관리</td><td>● 요양 종결 이후 상병의 재발 및 합병증 방지를 위한 진료비 및 약제비 지원</td></tr>
<tr><td>보상</td><td>장해급여</td><td>● 장해등급에 해당하는 지급일수에 평균임금을 곱하여 연금 또는 일시금을 지급
<table>
<tr><td>장해 1~3급</td><td>연금으로만 지급되며 1~4년 분의 50%에 해당하는 금액을 선지급 가능</td></tr>
<tr><td>장해 4~7급</td><td>일시금과 연금 중에 선택할 수 있으며, 연금으로 선택할 시 2년 분의 ½에 해당하는 금액을 선지급</td></tr>
<tr><td>장해 8~14급</td><td>55일~495일분의 일시금을 지급</td></tr>
</table></td></tr>
<tr><td rowspan="5">재활</td><td>직업훈련 지원</td><td>● 장해 1~12급 판정자(또는 해당 등급 판정 예정자)를 대상으로 직업능력개발훈련 및 한국장애인고용공단의 위탁 훈련 실시</td></tr>
<tr><td>직장복귀 지원</td><td>● 장해 1~12급 판정자인 산재근로자를 원직장에 복귀시켜 고용을 유지한 사업주에 대하여 급여 지급
<table>
<tr><td>직장복귀지원금</td><td>장해 1~3급, 4~9급, 10~12급에 대하여 월 45~80만원을 사업주에게 지급(최대 12개월까지)</td></tr>
<tr><td>직장적응훈련비</td><td>월 45만원 이내(최대 3개월까지)</td></tr>
<tr><td>재활훈련비</td><td>월 15만원 이내(최대 3개월까지)</td></tr>
</table></td></tr>
<tr><td>대체인력 지원</td><td>● 업무상 재해 발생 후 산재근로자의 대체인력을 고용하고 그 산재근로자를 원직장 복귀시킨 소규모 사업장 사업주에게 임금의 일부를 지원</td></tr>
<tr><td>재활스포츠</td><td>● 장해 1~14급 산재근로자의 회복 및 기능 강화를 도움</td></tr>
<tr><td>사회적응 프로그램</td><td>● 산재근로자의 사회복귀 및 직업복귀 촉진을 위하여 자기관리능력, 지역 사회적응능력, 직업적응능력 향상 프로그램 운영</td></tr>
<tr><td colspan="2">복지</td><td>● 생활안정자금 융자, 산재장학사업</td></tr>
</table>

출처: 고용노동부(2023) 및 근로복지공단 홈페이지(www.comwel.or.kr)

장해급여는 업무상 재해로 인한 노동력 상실에 대한 보상으로, 각 장해등급별로 지급 일수를 규정하고 여기에 평균임금을 곱하여 산정한다. 장해 1~3급은 장해연금, 장해 4~7급은 일시금 또는 연금 중에서 지급 방식을 선택할 수 있다. 장해 8~14급은 일시금으로 지급된다.

재활서비스 중 직업훈련사업 및 직장복귀지원사업은 장해 1~12급 판정자가 재취업과 원직장 복귀를 위한 지원을 받는다. 이 때 요양 기간 중에도 지원이 필요하다고 판단되는 경우는 장해등급 결정 전에 해당 구간의 장해등급으로 예상된다는 의학적 소견서를 토대로 자격을 부여받을 수 있다. 이 경우 장해 1~12급 판정이 기본적인 조건이지만, 요양 및 치료 기간 중에도 필요한 경우에는 해당 장해등급으로 예상되는 의학적 소견을 근거로 서비스 수급의 자격을 부여하는 것으로 명시되어 있다. 또한 재활스포츠 프로그램은 장해 1~14급인 경우에 참여할 수 있으며 재해 부위의 회복과 기능 강화를 위한 지원이 마련된다. 근로자 복지 분야에서 저리로 융자를 실행하는 '산재근로자 생활안정자금'의 경우, 월평균 소득이 중위소득 이하이면서 장해 1~9급자인 경우 대상에 포함된다.

2.2 산재보험 빅데이터 분석 및 머신러닝 기법 활용 선행연구

2.2.1. 업무상 재해 심각성 분석 및 예측과 관련한 선행연구

산재보험 데이터를 분석한 연구들 중에서 먼저 업무상 재해의 심각성에 대한 예측 모델을 개발한 연구를 살펴보겠다. Sarkar, et al(2020)은 머신러닝 기법을 적용해서 상해 정도에 대한 예측 모델을 개발하고 모델별 성능을 비교하였다. SVM, ANN 등 6종의 예측 알고리즘을 적용하여 인도에 소재한 철도 공장의 2010~2013년 데이터 1,897건을 분석하였다. Hallowell, et al(2017)은 중력 에너지 방출량에 따른 산재 심각성을 예측하였다. 미국의 281개 다국적 건설 단체와 산업안전 및 British Columbia 주의 산재보험 기관으로부터 확보한 데이터를 결합하고 산재 505건의 데이터셋을 구성하여 예측 모델을 생성하고 정책적 시사점을 제시하였다. Koklonis, et al.(2021)은 그리스의 Metaxa Cancer 병원에서 2014~2019년 중 발생한 의료시설 종사자의 업무상 재해 476건을 머신러닝 기법으로 분석하고 예측 모델을 개발하였다. 업무상 재해를 떨어짐・찔림 등의 5종으로 구분하여 머신러닝 기법을 적용한 MLP, BN, k-NN, NB 모델의 예측력을 비교하였다. 여기서 MLP 모델이 가장 우수한 예측력 및 정확성을 보고한 것으로 나타났다. Kakhki et al.(2019)은 미국 중서부 지역의 농업 분야에서 2008~2016년 간 발생한 3만 3천여 건의 업무상 재해 데이터를 활용하여 업무상 재해 예측 모델을 개발하였다. RBF 등의 머신러닝 기법을 적용하였고 상해 부위, 연령, 재해 부위 등의 요인을 사용하여 정확도 92-98% 수준의 예측 모델을 개발하였다.

업무상 재해자 특성을 분석한 연구로 장태용과 성윤정(2022)은 2022년 7~11월 간 국세청, 근로복지공단, 건설협회의 데이터를 활용해서 산재근로자의 재활 및 치료와 보험급여 수급 현황을 분석하였다. 한편 산재보험 정책 운영에 관한 선행연구로 신슬비(2022)는 2019~2021년에 발생한 뇌심혈관계 질병자 7천2백여 건의

산재보험 데이터를 활용해서 머신러닝 기반의 산재승인모형을 추정하고 승인 요인을 분석하였다. 안준모 등(2022)은 건설업종의 7,467개 산재보험료 청구 데이터에 머신러닝의 랜덤 포레스트 기법을 적용하여 산재・고용보험에 대한 허위 청구의 양상을 분석하였다. 유동희 등(2022)은 지리정보시스템을 적용하여 산재보험 빅데이터를 분석하였다. 시도별로 2015~2017년의 산업재해 발생 양상과 지정요양기관 특성 등을 분석하고, 중증도 지수와 산재 취약지수를 산출함으로써 인공지능 기술을 지리적으로 활용한 사례를 제시하였다.

2.2.2. 상해 심각성 분석 및 예측과 관련한 선행연구

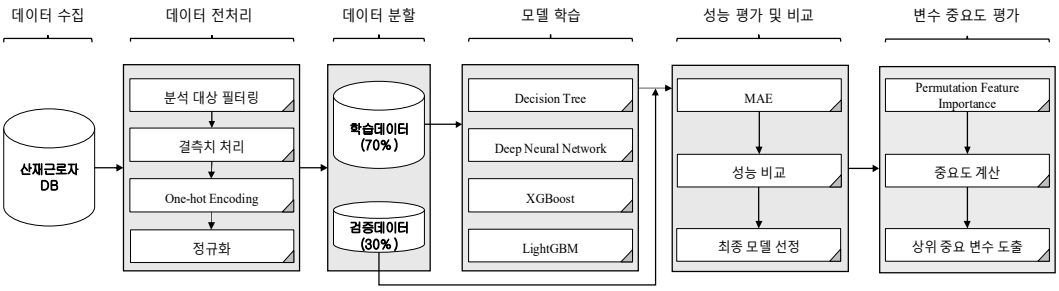
다음으로는 상해의 심각성에 대해서 머신러닝 기법을 적용한 선행연구를 살펴보았다. 상해 심각성 예측 모델을 개발한 선행연구로 고창완 등(2020)은 교통사고로 인한 상해의 심각성에 대한 예측 모형을 개발하고 모델별 성능을 비교하였다. 도로교통공단 및 교통안전정보관리시스템에서 2015~2017년 간 발생한 전국의 교통사고 67만여 건을 의사결정나무 등의 여러 가지 데이터 마이닝 기법으로 분석하였다. 이용범 등(2019)은 미국 교통부 교통안전국의 NASS-CDS 데이터베이스에서 2011~2015년 간 발생한 교통사고 4만5천여 건의 데이터를 추출하여 머신러닝 및 통계 기법을 적용한 상해 등급 예측 모델을 개발하고 정확도 등의 성능을 검토하였다. 해외 사례로 Yu, et al.(2020), Mafi, et al.(2018), Zhang, et al.(2018)은 공통적으로 차량 사고로 인한 부상 심각성 예측 모델을 다수 개발하고 모델별 성능을 비교・검토하였다. Yu, et al.(2020)은 워싱턴 교통안전국의 2010~2016년 간 발생한 차량 사고 데이터 3만 1천여 건을 활용하였다. Mafi, et al.(2018)은 미국 플로리다 주의 마이애미에서 2008~2012년 동안 발생한 5년치 사고 데이터3만 2천여 건을 분석하였으며, Zhang, et al.(2018)은 미국 플로리다 주에서 발생한 차량 충돌 사고 5천 4백여 건을 분석하였다.

또한 상해 치료와 관련한 예측 모델 개발 사례로 이덕규 등(2023)은 교통사고 후 경상환자들의 입원기간 예측 모델을 개발하고 모델별 성능을 비교하였다. 이를 위해서 2020년 건강보험심사평가원으로 자동차보험 심사가 청구된 자료 17만여 건을 머신러닝 기법으로 분석하였다. 박성호와 강성홍(2019)은 2006~2015년 간 질병관리본부 자료 2만 5천여 건을 머신러닝 기법으로 분석하여 중증도 보정 재원일수 예측 모델을 개발하였다.

Ⅲ. 연구 방법

3.1 연구의 흐름

본 연구의 흐름은 <그림 2>와 같다. 먼저 근로복지공단의 2018~2020년 요양종결자 및 재활서비스 결제 건을 취합하여 산재근로자 데이터베이스를 구성하였다. 이후 분석 기준에 맞는 분석 대상을 선별하고, 결측치를 처리한 후, One-hot Encoding 및 정규화(Min-Max Normalization)를 수행하는 데이터 전처리 작업을 진행하였다. 이러한 과정을 통해 구축한 분석 데이터셋을 학습데이터(70%) 및 검증데



<그림 2> 연구 프레임워크

이터(30%)로 분할 후, 4종류의 알고리즘을 이용하여 모델을 학습하고 모델별 성능을 평가·비교하였다. 추가로 Permutation Features Importance(PFI) 방법을 적용해서 예측 모델의 성능에 대한 변수별 상대적 기여도를 추정하고 중요 변수를 도출하였다.

3.2. 분석 데이터

본 연구에서 분석한 데이터는 업무상 재해로 인한 요양 신청 시에 제출하는 '산재요양급여 신청서'를 통해서 수집되는 근로복지공단의 2018~2020년 산재요양종결자 행정데이터이다. 완전한 구성의 고용·산재 데이터가 구축되는 데 소요되는 시간을 고려할 때, 이는 연구 활용이 가능한 비교적 최근의 데이터라 할 수 있다. 사망자, 소음성난청, 진폐 및 석면폐증, 만성폐쇄성폐질환, CS_2 질환자, 불법외국인근로자와 변수에 따른 결측치를 제외하는 전처리 과정을 거쳐서 최종적으로 290,157명의 데이터를 분석하였다. 학습데이터와 검증데이터는 각각 70%(203,110명) 및 30%(87,047명)로 무작위 분할하였다. 학습데이터에 4가지 머신러닝 기법을 적용하여 모델을 개발하였고, 개발된 모델에 검증데이터를 넣어 장해등급을 예측한 후, 실제 장해등급과 비교하여 모델의 성능을 평가하였다. 모델 개발 및 성능 평가는 python 프로그램을 이용하여 수행하였다.

3.3. 머신러닝 기법

3.3.1. Decision Tree Regressor

의사결정나무는 최대한 균일한 상태의 부분집합을 찾는 것을 목표로 학습해 나가는 알고리즘으로, 모델 학습에 사용하는 모든 변수들에 대해 변수별로 변수값에 따라 데이터를 분할하고 가장 균일한 상태의 부분집합을 생성해주는 변수를 선택하고, 이를 반복적으로 수행함으로써 트리를 형성해 나간다.

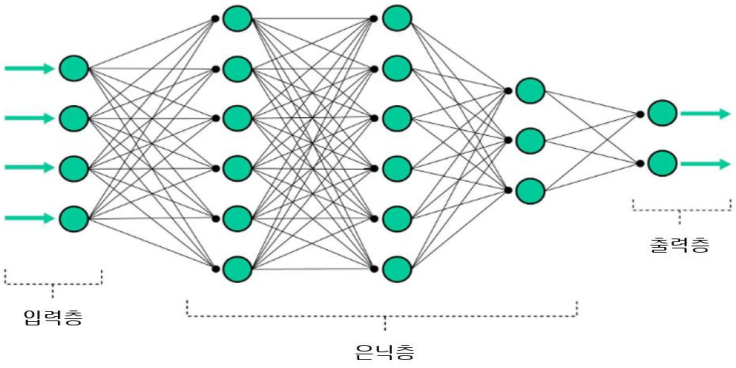
3.3.2. DNN(Deep Neural Network)

DNN은 스스로 학습하고 개선하는 대규모 신경 네트워크로서, 다층구조 형태의 신경망을 기반으로 하는 머신 러닝의 한 분야로 다량의 데이터로부터 높은 수준의 추상화 모델을 구축하고자 하는 기법이다. 신경망은 상호 연결된 일련의 노드로 구성되며, 이는 뉴런을 나타내는

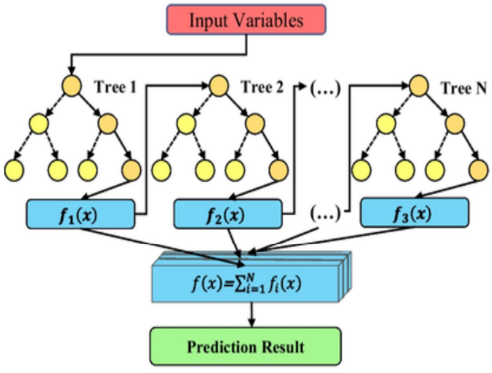
데 각 노드에는 가중치가 있으며, 가중치는 해당 노드가 네트워크의 출력에 얼마나 많은 영향을 미치는지를 결정하는 값으로, 가중치는 네트워크가 데이터를 학습하면서 시간이 지남에 따라 조정된다(<그림 3>).

3.3.3. XGBoost

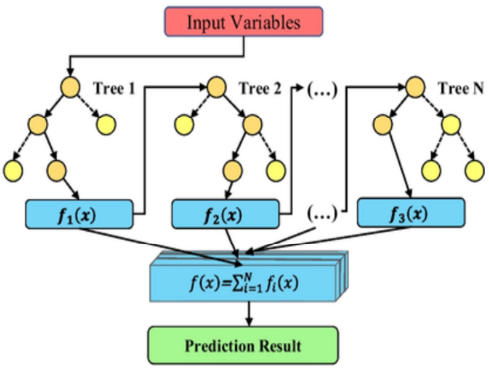
XGBoost는 속도와 정확도 면에서 GBM (Gradient Boosting Machine)을 개선한 부스팅 계열의 앙상블 알고리즘이다. 기본 학습기를 의사결정나무로 하며, 잔차(gradient)를 이용하여 이전 모델의 약점을 보완하는 방식으로 학습해 나간다. 기존의 GBM은 모든 변수들에 대해 모든 가능한 분기점들의 Information Gain을 추정하기 위해 모든 인스턴스를 사용하게 되는데, 모델 학습 시 계산의 복잡도는 변수와 인스턴스의 수에 비례하게 된다. XGBoost는 Split Finding 알고리즘을 통해 계산해야 할 분기점의 수를 줄이고 병렬 학습을 가능하게 함으로써, 속도 측면에서의 향상을 가져왔으며, 규제(regularization)를 추가한 손실함수를 통해 과잉적합을 완화함으로써 정확도 측면에서의 향상을 가져왔다(<그림 4>).



<그림 3> DNN(Deep Neural Network) 개요



<그림 4> XGBoost 개요(Wang, et al., 2023)



<그림 5> LightGBM 개요(Wang, et al., 2023)

3.3.4. LightGBM

LightGBM은 모델 학습 시 사용하는 인스턴스와 변수의 수를 줄이기 위해 GOSS (Gradient-based One-Side Sampling) 알고리즘과 EFB (Exclusive Feature Bundling) 알고리즘을 제안하였다. GOSS는 큰 잔차를 지닌 인스턴스(덜 학습된 인스턴스) 유지하고, 작은 잔차를 지닌 인스턴스(잘 학습된 인스턴스)를 무작위로 제거함으로써, 학습 시 사용되는 인스턴스의 수를 줄이는 방법이다. EFB는 변수들이 동시에 0이 아닌 값을 지니는 경우가 적고, 많은 변수들이 배타적이기 때문에 배타적인 변수들을 bundling 함으로써, 학습 시 사용하는 변수의 수를 줄이는 방법이다. LightGBM은 GOSS와 EFB 방법을 통해 XGBoost의 학습 속도를 더욱 개선하였다(<그림 5>).

본 연구에서 사용한 예측모델의 파라미터는 <표 3>과 같다.

3.4. 변수

본 연구의 타겟변수는 산재보험의 장해등급 체계인 '1~14급' 및 '무장해'이다. 장해등급을 카테고리 변수로 정의할 경우 모델의 성능 평가 시 카테고리 일치 여부만을 판단하기 때문에 크기를 지니는 장해등급의 특성을 잘 반영하지 못해 정확한 성능 평가가 어렵고 실무활용도 측면에서 제한이 있다. 때문에 본 연구에서는 타겟변수인 장해등급을 연속형 변수로 정의하였으며, '무장해'는 '15'로 정의하였다. 본 연구의 모델에서 장해등급의 예측값이 '15' 이상일 경우는 '15'로, 그 값이 '1' 이하일 때는 '1'로 정의하였다. 또한 예측값이 소수점 아래의 값을 가질 경우는 소수점 첫 번째 자리에서 반올림, 올림, 또는 버림을 하였으며 각각의 경우에 따른 정확도를 비교하였다.

독립변수는 <표 4>와 같다. 장해 수준을 예측하기 위한 변수로는 업종과 같은 사업장 특

<표 3> 예측 모델 개발 알고리즘 및 하이퍼 파라미터

❖ Decision Tree Regressor
- Criterion: Gini index
- Minimum number of samples in leaf nodes: 30

❖ DNN(Deep Neural Network)
- Batch size: 100
- Epoch: 10
- Kernel initializer: Xavier (glorot_uniform)
- Number of hidden-layers: 5
- Number of nodes in each hidden-layer: 500
- Drop-out rate: 0.5
- Activation function: relu
- Optimizer: adam
- Learning rate: 0.001
- Loss function: Mean absolute error

❖ XGBoost(Extreme Gradient Boosting) Regressor
- Number of estimators: 300
- Learning rate: 0.01
- Max depth: 10

❖ LightGBM(Gradient Boosting Machine) Regressor
- Max depth: 25
- Minimum number of child samples: 30
- Subsample: 0.8
- Number of leaves: 30
- Number of estimators: 200
- Learning rate: 0.1

<표 4> 장해등급 예측 모델 사용 변수

영역	변수명	변수 설명
사업장 특성	상시근로자 수	상시근로자 수(명)
	사업 종류	산재보험법 분류에 의한 10개의 사업 종류 (광업, 제조업, 전기・가스・증기・수도사업, 건설업, 운수・창고・통신업, 임업, 어업, 농업, 기타의 사업, 금융 및 보험업)
고용 특성	종사상 지위	상용직, 임시직, 일용직
	고용 형태	정규직, 비정규직
	고용 기간	채용일부터 산재 발생일까지의 기간(일)
재해 특성	재해 발생 형태	사고성 재해, 질병성 재해, 출퇴근 재해
의료 정보	주상병 코드	한국표준질병・사인분류(KCD)를 기본으로 하는 산재근로자의 주된 상병에 대한 코드
	주상병 부위	주상병 코드의 부위 (가슴・등, 귀, 눈, 다리, 두부, 목, 발・발가락, 복부, 복합부위, 비뇨・생식기관, 소화기관, 손・손가락, 순환기관, 신경계통, 안면부, 엉덩이, 전신, 팔, 허리, 호흡기관, 기타)
	중증도 지수	극도, 고도, 중등도, 경도, 경미, 무장해
개인 특성	연령	재해 발생 당시의 연령(세)
	성별	남성, 여성

성(최근호, 이승욱, 2016; Kakhki, etl al., 2019; Hallowell, et al., 2017), 고용 특성(신슬비, 2022; Sarkar, et al., 2020; Yu, et al., 2020), 산재근로자의 개인 특성(고창완 외, 2020; 박종호, 강성홍, 2019; Koklonis, et al., 2021; Mafi, et al., 2018)과 의료 정보(이용범 외, 2019; 최근호, 이승욱, 2016; Zoabi, 2021) 등을 꼽을 수 있다. 업무상 재해가 발생한 후 초기에 수집되는 정보를 중심으로, 실무자 회의 및 선행연구 검토를 거쳐 장해등급 예측에 영향을 미칠 가능성이 높은 변수를 선택하였다.

Ⅳ. 실험 결과

4.1 예측 모델의 성능 비교

Decision Tree, DNN, XGBoost, LightGBM의 예측 모델을 추정한 결과는 <표 5>와 같다. 예측 모델의 정확도는 실제 장해등급과 예측 장해등급의 절대값 차이의 평균을 의미하는 MAE(Mean Absolute Error)로 평가하였다. 예측 모델의 출력값은 실수이기 때문에 정수값(1~15)을 가지는 장해등급과 비교하기 위해, 예측 모델의 출력값을 반올림, 올림, 버림하여 정수값으로 변환 후 최종 예측값을 도출하였으며, 1보다 작은 출력값은 1, 15보다 큰 출력값은 15

<표 5> 모델 정확도

		Decision Tree	DNN	XGBoost	LightGBM
MAE	반올림	1.0507	0.7276	1.2557	0.7935
	올림	1.0497	0.7559	0.9144	0.7375
	버림	1.0517	1.1598	1.5846	1.1970

로 변환하였다. Decision Tree와 boosting 계열의 알고리즘인 XGBoost와 LightGBM은 출력값을 올림했을 때의 정확도가 가장 높았고, 버림했을 때의 정확도가 가장 낮았으나, DNN은 반올림했을 때의 정확도가 가장 높았고, 버림했을 때의 정확도가 가장 낮았다.

알고리즘별로 정확도를 비교해 보면, DNN 알고리즘을 사용한 예측 모델이 0.7276의 가장 낮은 MAE 값을 보였다. 이는 곧 실제 장해등급이 본 연구의 모델이 예측한 장해등급의 0.7276 범위 내의 값이 될 가능성이 크다는 것을 의미한다. 따라서 본 연구에서는 DNN 알고리즘이 가장 높은 정확도를 갖는 기법으로 추정되었다.

무장해를 포함한 총 15개 등급으로 구성된 장해등급에 대한 예측 오차가 평균 약 0.7임을 고려할 때, 현재 근로복지공단에서 활용 중인 중증도 지수 기반의 6개 등급 예측과 비교하여, 향후 장해등급을 보다 정밀하게 예측할 수 있음을 시사한다. 이는 요양 초기 단계에서 산재근로자의 장해 심각성을 예측함으로써, 서비스 제공에 있어 더욱 신뢰성 높은 대상자 선별이 가능할 것으로 기대된다.

4.2 변수 중요도 평가

본 연구에서는 예측 모델에서 변수의 중요도를 평가하기 위해 Permutation Feature Importance 알고리즘을 적용하였다. 분석 결과, 장해등급 예측에 중요한 변수 범주는 '중증도 지수', '주상병 코드', '주상병 부위', '재해 발생 당시 연령', '사업 종류'로 도출되었다(<표 6>). 예측 모델 성능에 가장 큰 영향을 미치는 변수는 '중증도 지수_무장해'였으며, 이는 산재근로자 중 무장해자의 비율이 가장 높다는 점에서 기인한 결과로 해석된다. 이어서 '중증도 지수_극도'와 '중증도 지수_고도' 역시 예측 성능에 큰 영향을 미치는 변수로 나타났는데, 이는 산재근로자 중 극도 및 고도의 비율이 상대적으로 낮기 때문으로 분석된다. 중요도가 높은 상위 3개의 변수가 모두 중증도 지수와 관련된 변수인 것은, 중증도 지수가 장해등급의 범위를 나타내는 종합 지표로서 예측에 핵심적인 역할을 하기 때문으로 해석할 수 있다.

또한, 산재의 의료적 특성을 보여주는 주상병 코드 및 주상병 부위 관련 변수들과 산재근로자의 개인적 특성을 반영하는 변수인 재해 발생 당시의 연령이 예측 모델의 성능에 큰 기여를 한 것으로 확인되었으며, 이어서 산재 발생 사업장의 특성을 나타내는 '사업 종류'가 중요한 변수로 나타났다.

반면, 상시근로자 수, 종사상 지위, 고용 형태, 고용 기간, 재해 발생 형태, 성별 등 변수들은 예측 모델 성능에 상대적으로 미미한 영향을 미치는 것으로 나타났다.

<표 6> 변수 중요도 평가: Permutation Feature Importance(PFI) 분석 결과 (상위 20개)

순위	변수명	중요도
1	중증도 지수_무장해	0.058587
2	중증도 지수_극도	0.034629
3	중증도 지수_고도	0.025713
4	주상병 코드_손목 및 손의 외상성 절단	0.017481
5	주상병 부위_허리	0.016250
6	주상병 코드_요추 및 골반의 관절 및 인대의 탈구, 염좌 및 긴장	0.013575
7	중증도 지수_중등도	0.004938
8	주상병 코드_늑골, 흉골 및 흉추의 골절	0.004264
9	주상병 부위_가슴, 등	0.003974
10	재해 발생 당시 연령	0.003665
11	중증도 지수_경미	0.003097
12	주상병 코드_요추 및 골반의 골절	0.003096
13	주상병 코드_대퇴골의 골절	0.002659
14	주상병 코드_손목 및 손의 열린 상처	0.002519
15	주상병 코드_발목을 제외한 발의 골절	0.002436
16	사업 종류_기타의 사업	0.002179
17	사업 종류_건설업	0.001765
18	주상병 코드_뇌경색증	0.001589
19	주상병 코드_발목을 포함한 아래 다리의 골절	0.001563
20	주상병 코드_손목 및 손 부위의 근육 및 힘줄의 손상	0.001544

Ⅴ. 결론 및 시사점

5.1. 연구결과 요약 및 시사점

본 연구는 고용·산재 데이터 29만여 건을 활용하여 머신러닝 기법을 적용해서 업무상 재해의 심각성에 대한 예측 모델을 개발하였다. 산재보험은 예측 모델을 기반으로 산재근로자의 향후 예상되는 장해등급을 사전에 예측함으로써, 향상된 선제적 대응을 지원하고 더욱 효과적이며 효율적인 정책 운영을 도모한다. 장해등급 15개 구간을 타겟변수로 정의하여 예측 모델을 개발한 결과, DNN에서 실제 장해등급과 예측 장해등급 간의 차이가 가장 적은 수준으로 추정되어 정확도가 가장 높은 알고리즘이라 평가할 수 있었다. 또한 PFI 평가를 통해서 의학 진단뿐만 아니라 개인 및 사업장의 특성이 예측 모델 생성에 중요한 요인으로 작용하였음을 검증하였다.

본 연구는 다음과 같은 학술적·정책적 시사점을 갖는다. 첫째, 산재 발생 초기 단계에 수집되는 정보를 기반으로 예측 모델 생성이 가능한 사례를 제시하였다. 사업장 특성, 고용 특성, 재해 특성, 의학적 특성, 개인 특성의 다섯 가지

영역의 변수를 활용해서 유용한 예측 모델을 개발하였으며, 이는 곧 업무상 재해가 체계적 경향성을 내재한다는 선행연구와 같은 맥락의 결과를 보여주었다. 여기서 PFI를 통해 장해등급 예측 모델을 생성할 때 의학적 진단 관련 변수(주상병 코드, 주상병 부위, 중증도 지수)와 더불어 개인 특성(재해 발생 당시 연령)과 사업장 특성(사업 종류)이 예측 모델의 생성에서 높은 중요도를 갖는 요인으로 추정된 결과는 큰 의미가 있다. 이는 다수의 선행연구에서 짚은 바와 같이 산재근로자의 인구・사회경제적 특성과 사업장 및 업무의 특성이 업무상 재해의 심각성과 체계적으로 연관되어 있음을 밝히는 결과이다.

둘째, 본 연구의 모델은 근로복지공단에서 사용 중인 중증도 지수를 고도화함으로써 더욱 세분화된 정보를 구축하는 방안을 제시하였으며, 이를 산재보험 정책의 다양한 영역에서 응용할 수 있는 가능성을 넓혔다. 예측 모델의 궁극적인 활용 목표는 장해 수준에 대한 조기 진단과 식별을 통해 산재보험 실무자, 의료 전문가, 재활기관 등의 산재보험 정책 담당자와 산재근로자에게 앞으로의 단계를 이행하는데 도움이 되는 필요한 정보를 제공하는 것이다. 이를 정책대상자에 대한 치료, 재활서비스 및 프로그램의 운영, 직업복귀 계획 수립의 기초자료로 활용하고, 요양-재활-직업복귀로 이어지는 정책 과정의 효율성과 효과성을 제고함으로써 정책대상자에 대한 실무자의 이해를 높이며 업무 부담 경감에 기여하는 효과를 기대할 수 있다. 또한 직업복귀, 맞춤형 재활서비스 계획 수립 등의 정책에서도 과학성과 객관성을 보강하는 방안으로 응용할 수 있을 것이다.

셋째, 우리나라에서는 아직 업무상 재해로 인한 장해의 심각성을 예측하는 연구가 활발하지 않은 편이다. 이러한 점에서 본 연구는 학술적으로도 의미 있는 결과물이라 평가할 수 있으며, 학술적 탐구에 그치지 않고 고용・산재 행정데이터의 연구 및 실무적 활용도를 높이고자 노력하였다. 본 연구의 시도는 사회정책 및 산업안전보건 분야에서 다양한 단계와 주제로 옮겨갈 수 있다. 주기적으로 데이터를 관리하고, 체계적인 데이터 연계 시스템을 구축하며, 선제적 대응이 필요한 사안에 대한 예측 모델을 개발하여 정책 문제를 관리하는 작업은 공통적으로 요구되는 노력이기 때문이다. 이와 같이 본 연구는 빅데이터 관리와 양적 연구의 중요성을 강조하고 향후 그 영역을 확장할 수 있는 예시를 제공하였다.

5.2. 연구의 한계 및 제언

본 연구의 한계를 꼽자면 무엇보다도 행정절차 상 수집되는 정보에 따라서 분석하는 데이터 범위의 제약이 있다는 점이다. 그리고 요양급여신청서 정보와 기존에 구축된 고용・산재 행정데이터에서 연구 목적을 위해 가공 및 활용 가능한 변수를 선택하는 과정에서 다수의 요인이 탈락하는 문제가 있었다. 또한, 비슷한 선행연구에서도 반복적으로 언급되듯, 분석 데이터 및 머신러닝 기법의 특성 상 데이터 분포에 따라서 다수의 너머에 있는 사례는 예측 모델에서 중요하지 않은 케이스로 처리될 수 있는 한계점이 있다. 때문에 예측 모델의 적극적인 활용을 위해서는 접근 가능한 행정정보의 범위, 그리고 사회 변화와 함께 역동적으로 변

모하는 업무상 재해의 특성을 주기적으로 고찰하고 기존의 체계를 검토・보완하는 작업이 요구될 것이다.

참고문헌

고용노동부, 산업재해현황, 고용노동부, 2024.

고용노동부, 2022년도 산재보험 사업연보, 고용노동부, 2023.

고창완, 김현민, 정영선, 김재희, “차대차 교통사고에 대한 상해 심각도 예측 연구”, 한국 ITS 학회 논문지, 제19권, 제4호, 2020, pp. 13-29.

근로복지공단, 2022년도 근로복지공단 통계연보, 근로복지공단 근로복지연구원, 2023.

박종호, 강성홍, “머신러닝을 이용한 신경계통의 질환 퇴원환자의 중증도 보정 재원일수 예측 모형 개발”, 한국보건사회연구, 제39권, 제1호, 2019, pp. 390-427.

송선영, 오종은, “산재신청 시 사업주 날인제도 폐지 정책의 효과성 분석”, 근로복지공단 근로복지연구원. 2023.

신슬비, 뇌심혈관계 질병 산재DB를 활용한 머신러닝 기반 모델 연구, 근로복지공단 근로복지연구원. 2022.

안준모, 문성욱, 이창용, “인텔리전트 규제: 인공지능을 활용한 산재보험 검증 사례를 중심으로”, 한국행정연구, 제31권, 제4호, 2022, pp. 27-50.

유동희, 정석훈, 이정화, 최근호, “산재보험 빅데이터를 활용한 산재 모니터링 지리정보시스템 개발”, 정보시스템연구, 제31권, 제2호, 2022, pp. 217-238.

이덕규, 남동현, 허성필, “자동차 사고 경상환자의 장기입원 예측 모델 개발”, 한국산업정보학회논문지, 제28권, 제6호, 2023, pp. 11-20.

이용범, 조은기, 윤창용, 박성근, “공개 교통사고 데이터베이스와 기계 학습을 이용한 탑승자의 상해 등급 예측에 관한 연구”, 전기학회논문지, 제68권, 제7호, 2019, pp. 866-871.

장태용, 성윤정, “공공 빅데이터를 이용한 산재환자의 재활 급여현황”, 고령자치매작업치료학회지, 재16권, 제2호, 2022, pp. 107-116.

전민성, 고재필, 최경주, “이미지 캡셔닝 기반의 새로운 위험도 측정 모델”, 정보시스템연구, 제32권, 제4호, 2023, pp. 119-136.

최근호, 이승욱, 산재환자 중증도 지수 개발, 근로복지공단 근로복지연구원, 2016.

Ayala, F., López-Valenciano, A., Martín, J. A. G., Croix, M. D. S., Vera-Garcia, F. J., del Pilar Garcia-Vaquero, M., Ruiz-Pérez, I., and Myer, G. D., “A Preventive Model for Hamstring Injuries In Professional Soccer: Learning Algorithms,” *International Journal of Sports Medicine*, Vol. 40, No. 5, 2019, pp. 344-353.

Kakhki, F. D., Freeman, S. A., and Mosher, G. A., “Evaluating Machine Learning Performance in Predicting Injury

Severity in Agribusiness Industries," *Safety Science*, Vol. 117, 2019, pp. 257-262.

Koklonis, K., Sarafidis, M., Vastardi, M., and Koutsouris, D., "Utilization of Machine Learning in Supporting Occupational Safety and Health Decisions in Hospital Workplace," *Engineering, Technology & Applied Science Research*, Vol. 11, No. 3, 2021, pp. 7262-7272.

Hallowell, M. R., Alexander, D., and Gambatese, J. A., "Energy-based Safety Risk Assessment: Does Magnitude and Intensity of Energy Predict Injury Severity?," *Construction Management and Economics*, Vol. 35, No. 1-2, 2017, pp. 64-77.

Kijowski, R., Liu, F., Caliva, F., and Pedoia, V., "Deep Learning for Lesion Detection, Progression, and Prediction of Musculoskeletal Disease." *Journal of Magnetic Resonance Imaging*, Vol. 52, No. 6, 2020, pp. 1607-1619.

Mafi, S., AbdelRazig, Y., and Doczy, R., "Machine Learning Methods to Analyze Injury Severity of Drivers from Different Age and Gender Groups," *Transportation Research Record*, Vol. 2672, No. 38, 2018, pp. 171-183.

Sarkar, S., Pramanik, A., Maiti, J., and Reniers, G., "Predicting and Analyzing Injury Severity: A Machine Learning-based Approach using Class-imbalanced Proactive and Reactive Data," *Safety Science*, Vol. 125, 2020, 104616.

Sayed, S. A. F., Elkorany, A. M., and Mohammad, S. S., "Applying Different Machine Learning Techniques for Prediction of COVID-19 Severity," *Ieee Access*, Vol. 9, 2021, pp. 135697-135707.

Siyuan W., Ying R., and Bisheng X., "Estimation of Urban AQI based on Interpretable Machine Learning," *Environmental Science and Pollution Research*, Vol. 30, No. 42, 2023, pp. 1-13.

Yu, H., Yuan, R., Li, Z., Zhang, G., and Ma, D. T., "Identifying Heterogeneous Factors for Driver Injury Severity Variations in Snow-related Rural Single-vehicle Crashes," *Accident Analysis & Prevention*, Vol. 144, 2020, 105587.

Zhang, J. Z. Li, Z. Pu and Xu, C., "Comparing Prediction Performance for Crash Injury Severity Among Various Machine Learning and Statistical Methods," in *IEEE Access*, Vol. 6, 2018, pp. 60079-60087.

Zoabi, Y., Deri-Rozov, S., and Shomron, N., "Machine Learning-based Prediction of COVID-19 Diagnosis based on Symptoms," *npj Digital Medicine*, Vol. 4, No. 1, 2021, pp. 1-5.

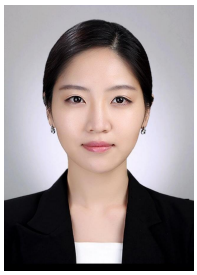
최 근 호 (Choi, Keunho)

고려대학교에서 경영학 박사학위를 취득하였다. 현재 국립한밭대학교 융합경영학과 부교수로 재직하고 있으며, 주요 관심 분야는 인공지능, 데이터 마이닝, 추천 시스템 등이다.

김 민 정 (Kim, Min Jeong)

성균관대학교에서 행정학 박사학위를 취득하였다. 현재 한국소비자원 정책연구실 책임연구원으로 재직하고 있으며, 주요 관심 분야는 시민사회, 민관협력, 거버넌스, ESG 경영 등이다.

이 정 화 (Lee, Jeonghwa)



성균관대학교에서 행정학 박사학위를 취득하였다. 현재 근로복지공단 근로복지연구원 책임연구원으로 재직하고 있으며, 주요 관심 분야는 사회보장, 노동시장 정책, 여성·가족 정책 등이다.

<Abstract>

Development of a Predictive Model for Occupational Disability Grades Using Workers' Compensation Insurance Data

Choi, Keunho · Kim, Min Jeong · Lee, Jeonghwa

Purpose

A prediction model for occupational injuries can support more proactive, efficient, and effective policy-making. This study aims to develop a model that predicts the severity of occupational injuries, classified into 15 disability grades in South Korea, using machine learning techniques applied to COMWEL data. The primary goal is to improve prediction accuracy, offering an advanced tool for early intervention and evidence-based policy implementation.

Design/methodology/approach

The data analyzed in this study consists of 290,157 administrative records of occupational injury cases collected between 2018 and 2020 by the Korea Workers' Compensation & Welfare Service, based on the 'Workers' Compensation Insurance Application Form' submitted for occupational injury treatment. Four machine learning models – Decision Tree, DNN, XGBoost, and LightGBM – were developed and their performances compared to identify the optimal model. Additionally, the Permutation Feature Importance (PFI) method was used to assess the relative contribution of each variable to the model's performance, helping to identify key variables.

Findings

The DNN algorithm achieved the lowest Mean Absolute Error (MAE) of 0.7276. Key variables for predicting disability grades included the severity index, primary disease code, primary disease site, age at the time of the injury, and industry type. These findings highlight the importance of early policy intervention and emphasize the role of both medical and socioeconomic factors in model predictions. The academic and policy implications of these results were also discussed.

Keyword: Workers' Compensation Insurance, Occupational Disability Grades, Machine Learning, DNN, Prediction Model

* 이 논문은 2024년 8월 5일 접수, 2024년 8월 31일 1차 심사, 2024년 9월 11일 게재 확정되었습니다.