연구보고서

고위험 사업장 선정 모델 개선을 통한 감독·점검 효과성 제고방안 연구

주식회사 사미텍

산업재해예방
안전보건공단
산업안전보건연구원

제 출 문

산업안전보건연구원장 귀하

본 보고서를 “고위험 사업장 선정 모델 개선을 통한 감독·점검 효과성 제고방안 연구”의 최종 보고서로 제출합니다.

2024년 10월

연 구 진

연 구 기 관 : **주식회사 사미텍**

연구책임자 : 김재두 (연구소장, ㈜사미텍 기업부설연구소)
연 구 원 : 김기형 (선임연구원, ㈜사미텍)
연 구 원 : 우은경 (연구원, ㈜사미텍)
연 구 원 : 김수경 (감사, ㈜사미텍)
연 구 원 : 이형용 (선임연구원, ㈜사미텍)
연구보조원 : 양지호 (연구원, ㈜사미텍)
연구보조원 : 박주연 (연구원, ㈜사미텍)
연구보조원 : 오민근 (연구원, ㈜사미텍)
연구보조원 : 김혜민 (대표이사, ㈜사미텍)

목 차

요약문 ····· 1

Ⅰ. 연구개요 ····· 9

1. 연구배경 및 필요성 ····· 9

2. 연구목적 ····· 18

3. 선행 연구 및 관련 기술 ····· 21

4. 연구내용 및 방법 ····· 31

Ⅱ. 연구 수행 체계 ····· 63

1. 연구 추진조직 ····· 63

2. 연구 추진 경과 ····· 65

목 차

Ⅲ. 연구추진 및 실적 ····· 81

1. 산업안전 데이터 수집 및 전처리 ····· 81

2. 기존 고위험사업장 선정 모델 및 데이터 분석 ····· 118

3. 고위험사업장 선별 모델 설계 및 개발 ····· 166

4. 산업안전 도메인 맞춤형 언어모델 개선 및 모델 성능 비교 분석 ···· 180

Ⅳ. 결론 및 제언 ····· 205

1. 결론 ····· 205

2. 제언 ····· 207

참고문헌 ····· 213

표 목차

〈표 1-1〉 SVM의 주요 특징 ····· 23
〈표 1-2〉 주요 앙상블 방법 비교 ····· 25
〈표 1-3〉 보팅의 주요 특징 ····· 26
〈표 1-4〉 배깅 및 페이스팅의 주요 특징 ····· 27
〈표 1-5〉 부스팅의 주요 특징 ····· 29
〈표 1-6〉 딥러닝 모델의 주요 특징 ····· 30
〈표 1-7〉 데이터 전처리 방법 ····· 36
〈표 1-8〉 이상치 처리 방법 비교 분석 ····· 38
〈표 1-9〉 데이터 불균형 처리 방법 비교 분석 ····· 39
〈표 1-10〉 코드값 변환 처리 비교 분석 ····· 40
〈표 1-11〉 결측치 처리 방법 비교 분석 ····· 42
〈표 1-12〉 특성분포 불균형 처리 방법 비교 분석 ····· 43
〈표 1-13〉 데이터 범주화 처리 비교 분석 ····· 44
〈표 1-14〉 특성 스케일링 처리 비교 분석 ····· 45
〈표 1-15〉 데이터 분할 처리 비교 분석 ····· 46
〈표 1-16〉 XGBoost 주요 하이퍼파라미터 ····· 48
〈표 1-17〉 Confusion Matrix 구조 ····· 49
〈표 1-18〉 범주별 데이터 학습 처리 장단점 ····· 52
〈표 1-19〉 모델별 특징 및 적용 예 ····· 53
〈표 1-20〉 모델선정 기준 ····· 54
〈표 1-21〉 앙상블 모델 적용 예 ····· 55

표 목차

〈표 1-22〉 여러 모델 학습 결과 비교 ····· 56

〈표 2-1〉 기존모델 및 데이터 분석 결과 ····· 65
〈표 2-2〉 모델 및 데이터 전처리 영향 확인 결과 ····· 67
〈표 2-3〉 학습데이터 재조정 후 학습 결과 ····· 68
〈표 2-4〉 승인통계, 사업장 위험 수준 현장평가 데이터 활용 학습 결과 ····· 69
〈표 2-5〉 근로손실일수 상관분석결과 반영 학습 결과 ····· 70
〈표 2-6〉 특성중요도가 높은 특성을 반영한 학습 결과 ····· 71
〈표 2-7〉 클러스터링 및 군집분석 결과 반영 학습 결과 ····· 72
〈표 2-8〉 REF 적용 특성수 조절을 통한 학습결과 ····· 73
〈표 2-9〉 다중분류학습 적용 결과 ····· 74
〈표 2-10〉 저위험 사업장 교집합처리 반영 학습 결과 ····· 75

〈표 3-1〉 업종별 '고위험사업장 선정 모델'에 활용된 사업리스트 ····· 82
〈표 3-2〉 제조업 특성명별 데이터 타입 ····· 83
〈표 3-3〉 서비스업 특성명별 데이터 타입 ····· 93
〈표 3-4〉 데이터 전처리 대상 특성 데이터 ····· 105
〈표 3-5〉 이상치 처리 결과, 제거된 데이터 ····· 107
〈표 3-6〉 데이터 불균형 처리 결과, 증강된 학습데이터 ····· 109
〈표 3-7〉 코드값 변환(타겟 인코딩 기법) 처리 결과 ····· 110
〈표 3-8〉 사업장 위험 수준 현장평가 결합 후, 결측치 처리 결과 ····· 111

표 목차

〈표 3-9〉 특성분포 불균형 처리 결과 ····· 112
〈표 3-10〉 데이터 범주화 처리 결과 ····· 114
〈표 3-11〉 특성 스케일링 처리 결과 ····· 115
〈표 3-12〉 데이터 분할 처리 현황 ····· 116
〈표 3-13〉 XGBoost 하이퍼파라미터 적용 범위 ····· 118
〈표 3-14〉 제조업 특성 불균형 목록 ····· 123
〈표 3-15〉 서비스업 특성 불균형 목록 ····· 134
〈표 3-16〉 업종별 근로손실일수와 특성간 상관분석 결과 상위 50개 ····· 149
〈표 3-17〉 업종별 특성별 기여도 분석 결과 상위 50개 ····· 153
〈표 3-18〉 데이터 추가 및 실험 내용 ····· 167
〈표 3-19〉 언어모델 학습 방법의 비교 표 ····· 181
〈표 3-20〉 언어모델 학습서버 구성 정보 ····· 183
〈표 3-21〉 Gemma모델 LoRA학습 설정 내용 ····· 183
〈표 3-22〉 언어모델에 사용하기 위한 데이터 선별 및 그룹화 ····· 185
〈표 3-23〉 언어모델 입력 프롬프트 예시 일부 ····· 187
〈표 3-24〉 언어모델 출력 결과 예시 일부 ····· 192
〈표 3-25〉 언어모델 및 수치적 해석의 특장점 비교 표 ····· 201
〈표 3-26〉 언어모델 및 수치적 해석의 한계점 비교 표 ····· 202

그림목차

[그림 1-1] 국내 전산업 사고사망자 및 사고사망만인율과 주요 정책 관계 (2003년~2021년) … 9
[그림 1-2] 우리나라와 주요 선진국과의 사고사망 현황(2003년~2021년) …… 10
[그림 1-3] 중대재해 감축 추진 방향 총괄 …… 11
[그림 1-4] 산업재해 발생 위험도에 따른 사업장 분포 공시 …… 12
[그림 1-5] '23년도 빅데이터분석 기반 사업장 점검.감독 계획시 …… 14
[그림 1-6] '23년도 산업재해 고위험요인 활용자료 …… 15
[그림 1-7] 사업장 선정 방식 개선안 …… 17
[그림 1-8] 데이터 기반의 고위험 사업장 선정 프레임워크 …… 19
[그림 1-9] 연구목표 …… 20
[그림 1-10] 보팅(Voting) 예측 방법 …… 26
[그림 1-11] 배깅(Bagging)과 페이스팅(Pasting) 예측 방법 …… 27
[그림 1-12] 부스팅(Boosting) 예측 방법 …… 28
[그림 1-13] 산업재해 관련 보고 단계별 메타데이터 …… 32
[그림 1-14] 산업재해 관련 보고서 양식 …… 33
[그림 1-15] 재해사례 예시 …… 34
[그림 1-16] 학습 회차별 모델 Accuracy 값 변화 …… 56
[그림 1-17] 연구 흐름도 …… 57

[그림 2-1] 연구 추직 조직도 …… 63

[그림 3-1] 특성 불균형 분석 결과 …… 121

그림목차

[그림 3-2] 데이터 불균형 분석 결과 ····· 147
[그림 3-3] 다중공선성 분석 결과 ····· 148
[그림 3-4] SHAP을 활용한 특성 기여도 분석 결과 ····· 152
[그림 3-5] 딥러닝 모델 구성 ····· 162
[그림 3-6] 언어모델 LoRA 튜닝 결과 API 구성 및 테스트 결과 ····· 200

요 약 문

● **연 구 기 간** 2024년 04월 ~ 2024년 10월

● **핵 심 단 어** 고위험사업장 선정모델, 고위험요인, 산업안전 데이터, 감독·점검 체계 지원, 언어모델, 인공지능

● **연구과제명** 고위험 사업장 선정 모델 개선을 통한 감독·점검 효과성 제고방안 연구

1. 연구배경 및 목적

○ 지난 20년간 정부가 많은 노력을 기울인 결과 국내 산업재해로 인한 사고사망만인율은 감소하였으나, 주요 선진국의 사고사망만인율을 훨씬 상회하는 것으로 경제협력개발기구 (OECD) 회원 38개국 중 34위인 것으로 조사되었고, 주요 선진국과의 사고사망 현황을 비교한 결과로 OECD 평균과 비교할 때, 거의 2배에 해당하는 수준임.

○ 정부는 2022년 '중대재해 감축 로드맵'에서 수동적·타율적 규제인 '처벌과 감독 단계'를 넘어 『자기규율 단계』에 진입하고, 『안전문화 내면화 단계』를 지향하는 대책을 제시하였음.

○ 특히 정부는 빅데이터 분석으로 고위험사업장 8만 곳을 선별하여 집중관리 하기 위한 체계를 발표하였는데, 빅데이터 분석 결과에 따라 산업재해 발생 가능성이 크다는 '위험경보서'를 최초로 교부·설명하면서, 기업의 산재 발생 위험도에 대한 경각심을 높이고 노사참여와 협력 기반의 사전 예방 체계를 갖추기 위한 방안을 발표하였음.

○ 정부는 사업장 점검.감독을 위한 계획으로 빅데이터분석으로 고위험 사업장을 선정하여 '감독대상 선정', '감독방향 설정', '교육대상 선정'을 기반으로 (1)위험성평가 특화점검, (2)일반감독, (3)특별감독 등의 기준

정보로 활용되는 정책을 수립하였음.

○ ‘23년도에 고용노동부와 한국산업안전보건공단은 최근 6년간 4,432건의 사고사망사례를 분석한 결과를 발표하였고 이를 기반으로 사업장이 산업안전에 참조할 수 있도록 하였음.

○ 산업재해에 연관된 복잡하고 다양한 정보들을 결합하고 연계하여 사업장의 위험 수준을 분류하고 선정하는 것은 결국 고위험사업장을 선정과 같은 맥락이며, 산업재해 고위험요인을 분석하여 반영하고 고위험사업장 내용 및 이력을 사업장 선정 모델 저장소에 저장하여 관리하는 방식으로 개선할 필요가 있음.

○ 현재 고위험 사업장 선정 및 관리는 산업재해 보고나 사업장 특성과 같이 후발적 정보가 반영되는 것으로서 사업장의 업무 환경이나 장비/도구/인력 등의 빈번하게 변화하는 상황을 적시에 반영하지 못하고 있기 때문에 사업장 선정의 결과의 객관성 및 신뢰성이 하락하는 한계가 있어 재해발생 가능성이 높은 사업장의 주기적 또는 실시간적 예측 체계와 산업별로 재해발생 요인과 상이한 상황정보들을 적시에 반영할 수 있는 체계지원 필요.

○ 또한, 산업재해 연관 메타정보의 필터링 방식의 고위험사업장 후보를 선정하는 방식과 복잡하고 유기적인 재해 유발 데이터들이 반영된 인공지능 모델 개발이 필요.

○ 특히, 자연어처리 관리 데이터 처리 기술의 한계로 산업재해 관련 데이터들의 유형적(비정형성) 특성과 의미적(문장, 맥락 등) 특성 처리가 부족하여 이를 해결할 수 있는 모델 개발이 필요함.

○ 본 연구에서는 데이터 기반의 감독·점검을 위한 기본 프레임워크로, 데이터와 고위험사업장 선정 모델을 이용한 “데이터기반 감독.점검 체계”를 바탕으로 [1]사업장 선정·배포, [2]지방관서 지시.감독 실시, [3]결과 분석·평가, [4]다음연도 계획 수립하는 데이터 기반 감독·점검체계를 구

축하는 것이 본 연구를 통해 달성하고자 하는 핵심 목적임.

○ 본 연구의 목적을 달성하기 위하여 고위험사업장의 선정 방식을 개선하고, 선제적인 재해발생 사업장 예측을 지원하며, 다양한 인공지능 모델 개발 및 활용을 위해 '데이터 기반 고위험 사업장 선정', '데이터 기반 감독·점검 체계 완성', '사업장 자율점검 체계 지원', '선제적 사업장 위험 분류'등 연구목표를 설정함.

○ 연구목표를 달성하기 위해 '산업안전 데이터 수집 및 전처리', '기존 고위험사업장 선정 모델 및 데이터 분석', '고위험 사업장 선별 모델 설계 및 개발', '산업안전 도메인 맞춤형 언어모델 개선 및 모델 성능 비교 분석' 등 과업을 수행함.

2. 주요 연구내용

본 연구의 목적을 위해 설정한 연구목표를 달성하기 위하여 다음과 같은 과업을 수행하였음.

1) 산업안전 데이터 수집 및 전처리

○ 산업안전 데이터를 수집하고, 전처리 과정을 통해 모델 학습에 적합한 데이터셋을 구성하는데 중점을 두었음. 이를 위해 모델에서 처리되는 이상치 처리, 코드값 변환, 결측치 처리 등 기본적인 전처리 기법 외에도, 사업장 데이터의 특성에 맞춘 추가 전처리 방법들을 적용하여 데이터를 정제하였음. 이러한 정제된 데이터셋은 모델의 기본적인 성능확보와 학습을 수행하는 필수적인 단계로 작용하였음.

2) 기존 고위험사업장 선정 모델 및 데이터 분석

○ 기존의 고위험사업장 선정 모델은 XGBoost 알고리즘을 사용하여

Confusion Matrix를 기반으로 정밀도, 재현율, F1 Score 등의 지표를 통해 모델의 성능을 평가하였음. 초기 모델 학습 및 데이터를 분석하여 안전사업장과 고위험사업장의 불균형, 특성값의 불균형, 특성간 상관관계 등이 확인되었으며, Feature Importance와 SHAP 분석을 통해 모델이 일부 특성에 과도하게 의존하는 경향도 확인하였음.

○ XGBoost 외에도 LGBM(LightGBM), CatBoost, 그리고 딥러닝 모델을 추가로 적용하여 다른 모델에서도 데이터 학습 성능을 확인하였음. 최종적으로 LGBM 모델이 정밀도, 재현율이 균형이 있으며 전반적으로 높은 성능을 보임.

3) 고위험사업장 선별 모델 설계 및 개발

○ 데이터의 적용 방식에 따라 모델 성능의 변동을 확인하기 위해 산재발생 데이터와 위험 수준 평가 데이터를 각각 또는 결합하여 모델을 학습하였음. 그 결과, 전체 데이터셋으로 학습한 것보다 성능이 향상되었음을 확인하였고, 또한 분석을 통해 상위 50개의 주요 특성만을 적용하여 학습했을 때, 성능이 더 개선되는 결과를 얻었음. 학습에 필요한 주요 특성의 선별과 데이터셋의 품질이 향후에도 중요한 역할을 할 수 있음을 확인함.

4) 산업안전 도메인 맞춤형 언어모델 개선 및 모델 성능 비교 분석

○ 기존의 수치적 해석 방식 대신, 본 연구에서는 언어모델을 도입하여 비전문가도 쉽게 이해할 수 있는 설명 방식을 개발하였음. 고위험사업장 선정 모델에서 산출한 예측 결과를 기반으로, LoRA(Low-Rank Adaptation)기법을 적용하여 학습된 언어모델을 통해 고위험사업장 판단 이유, 위험요인, 개선사항 등을 자연어로 설명함으로써 접근성과 설명력을 높임.

○ 언어모델을 활용한 해석 방식은 비전문가도 이해할 수 있는 텍스트 기반

설명을 제공하여 현장 적용 가능성을 높였으며, 이는 Feature Importance나 SHAP과 같은 수치적 해석과 함께 보완할 수 있는 도구로 활용될 수 있음.

3. 연구 활용 및 제언

○ 본 선정 모델은 기존의 사업장 감독·점검 시 생성한 데이터를 활용하여 데이터 기반 고위험 사업장 선정 모델을 통하여 실제 산업재해가 발생할 가능성이 얼마나 되는지, 발생할 수 있는지, 또는 사업장이 어느 정도의 산재위험에 노출되었는지을 판단하는데 활용할 수 있음.

○ 또한, “데이터기반 감독.점검 체계” 프레임 워크로 사업장 선정·배포, 지방관서 지시.감독 실시, 결과 분석·평가, 다음연도 계획 수립하는데 활용할 수 있음.

○ 그러기 위해서는 데이터의 양적 및 질적 확보를 통한 모델 성능을 더욱 향상시켜야 하며 양질의 데이터 확보가 필수적임. 특히 산재 발생 이력이 부족하거나 산재에 관련성이 떨어지는 특성, 결측된 데이터는 예측 성능을 저하시킬 수 있으므로, 많은 데이터를 확보할 수 있는 방안이 필요함.

○ 라벨링의 기준 설정은 모델의 예측 성능에 큰 영향을 미칠 수 있어, 안전/고위험 사업장으로 분류되는 기준이 명확하지 않거나 모호할 경우, 모델 성능의 저하로 이어질 수 있음. 조건에 부합하지 않거나 부여할 수 없는 데이터에 대한 처리 기준을 정립하여 불확실성을 최소화하고, 데이터의 일관성을 유지하도록 라벨링 기준을 개선해야 함.

○ 이번 연구에서 사용된 언어모델은 비전문가도 쉽게 이해할 수 있는 방식으로 고위험 사업장 판단 근거를 제공함으로써 현장 적용 가능성을 높였으며, 향후 더 다양한 언어모델과 인프라를 도입하여, 각 사업장의 구체

적 상황에 맞춘 맞춤형 해석을 제공할 수 있도록 발전시킬 필요가 있음.

○ 비정형 데이터는 사업장의 텍스트 기록, 보고서, 현장 평가 내용 등을 포함하며, 이를 효과적으로 처리할 수 있는 시스템 및 기존 자료를 전산화할 수 있는 방법이 필요함. 특히 AI기반 자연어 처리(NLP) 기술을 활용하여 비정형 데이터를 정형 데이터로 변환함으로써, 보다 정교한 분석과 예측이 가능해질 것임.

○ 본 연구의 결과로 데이터 기반 감독·점검체계 구축을 완성하기 위하여 산재발생 현황 모니터링 웹페이지 개발 시 분석 및 시각화 데이터로 활용할 필요가 있음.

4. 연락처

○ 주식회사 사미텍: T 042)826-7336, E acroie@samitech.kr

Ⅰ. 연구개요

Ⅰ. 연구개요

1. 연구배경 및 필요성

1) 연구배경

(1) 국내 산업재해 현황의 이해

○ 우리나라가 급속한 산업화를 이룸과 동시에 큰 국가적·사회적 문제로 떠오른 산업재해를 해결하기 위해 [그림 1-1]에서 보듯이 지난 20년간 정부가 많은 노력을 기울인 결과 2022년 국내 사고사망만인율[1]을 203년 1.24%에서 65.3% 감소한 0.43%로 감소하였음.

[그림 1-1] 국내 전산업 사고사망자 및 사고사망만인율과 주요 정책 관계 (2003년~2021년)[2]

1) 사고사망만인율 : 임금근로자수 10,000명당 발생하는 사망자수의 비율(사망자수 / 임금근로자수) × 10,000)
2) 출처 : 2022.11, '중대재해 감축 로드맵'

○ 국가가 특별한 계획과 법령 등을 도입하여 노력할 때마다 사고사망자나 사고사망만인율이 감소하는 효과를 보여주고 있으나, 중대재해의 경우 감소 수준이 정체되는 것으로 보고되고 있음.

○ 국내 산업재해로 인한 사고사망만인율은 감소하였으나, 주요 선진국의 사고사망만인율을 훨씬 상회하는 것으로 나타남. 2022년에 고용부가 발표한 '중대재해 감축 로드맵'에 따르면 경제협력개발기구 (OECD) 회원 38개국 중 34위인 것으로 조사되었고, 2020년 독일의 사고사망만인율 0.07%와 비교할 때, 여전히 우리나라의 중대재해는 높은 상황임을 알 수 있음.

○ [그림 1-2] 우리나라와 주요 선진국과의 사고사망 현황을 비교한 결과로 OECD 평균과 비교할 때, 거의 2배에 해당하는 수준임을 알 수 있음.

[그림 1-2] 우리나라와 주요 선진국과의 사고사망 현황(2003년~2021년)[3)]

○ 사고사망만인율은 2019년 0.46%에서 2022년 0.43%로 미비한 수준의 감소가 이뤄지고 있으나 더 중요한 사항은 사고사망자의 수가 감소되지 않고 850명을 기준으로 사망자 수의 증감이 반복되고 있음.

3) 출처 : 2022.11, '중대재해 감축 로드맵'

(2) 국내 산업재해 대책 방향의 전환

○ 정부는 그동안 중대재해 예방을 위해 산업안전보건과 관련된 법이나 제도를 강화하는 형태의 중대재해 예방 정책을 수립하였으나, 이런 정책이나 처벌을 강화하는 대책이 아닌 새로운 형태의 중대재해 예방정책을 수립하고 있으며, 2022년 '중대재해 감축 로드맵'에서 수동적·타율적 규제인 '처벌과 감독 단계'를 넘어 『자기규율 단계』에 진입하고, 『안전문화 내면화 단계』를 지향하는 대책을 제시하였음.

[그림 1-3] 중대재해 감축 추진 방향 총괄

○ [그림 1-3]와 같이 '중대재해 감축 추진 방향 총괄'에 제시된 핵심 전략은 단순히 법령이나 정책의 뒷받침이 아닌, "위험성평가 중심의 자기규율 예방체계 확립"으로서, 이를 위한 핵심 수단은 이전까지의 현장 점검이나 법적 체계가 아닌 예방과 재발방지를 위한 핵심수단으로 위험성평가를 개편하는 것이며, 위험성평가를 위한 방법으로 『빅데이터로 위험요인을 분석해 초고위험사업장 2만 개소를 포함한 고위험사업장 8만개소를 선별·집중 관리』를 통해 자기규율 예방체계 구축 지원을 위한 '위험성평가 특화점검'을 새롭게 도입하였음.

○ ‘22년도 산업재해 발생 기준에 따라 사업장의 산업재해 발생 가능성 위험도를 도식한 [그림 1-4]에서 보는 바와 같이 우리나라의 사업장은 대부분은 중위험 아래 놓여 있지만, 고위험과 초고위험 사업장도 분포되어 있음을 알 수 있음.

○ 특히 정부는 빅데이터 분석으로 고위험사업장 8만 곳을 선별하여 집중관리 하기 위한 체계를 발표하였는데, 빅데이터 분석 결과에 따라 산업재해 발생 가능성이 크다는 ‘위험경보서’를 최초로 교부·설명하면서, 기업의 산재 발생 위험도에 대한 경각심을 높이고 노사참여와 협력 기반의 사전 예방 체계를 갖추기 위한 방안을 발표하였음.

[그림 1-4] 산업재해 발생 위험도에 따른 사업장 분포 공시[4)]

○ 즉, 고용노동부는 그간 개인정보보호 등의 문제로 산재예방에 있어 근로복지공단의 홈페이지를 통해 비정형파일 형식으로 산업재해 사례를 제한적으로 공개하여, 사업장이 직접적인 참조나 판단을 하는데 큰 도움을 주지 못하는 상황이었으나, 그동안 축적된 정보들을 빅데이터로 구성하여 산재 발생 위험도를 미리 예측하고, 자가 점검에 활용할 수 있도록 지원함과 동시에 정책 방향 설정 등의 참조할 것이라 발표하였고, ‘23년에는 이와 관련되어 인공지능을 이용하여 산업재해를 분류하는 연구 등의 정책 연구를 시행하였음.

4) ‘23.1, 대한민국 정책 브리핑 “빅데이터로 위험요인 분석....고위험사업장 8만개 선별”, 고용노동부

(3) 중대재해 및 고위험사업장 관리

○ 20년간 지속된 정부의 정책과 관리에 따라 산업재해의 많은 부분의 개선되었으나, 2022년 우리나라의 업무상 사고사망자는 874명으로 아직도 하루 평균 2.4명의 근로자가 산업재해로 사망하고 있으며, 비슷한 경제규모의 다른 나라와 비교할 때, 위험도가 높은 산업재해는 높은 수준임.

○ 산업재해를 줄이기 위한 새로운 접근방법이 시도되고 있으며, 앞에서 기술한 바와 같이 "중대재해감축 로드맵"은 처벌·감독을 통한 타율적 규제가 아닌 안전주체들의 책임에 기반한 '자기규율 예방체계'를 중심으로 산업재해를 예방한다는 것이 주된 내용으로, 중대감축 로드맵에는 산업재해 통계와 관련하여 아래와 같은 내용을 제시하였음.

고위험 기업 대상 선정
• 산재통계(보상) 분석 등을 통해 재해 발생 경향성을 사전에 확인 후 감독 방향 설정 • 고위험 기업 자동 선정 ('23년 방향성)

○ 위 방향은 재해 발생 경향성을 사전에 확인한 후 감독 방향을 설정하고, 고위험 기업을 자동 선정한다는 것은 지방관서가 아닌 고용노동부 본부에서 사전에 감독 대상을 선정한다는 의미를 갖는 것이며, 향후 주제에 맞는 감독 대상을 일괄적으로 선정하고 이를 지방관서에 시달하는 방법이 주로 사용되는 정책의 방향을 변경시키고 있으며, 산업안전보건 분야에서 '감독'은 가장 주요한 정책 수단 중의 하나이기 때문에 감독 대상 선정 방법을 바꾼다는 것은 정책 효과에 큰 변화를 줄 수 있는 사안이며, 정부는 보다 정확한 산재통계 분석을 위해 여러 가지 노력을 하고 있으며, 최근 발표한 '산업재해 고위험요인 분석'도 그러한 시도 중의 하나임.

[그림 1-5] '23년도 빅데이터분석 기반 사업장 점검.감독 계획시[5)]

○ [그림 1-5]에서 정부는 사업장 점검.감독을 위한 계획으로 빅데이터분석으로 고위험 사업장을 선정하여 '감독대상 선정', '감독방향 설정', '교육대상 선정'을 기반으로 (1)위험성평가 특화점검, (2)일반감독, (3)특별감독 등의 기준 정보로 활용되는 정책을 수립하였음.

○ '23년도에 고용노동부와 한국산업안전보건공단은 최근 6년간 4,432건의 사고사망사례를 분석한 결과를 발표하였고 이를 기반으로 사업장이 산업안전에 참조할 수 있도록 하였음. 이 내용에서 재해의 개요, 기인물, 고위험 작업/상황, 재해유발요인, 위험성감소 대책까지 일목요연하게 정리하여 제공된 신뢰있는 정보이나, 담당자가 직접 읽어보고 스스로 사업장의 상황과 비교하고 판단하도록 지원함.

○ 아래의 [그림 1-6]은 '23년 5월에 고용노동부가 제시한 산업재해 고위험요인 활용자료임.

5) '23.5, 산업안전보건 종합계획, 고용노동부

□ 제조업 등[건설업 외 업종]

산재업종 (대분류)	산재업종 (중분류)	산재업종 (소분류)	재해개요	기인물	고위험작업 상황	재해유발요인	위험성 감소대책(예시)
제조업	전기·가스·증기·수도 사업	전기 자재·기구 제조업	2015년 9월경 OO업체 주차장에서 작업자가 물건을 적재하기 위해 지게차 운전석에 앉아 주차장내의 협소한 공간을 정차 하던 중 지게차 앞쪽 운전석 문이 열리면서 지게차 운전자가 차량 밖으로 떨어지면서 지게차 마스트에 머리를 부딪혀 사망	지게차 운전석	지게차 주차	▸ 운전석 문이 열리지 않도록 잠금장치 등 안전조치 미흡	▸ 운전석 문이 열리지 않도록 잠금장치 등 안전조치 실시
기타의 사업	시설관리 및 사업지원 서비스업	시설물 유지관리 서비스업	2015년 10월경 OO공원에서 설치된 지하주차장 출입구 계단(높이 2.4m)에 설치된 CCTV 배선을 점검하고자 작업자가 A형사다리를 이용하여 계단 중간에 올라서서 작업하던 중 균형을 잃고 떨어져 사망	사다리(각립식(A형))	사다리를 이용한 작업 또는 점검	▸ 사다리 기대 놓아가 작업, 사다리 안전 조치를 취하지 않고 작업하다 떨어짐으로 발생함	▸ 3.5m 초과 작업 시 A작업 사다리만 사용(고용노동부 고시 제2013-15호) ▸ 아웃트리거 설치 및 전환대, 안전대 등 개인보호구 착용
기타의 사업	도소매업 및 숙박업	도소매업	2015년 3월경 OO시장 3층 창고에서 작업자가 입원하였던 병원 기프트샵 확장 공사 중 문 안쪽에 설치된 가로등 센서를 점검하기 위해 사다리를 이용하여 작업하던 중 균형을 잃고 떨어져 사망	각종 프레임 구조물 및 부품	사업장 작업시설물 설치 및 점검	▸ 설치 기준을 정하지 않거나 준수하지 않은 상태로 사다리 형식 및 종류를 작업자가 자의적으로 결정	▸ 설치 기준을 이해할 수 있는 작업자용 안전보건 교육 실시 ▸ 작업계획서 작성하여 준수하도록 조치 ▸ 사다리 안전 사용 및 점검 체크리스트 작성 및 활용
기타의 사업	시설관리 및 사업지원 서비스업	기술장치종합관리사업	2015년 3월경 OO빌딩 옥상에서 옥상 창문 상부에서 작업하기 위해 발코니 외벽을 이용하여 작업하던 중 약 10.0m 아래로 추락하여 사망	입목 불가	사업장 작업시설물 설치 및 점검	▸ 작업을 위해 필요한 구조물이 파손되거나 균형을 잃기 쉬운 불안정 상태로 작업하게 설치되거나 방치됨	▸ 건축난간이 설치된 작업발판을 설치·사용 ▸ 고소작업대 사용 또는 3.5m초과한 A작업 사다리 사용(고용노동부 고시 제2013-15호) 준수 ▸ 안전대 및 안전모 지급착용 등 준수사항 이행 확인
건축물 및 토목구조물 건설업	특수건설업	특수건설업	2015년 4월경 OO빌딩에서 작업자가 엑스레이와 초음파를 검사 수리 작업을 하던 중 기계실 내부 고정부와 엑스레이와의 방호 사이에 끼여 사망	엑스레이와	유지보수 작업(정비, 보수)	▸ 설치 기준을 정하지 않거나 준수하지 않은 상태로 사다리 형식 및 종류를 작업자가 자의적으로 결정	▸ 설치 기준을 이해할 수 있는 작업자용 안전보건 교육 실시 ▸ 작업계획서 작성하여 준수하도록 조치 ▸ 사다리 안전 사용 및 점검 체크리스트 작성 및 활용
건축물 및 토목구조물 건설업	철도 및 궤도 건설업	철도 궤도건 설업	2015년 5월경 OO역 플랫폼에서 작업자가 스크린도어 센서를 점검하던 중 진입하는 열차에 치여 사망	기차 차량(열차/기차)	사업장 작업(점검)	▸ 설치 기준을 정하지 않거나 준수하지 않은 상태로 사다리 형식 및 종류를 작업자가 자의적으로 결정	▸ 설치 기준을 이해할 수 있는 작업자용 안전보건 교육 실시 ▸ 작업계획서 작성하여 준수하도록 조치 ▸ 사다리 안전 사용 및 점검 체크리스트 작성 및 활용
기타의 사업	시설관리 및 사업지원 서비스업	기술장치종합관리사업	2016년 6월경 OO빌딩에서 작업자가 건장에 전진에 스프링클러 배관부식 발 교체하고 나온 한 후 작업이 실패가 전기에서 불이 중전부에 접촉에 감전되어 사망	전기 장치부	전기 설비 작업 점검	▸ 작업계획 및 전기점검 등 조작 실시 ▸ 절연장갑 등 안전보호구 착용	▸ 전기 설비 작업 시 전기 차단 및 잠금장치 설치
기타의 사업	도소매업 및 숙박업	도소매업	2016년 7월경 OO빌딩에서 작업자가 건장의 금형이를 제거하기 위해 작업 중 사다리(A형)에 올라가 금형이를 제거 중 중심을 잃고 2.2m 아래로 추락하여 사망	사다리 각립식(A형)	사다리를 이용한 작업 또는 점검	▸ 사다리 기대 놓아가 작업, 사다리 안전 조치를 취하지 않고 작업하다 떨어짐으로 발생함	▸ 3.5m 초과 작업 시 A작업 사다리만 사용(고용노동부 고시 제2013-15호) ▸ 아웃트리거 설치 및 전환대, 안전대 등 개인보호구 착용
기타의 사업	시설관리 및 사업지원 서비스업	기술장치종합관리사업	2016년 9월경 OO공장 지하에서 작업자가 정소작업을 하던 중 무게에 의해 설치된 배관과 구조물 사이 틈을 청소하기 위해 약 1.65m 아래 계단으로 추락하여 사망	입목 불가	작업장소 정돈/청소	작업 중 '여우굴 정소의 요철부위 걸려 넘어지거나 실족하여 떨어짐	▸ 바닥에 돌출부위는 기초가 될 때까 조치 ▸ 위험 예산을 방지하기 위한 개선 실시
기타의 사업	시설관리 및 사업지원 서비스업	기술장치종합관리사업	2016년 10월경 OO빌딩 지하 주차장에서 작업자가 설비 점검을 하던 중 추락 또는 5.3m 사망	지게차 운전석	지게차 주차 설비	▸ 정비작업 중 작업자 보호구(안전모, 안전화 등) 미착용	▸ 생산 설비의 불시 가동을 방지하기 위한 LOTO(Lock Out_잠금장치, Tag Out_표찰 후 조치)프로그램 시행 ▸ 안전장비 착용 철저
기타의 사업	국가기술자격 검정	국가기술자격검정	2016년 10월경 OO기계 설계에서 작업자가 수력 가이가 제작한 조립가정 차선 차 카고트럭) 상에서 내려가기 정지작업을 하던 중 몸이 중심을 잃고 2.5m 아래 콘크로 추락하여 사망	조립 콘베이 작업 차 (그로 차)	상세한 작업 또는 계획된 작업	▸ 계획된 작업 계로 또는 조립 작업절 정해 놓지 않았거나 중균형을 잃고 떨어짐	▸ 작업계획서 도면내용기 위한 설비 설치사용 ▸ 작업 시 안전대, 안전모 등 보호구 착용
제조업	복합물류관련업	금융지원 및 기타	2016년 2월경 OO사업장 지하 창고에서 작업자가 작업지휘를 하던 중 지상 1층에 적치해 있던 리프트 운반구가 떨어지면서 작업자를 덮쳐 사망	리프트(동력차 장비)	사업장 작업(이동) 및 적치 정소	▸ 화물이 넘어지거나 적재상태가 불량한 상태로 방치	▸ 주변의 중량물에 대한 충격 방지를 위해 중량물 적치 후 작업 ▸ 운행을 지시하고 설치 후 수리 교체작업 실시
기타의 사업	시설관리 및 사업지원 서비스업	기술장치종합관리사업	2016년 3월경 OO아파트 현장에서 작업자가 작업차 사다리를 이용하여 조경수 전지작업을 하던 중 약 3.6m 아래 바닥으로 떨어져 사망	사다리(이동식/접개형)	사다리를 이용한 작업 또는 점검	▸ 사다리 기대 놓아가 작업, 사다리 안전 조치를 취하지 않고 작업하다 떨어짐으로 발생함	▸ 3.5m 초과 작업 시 A작업 사다리만 사용(고용노동부 고시 제2013-15호) ▸ 아웃트리거 설치 및 전환대, 안전대 등 개인보호구 착용
기타의 사업	도소매업 및 숙박업	도소매업	2016년 4월경 OO빌딩에서 작업자가 3층 외벽에 설치된 실외기 거치대의 설치기둥을 철점하던 중 안전난간이 설치되지 않은 옥상 바닥으로 추락하여 사망	안전난간	설비 및 지정 등 높은 장소에서의 작업	▸ 작업을 위해 필요한 구조물이 파손되거나 균형을 잃기 쉬운 불안정 상태로 작업하게 설치되거나 방치됨	▸ 건축난간이 설치된 작업발판을 설치·사용 ▸ 고소작업대 사용 또는 3.5m초과한 A작업 사다리 사용(고용노동부 고시 제2013-15호) 준수 ▸ 안전대 및 안전모 지급착용 등 준수사항 이행 확인
기타의 사업	시설관리 및 사업지원 서비스업	기술장치종합관리사업	2016년 8월경 OO아파트 재활용지 적재작업 교체 굴절 보수작업을 하기 위해 설비계에 탑승하던 중 추락하여 사망 (3.6m) 사망	입목 불가	설비 및 지정 등 높은 장소에서의 작업	▸ 안전난간이 설치되지 않은 상태로 사다리 형식 및 종류를 작업자가 자의적으로 결정	▸ 지정 작업 확인 및 작업 안전보건 교육 후 작업 ▸ 지정 장소 외에 작업 시 안전보건 조치 후 작업 ▸ 지정 설비부재 안전난간 설치

[그림 1-6] '23년도 산업재해 고위험요인 활용자료[6)]

○ 우리나라의 산업재해와 중대재해 관리 현황을 다시 한번 요약하면 다음과 같음.

6) '23.5, 산업안전보건 종합계획, 고용노동부

우리나라의 산업재해와 중대재해 관리 현황

- 우리나라는 산업재해의 예방을 위한 지속적 노력으로 20년간 크게 감소
- 전체적인 산업재해는 크게 감소하였으나 현재도 중대재해는 선진국에 비해 거의 2배 수준이라 분석
- 정부는 중대재해를 감축하기 위해 '22년 중대재해 감축 로드맵'을 발표하였고, '23년 산업안전보건 종합계획을 통해 고위험 사업장을 선제적으로 파악하여 관리하기 위한 다양한 방안 제시
- '23년도는 빅데이터 분석 고위험 사업장 선정을 통해 점검과 감독을 계획하였으며, '23년도 산업재해 고위험요인 활용자료를 배포
- 고용부나 관련 기관은 산업 특성별로 선정된 고위험 사업장을 대상으로 지도점검을 시행하고 있음

- 빅데이터 분석 기반의 최신 기술을 도입하고 있으나, 정부가 매해 점검해야 될 사업장의 규모가 20,000여개가 넘으며 동시에 이 모든 사업장을 지도·점검 하는 것은 현재의 방법으로는 어려운 상황으로 이를 해결할 적극적 방법이 필요

2) 연구 필요성

(1) 고위험 사업장 선정 방식의 한계

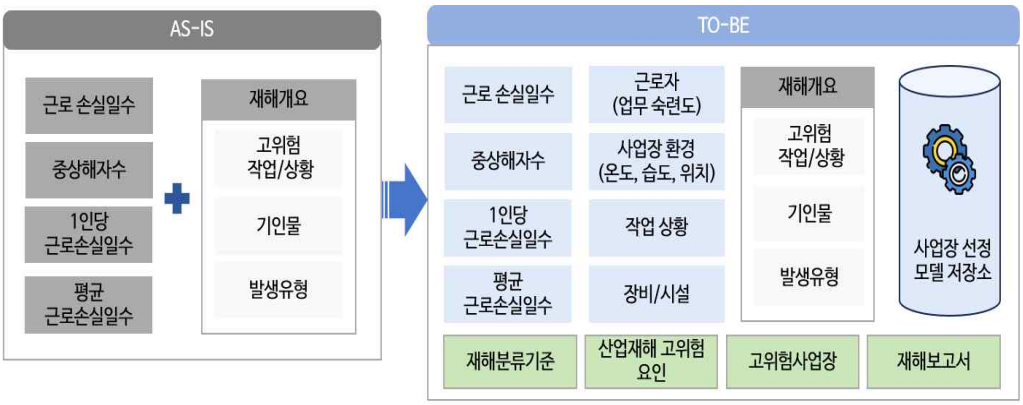
○ 산업재해에 연관된 복잡하고 다양한 정보들을 결합하고 연계하여 사업장의 위험 수준을 분류하고 선정하는 것은 결국 고위험사업장을 선정과 같은 맥락이며, 고위험사업장 선정방식은 '22년부터 개선되고 있으나 아직 기초단계임.

○ 현재 고위험사업장 후보를 도출하는 기본 정보는 근로손실일수, 중상해자수, 1인당 근로손실일수, 평균근로손일일수 4가지를 기준으로 고위험 사업장을 선정하고 있으며, 외에도 추가적으로 필요하다고 판단되는 기준이 있을 때 직접 추가하여 선정하며, 업종으로 MSDS 관련 사업장,

공정안전관리 대상 사업장, 질식위험 사업장, 고독성 화학물질 취급 사업장 등 고위험사업장 후보 목록을 구성하여 관리하고 있음.

○ 그러나, 산업재해 고위험요인을 분석하여 반영하고 고위험사업장 내용 및 이력을 사업장 선정 모델 저장소에 저장하여 관리하는 방식으로 개선할 필요가 있음.

○ 아래 [그림 1-7]은 고위험 사업장 선정 방식 개선을 도식화한 것임



[그림 1-7] 사업장 선정 방식 개선안

(2) 선제적 재해발생 사업장 예측 부재

○ 정부는 최근 빅데이터나 인공지능과 같은 기술을 활용해 중대재해를 효과적 감독·점검하고 선제적 예방의 수단으로 큰 관심을 갖고 있으며, 재해분류, 빅데이터를 이용한 사고사례 분석과 같은 기본적 연구를 '23년부터 추진하였음.

○ 현재 고위험 사업장 선정 및 관리는 산업재해 보고나 사업장 특성과 같이 후발적 정보가 반영되는 것으로서 사업장의 업무 환경이나 장비/도구/인력 등의 빈번하게 변화하는 상황을 적시에 반영하지 못하고 있기 때문에 사업장 선정의 결과의 객관성 및 신뢰성이 하락하는 한계가 있음.

○ 재해발생 가능성이 높은 사업장의 주기적 또는 실시간적 예측 체계와

산업별로 재해발생 요인과 상이한 상황정보들을 적시에 반영할 수 있는 체계지원 필요.

○ 또한, 기업이 정부 기준을 기반으로 자체규범을 마련하고 근로자와 함께 사업장 위험요인을 발굴하고 제거할 수 있도록 쉽게 사업장의 자율점검체계를 확인할 수 있는 지원 체계 필요.

(3) 고위험 사업장 선정 인공지능 모델 부재

○ 산업재해를 일으키는 상황은 매우 다양하고 복잡하며, 이런 상황들이 서로 연계되어 산업재해가 발생될 수 있고, 또 발생된 후 그 당시의 상황을 정확하게 기록하고, 재해의 원인과 결과를 결정하는 것은 더욱 어려운 일이라 할 수 있음. 그러므로 산업재해 연관 메타정보의 필터링 방식의 고위험사업장 후보를 선정하는 방식과 복잡하고 유기적인 재해 유발 데이터들이 반영된 인공지능 모델 개발이 필요

○ 특히, 자연어처리 관리 데이터 처리 기술의 한계로 산업재해 관련 데이터들의 유형적(비정형성) 특성과 의미적(문장, 맥락 등) 특성 처리가 부족하여 이를 해결할 수 있는 모델 개발이 필요함.

2. 연구목적

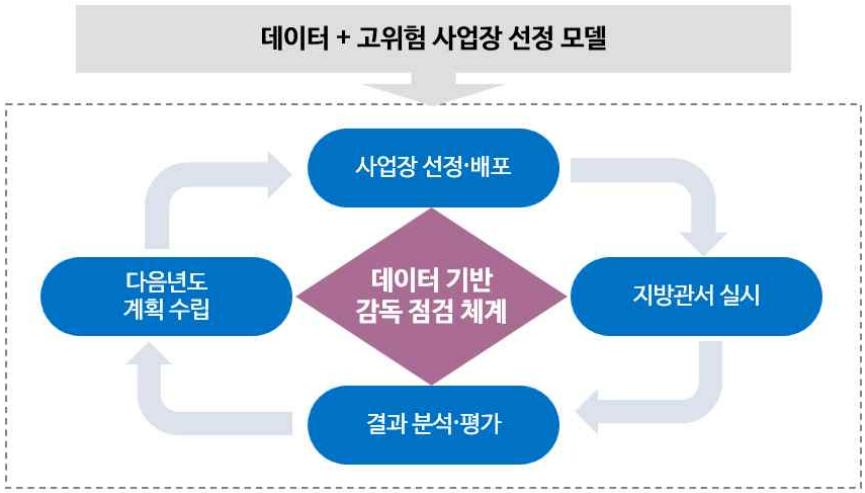
1) 연구목적

○ 현재까지 사업장을 감독·점검할 때 감독관(또는 공단 직원)은 복잡한 산업재해 관련 법령과 지침 등을 기준으로 일일이 감독하고 관리하고 산업재해가 발생했을 때, 그에 대한 내용을 조사하고 판단하여 보고를 작성하는 방식으로 관리되고 있으며, 재해를 분류하고, 고위험 사업장을 판단하기 위해 가장 중요한 자료를 생산하고 있음.

○ 이러한 데이터를 활용하여 데이터 기반 고위험 사업장 선정 모델을 구축하고 감독·점검을 위한 기본 프레임워크로 사용하고자 함.

○ 데이터기반의 고위험 사업장 선정을 위한 핵심은 데이터와 인공지능 모델을 통해 사업장의 상태를 “저위험, 중위험, 고위험, 초고위험”으로 정확히 분류하는 기술을 개발하는 것이나, 실제 산업재해가 발생할 가능성이 얼마나 되는지, 발생할 수 있는지, 또는 사업장이 어느 정도의 산재위험에 노출되었는지 판단하는 것은 매우 복잡한 과제임.

○ 그러므로 본 연구에서는 [그림 1-8]같이 데이터 기반의 감독·점검을 위한 기본 프레임워크로, 데이터와 고위험사업장 선정 모델을 이용한 “데이터기반 감독.점검 체계”를 바탕으로 [1]**사업장 선정·배포**, [2]**지방관서 지시.감독 실시**, [3]**결과 분석·평가**, [4]**다음연도 계획 수립**하는 **데이터 기반 감독·점검체계를 구축**하는 것이 본 연구를 통해 달성하고자 하는 핵심 목적이라 할 수 있음.



[그림 1-8] 데이터 기반의 고위험 사업장 선정 프레임워크

2) 연구 목표

○ 데이터 기반 감독·점검체계 구축이 본 연구의 목적으로 이를 달성하기 위하여 고위험사업장의 선정 방식을 개선하고, 선제적인 재해발생 사업장 예측을 지원하며, 다양한 인공지능 모델 개발 및 활용을 위해 다음과 같은 연구목표를 설정함.

- 데이터 기반 고위험 사업장 선정
- 데이터 기반 감독·점검 체계 완성
- 사업장 자율점검 체계 지원
- 선제적 사업장 위험 분류

○ 연구목표 달성을 위해 필요한 고위험 사업장 선정 방식의 한계 극복, 선제적 재해발생 사업장 예측 부재와 고위험 사업장 선정 인공지능 모델 부재로 인한 산업재해 관리·감독의 어려움 해소를 위해 다음과 같이 연구과제를 수행.

- 산재데이터의 포괄적 탐색 및 구조화(데이터 수집 및 전처리)
- 기존 고위험사업장 선정 모델 및 데이터 분석
- 고위험 사업장 선별 모델 설계 및 개발
- 산업안전 도메인 맞춤형 언어모델 개선 및 모델 성능 비교 분석

○ [그림 1-9]은 본 연구과제의 필요성에 따라 연구목표를 도식화한 것임.

[그림 1-9] 연구목표

3. 선행 연구 및 관련 기술

1) 선행 연구 내용 분석

(1) 인공지능 알고리즘을 활용한 재해개요 분류모델 시범 개발

○ 산업재해의 감축의 중요성을 인식하고, 기존의 감독 및 점검 프로세스의 한계를 극복하기 위해 인공지능과 자연어처리 기술을 활용한 재해개요 자동 분류 모델을 개발하는 것을 목적으로 함.

○ Word2Vec, BERT, GPT 등 인공지능 모델을 사용하여 비구조화된 산업재해 보고서와 같은 복잡한 텍스트 데이터를 처리하고 분석하며, 이를 통해 재해개요 데이터의 품질진단을 실시하고 품질을 개선하여 데이터 가치를 극대화함.

○ 연구를 통해 개발된 재해개요 분류 모델은 기존의 수동적이고 시간이 많이 소요되는 부분을 자동화되고 효율적인 방법을 제시함.

○ 인공지능 기술, 특히 자연어 처리와 머신러닝 기법을 활용하여 산업안전과 관련한 데이터 분석의 자동화 및 효율성 증진을 목적으로 함.

○ 비정형 데이터(재해 보고서, 사업장 작업 환경 설명 등) 및 구조화된 데이터의 처리와 분석을 통해 산업재해 예방 및 고위험 사업장 선정의 정확성을 높이는데 중점으로 함.

○ 본 연구는 사업장의 위험도 평가에 중점을 두는 반면 재해개요 분류모델 시범 개발은 산업재해 데이터의 자동 분류를 중점으로 함.

(2) 사업장 정량정보를 활용한 산재고위험사업장 선별 효과성 평가 및 개선방안

○ 고위험사업장을 효과적으로 선별하기 위해 빅데이터와 기계학습 기술의

적용 가능성을 탐색하고, 실제 현장에서 시범적으로 적용하여 감독 효율성과 정확성 향상시키는 것을 목적으로 함.

○ 제조업 중심 다양한 산업분야에서 고위험 사업장 데이터를 수집하여 빅데이터와 기계학습 알고리즘을 활용한 고위험사업장 선별모델 개발

○ 빅데이터와 기계학습 기술을 활용한 고위험사업장 선별의 초기 단계 연구로서, 기존 연구와 달리 실제 현장에 모델을 적용하고 가능성과 효과를 확인함.

○ 고위험사업장을 효과적으로 선별하기 위한 모델 개발에 중점을 두며, 이를 위해 빅데이터와 기계학습 기술을 활용함.

○ 본 연구는 모델의 구체적인 개선 방안에 중점을 두고, 다양한 데이터 소스와 모델링 기법의 탐색 및 적용에 초점을 맞추며 고위험사업장 선정 과정의 전반적인 최적화에 더 광범위한 접근 시도.

○ 사업장 정량정보를 활용한 산재 고위험사업장 선별 효과성 평가 및 개선 방안 연구는 고위험사업장 선별 모델의 초기 단계 연구로서, 제조업 중심으로 실제 현장에서의 모델 적용과 효과 검증에 집중함.

(3) 한국의 산업별 산업재해 발생 추이와 경기적 영향 요인 연구

○ 경기적 요인이 산업재해 발생에 미치는 영향을 실증적으로 분석하여, 산업재해 예방을 위한 맞춤형 전략 수립에 기여를 목적으로 함.

○ 경기적 요인과 산업재해 발생률 간의 상관관계를 분석하여, 경기 상황에 따른 산업재해 경향을 파악하고 분석 결과를 바탕으로 산업재해 예방 전략과 정책 설계.

○ 경기적 요인을 산업재해 예방 정책에 통합하는 새로운 접근 방식을 제시하며, 기존 산업재해 연구들이 주로 작업 환경이나 작업자의 행동과 같은 내부 요인에 집중된 반면, 경제환경이라는 외부 요인의 영향을 분석

함으로써 산업재해 예방에 대한 보다 넓은 관점 제공.

○ 산업재해의 예방 및 관리 개선을 목적으로 하며, 산업재해와 관련된 데이터 분석과 모델링에 중점을 두어 발생의 패턴을 식별하고 예방 전략 개발 접근 방식 공유.

○ 본 연구는 고위험사업장 식별 모델의 개선에 초점을 맞추며 산업별 산업재해 발생 추이와 경기적 영향요인 연구는 경제환경이라는 외부 요인을 중점으로 산재 발생에 미치는 영향을 분석함.

2) 관련 기술 조사

(1) Support Vector Machine (SVM)

○ SVM은 분류(Classification)와 회귀(Regression) 문제에 널리 사용되는 머신러닝 모델이며, 특히 이진 분류 문제에서 높은 성능을 보이고 선형 또는 비선형 결정 경계를 찾는데 사용됨.

○ SVM의 기본원리, 커널종류, 장·단점 등 주요 특징은 〈표 1-1〉와 같음.

〈표 1-1〉 SVM의 주요 특징

구분	내용
기본 원리	• **최대 마진 분류기**: 데이터 포인트들을 분류하는 결정 경계(Decision Boundary)를 찾되, 가장 가까운 트레이닝 데이터 포인트로부터 최대한 떨어지게 함. 이로써 일반화 오류가 적은 모델이 생성됨 • **커널 트릭**: 선형으로 분리가 불가능한 데이터에 대해서, 커널 함수를 사용하여 고차원 공간으로 매핑한 뒤, 고차원에서 선형 분리를 시도함

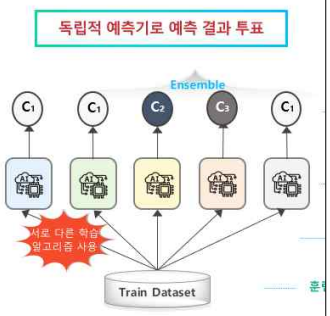
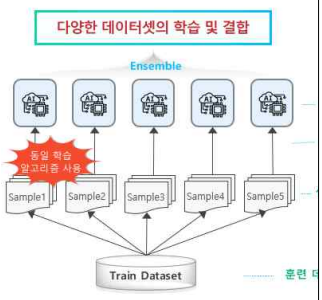
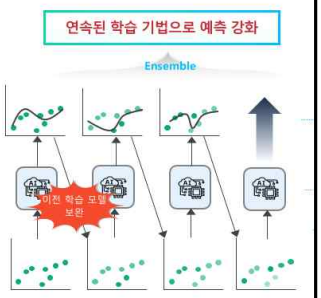
구분	내용
커널 종류	• **선형 커널(Linear Kernel)**: 선형 결정 경계를 필요로 하는 상황에서 사용 • **다항식 커널(Polynomial Kernel)**: 데이터를 다차원 공간으로 매핑하여 비선형 결정 경계를 찾는데 사용 • **RBF 커널(Radial Basis Function Kernel):** 각 데이터 포인트를 중심으로 하는 고차원 공간으로 매핑하여, 더욱 복잡한 비선형 결정 경계 찾을 수 있음
장점 및 단점	• **장점** - 고차원 데이터에 대해서도 뛰어난 성능을 보여주며, 특히 차원의 저주에 강함 - 다양한 커널 함수를 사용하여 다양한 형태 데이터에 대응할 수 있음 - 마진을 최대화하는 원리 덕분에, 과적합을 방지하며 좋은 일반화 성능을 보임 • **단점** - 훈련 과정에서 모든 데이터 포인트 간의 거리를 계산하여야 하기 때문에, 데이터셋 크기가 매우 클 경우, 훈련 시간이 오래 걸릴 수 있음 - 결정 과정이 블랙박스와 같아 해석이 어려움 - 커널 종류와 커널 매개변수, 정규화 매개변수 등 매개변수 설정에 성능이 크게 의존하기 때문에, 추가 작업이 필요함

(2) 앙상블 학습

○ 앙상블은 여러 머신러닝 모델을 묶어 더 강력한 모델을 만드는 기법을 총칭하며, 다양한 분류 및 회귀 문제에서 효과적임을 입증하였으며 연구 및 프로젝트 시, 만들어진 여러 괜찮은 성능의 모델들을 활용하여 더 강력한 모델을 구성할 수 있음.

○ 〈표 1-2〉은 다양한 모델들을 결합하여 더 강력한 모델 생성하는 앙상블 학습 방법인 보팅(Voting), 배깅(Bagging)과 페이스팅(Pasting), 부스팅(Boosting)을 비교하여 놓았음.

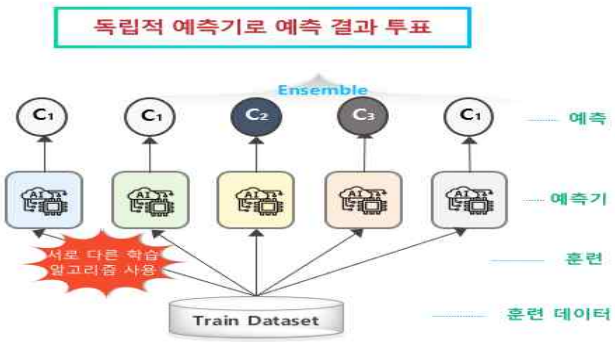
〈표 1-2〉 주요 앙상블 방법 비교

다양한 모델들을 결합하여 더 강력한 모델 생성		
보팅(Voting)	배깅(Bagging)과 페이스팅(Pasting)	부스팅(Boosting)
		
• 여러 개의 분류기의 예측 결과로 투표를 진행 후 최종 예측 결과를 결정하는 방법 • 다양한 학습 알고리즘 적용 및 개별 모델 보완으로 예측 결과 지속 보완 가능	• 데이터 샘플링으로 데이터를 분리 후, 동일 학습 알고리즘으로 모델을 학습시키고 결과를 집계하는 방법 • 병렬 학습으로 높은 확장성 제공	• 이전 학습 모델을 보완하며 강한 예측기를 만드는 방법 • 잘못 분류된 훈련 샘플의 가중치를 상대적으로 높이고 재학습으로 예측기 강화

가) 보팅(Voting)

○ 앙상블 방법 중 하나인 보팅은 각 분류기의 예측을 모아서 가장 많이 선택된 클래스를 예측하는 비교적 간단한 방법이며, 개별 분류기 중 가장 뛰어난 것보다 정확도가 높음.

○ [그림 1-10]은 보팅(Voting) 예측 방법을 도식화한 것이며, 보팅(Voting)의 주요 특징은 〈표 1-3〉와 같음.



[그림 1-10] 보팅(Voting) 예측 방법

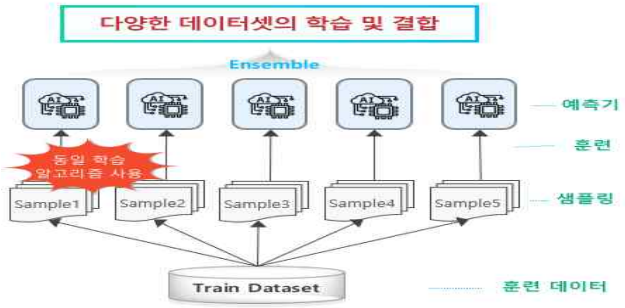
〈표 1-3〉 보팅의 주요 특징

구분	내용
보팅 종류	• Hard Voting:각 모델의 예측값(분류 클래스) 중 가장 많이 선택된 클래스를 최종 예측값으로 결정하는 방식이며, 주로 분류 문제에 주로 사용됨 • Soft Voting: 각 모델의 예측 확률을 평균내어, 가장 높은 평균 확률을 가진 클래스를 최종 예측값으로 결정. 각 모델의 예측에 대한 확률을 고려하므로 Hard Voting보다 더 정확한 예측이 가능할 수 있음
장점 및 단점	• **장점** - 서로 다른 알고리즘을 결합함으로써, 각 모델의 강점을 활용하고 약점을 보완할 수 있음 - 단일 모델보다 일반적으로 더 나은 예측 성능 달성할 수 있음 - 여러 모델의 예측을 결합함으로써 과적합 위험 줄일 수 있음 • **단점** - 여러 개의 모델을 동시에 훈련하고 관리해야 하므로, 학습과 예측 과정이 더 복잡하고 시간이 더 오래 소요될 수 있음 - 개별 모델의 예측 결과를 조합하기 때문에, 최종 예측이 어떻게 결정되었는지 어려울 수 있음

나) 배깅(Bagging, Bootstrap Aggregating)과 페이스팅(Pasting)

○ 배깅과 페이스팅은 여러개의 학습 알고리즘을 동일하게 적용하여 개별적으로 학습시킨 후, 그 결과를 집계하는 방법임.

○ [그림 1-11]는 배깅과 페이스팅의 예측 방법을 도식화한 것이며, 배깅과 페이스팅의의 주요 특징은 〈표 1-4〉와 같음.



[그림 1-11] 배깅(Bagging)과 페이스팅(Pasting) 예측 방법

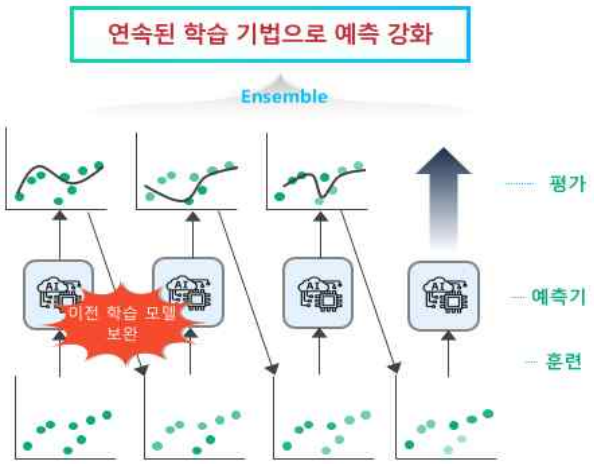
〈표 1-4〉 배깅 및 페이스팅의 주요 특징

구분	내용
배깅과 페이스팅 차이점	• **배깅(Bagging)**: 부트스트랩 샘플링(Bootstrap sampling)을 사용하여 원본 학습 데이터셋에서 여러 개의 서브셋을 무작위 생성. 하나의 샘플이 여러개 선택 될 수 있으며 각 서브셋으로 독립적인 모델 학습시킨 후 모든 모델의 예측을 집계하여 최종 예측 결정 • **페이스팅(Pasting)**: 배깅과 유사하지만, 부트스트랩 샘플링 대신 비복원 추출 방식을 사용하여 서브셋을 생성함. 중복 없이 무작위로 선택된 샘플들로 구성되며 이를 개별 모델로 독립적으로 학습시키고, 배깅과 같은 방법으로 모든 모델의 예측을 집계하여 최종 예측 결정
배깅과 페이스팅 대표 알고리즘 종류	• **랜덤 포레스트(Random Forest)**: 결정트리(Decision Tree)를 기반으로 하며 각 트리는 데이터셋의 랜덤한 샘플로 사용해 독립적으로 학습함. 분류와 회귀 모두 사용 가능하며 특성 중요도 평가 시 유용함 • **엑스트라 트리(Extra Tree)**: 랜덤 포레스트와 달리 최적의 분할을 찾는 대신 완전히 무작위로 분할하며, 이로 인해 훈련 속도가 빨라질 수 있음
장점 및 단점	• **장점** – 여러개의 모델을 조합함으로써 예측의 분산을 줄이고, 일반화 성능을 향상시킴 – 개별 모델이 훈련 데이터의 일부만 사용하므로 과적합 위험이 줄어듬 – 각 모델은 독립적으로 학습되므로 병렬처리가 가능하여 학습속도를 향상시킬 수 있음 • **단점** – 여러개의 모델을 학습시키므로 단일 모델보다 더 많은 계산량이 필요함 – 개별 모델 예측을 집계하여 최종 예측을 내므로, 단일 모델보다 해석이 복잡해질 수 있음

다) **부스팅(Boosting)**

○ 부스팅 방법은 약한 학습기 여러개를 순차적으로 학습시켜 각 학습기 예측을 결합하여 최종적으로 더 강력한 모델을 만드는 기법이며, 연속된 학습과정에서 이전 학습기의 오류를 줄여 나가는 방식으로 성능을 향상시킴.

○ [그림 1-12]는 부스팅의 예측 방법을 도식화한 것이며, 부스팅의 주요 특징은 〈표 1-5〉와 같음.



[그림 1-12] 부스팅(Boosting) 예측 방법

○ 부스팅의 대표 알고리즘 중 에이다부스트 (AdaBoost)는 이전 학습기가 잘못 분류한 샘플에 더 큰 가중치를 부여함으로써, 새로운 학습기가 해당 샘플을 올바르게 분류할 수 있도록 하며 과정을 반복하며 성능을 향상시킴.

○ 그레디언트 부스팅(Gradient Boosting)은 손실 함수의 기울기를 이용하여 약한 학습기를 학습시키는 방법이며, 각 단계에서 손실 함수를 최소화하는 방향으로 모델을 업데이트하며 점진적으로 성능을 개선함.

○ XGBoost, LightGBM은 그레디언트 부스팅을 기반으로 한 고성능 알고

리즘이며, 대규모 데이터셋과 복잡한 데이터 구조에서도 뛰어난 성능을 보이며, 과적합 방지, 범주형 변수 자동 처리 등 개선된 기능 제공함.

○ 이들 부스팅 방법은 복잡한 분류 및 회귀에서도 강력한 성능을 제공하고 중요도가 낮은 특성을 자동으로 걸러내며 규제를 사용함으로써 과적합을 방지하여 높은 예측 정확도를 제공하는 장점이 있음.

○ 반면, 여러 모델을 순차적으로 학습하므로 계산 비용이 높으며, 매개변수의 수와 데이터 크기의 수가 클 경우 학습 시간이 오래 걸리고, 하이퍼파라미터 조정이 성능에 큰 영향을 미칠 수 있으며, 최적의 매개변수를 찾기 위한 노력이 필요하다는 단점이 있음.

〈표 1-5〉 부스팅의 주요 특징

구분	내용
부스팅 대표 알고리즘 종류	• **에이다부스트 (AdaBoost)**: 이전 학습기가 잘못 분류한 샘플에 더 큰 가중치를 부여함으로써, 새로운 학습기가 해당 샘플을 올바르게 분류할 수 있도록 함. 과정을 반복하며 성능을 향상시킴 • **그레디언트 부스팅(Gradient Boosting)**: 손실 함수의 기울기를 이용하여 약한 학습기를 학습시키는 방법이며, 각 단계에서 손실 함수를 최소화하는 방향으로 모델을 업데이트하며 점진적으로 성능을 개선함 • **XGBoost, LightGBM** : 그레디언트 부스팅을 기반으로 한 고성능 알고리즘이며, 대규모 데이터셋과 복잡한 데이터 구조에서도 뛰어난 성능을 보이며, 과적합 방지, 범주형 변수 자동 처리 등 개선된 기능 제공함
장점 및 단점	• **장점** - 높은 예측 정확도를 제공하며, 복잡한 분류 및 회귀에서도 강력한 성능을 제공함 - 중요도가 낮은 특성을 자동으로 걸러내며 규제를 사용함으로써 과적합을 방지함 • **단점** - 여러 모델을 순차적으로 학습하므로 계산 비용이 높으며, 매개변수의 수와 데이터 크기의 수가 클 경우 학습 시간이 오래걸림 - 하이퍼파라미터 조정이 성능에 큰 영향을 미칠 수 있으며, 최적의 매개변수를 찾기 위한 노력이 필요함

(3) 딥러닝 모델

○ 딥러닝은 기계학습의 한 분야로, 다층으로 이루어진 신경망을 통해 복잡한 패턴을 학습하는 알고리즘이며, 데이터로부터 특징을 자동 추출하고 이를 기반으로 분류, 회귀, 패턴 인식 등 다양한 문제를 해결할 수 있음.

〈표 1-6〉 딥러닝 모델의 주요 특징

구분	내용
딥러닝 모델 기본구조	• **인공 신경망(Artificial Neural Networks, ANN)**: 딥러닝 기본 모델로, 입력층과 하나 이상의 은닉층, 출력층으로 구성되며 각 층은 노드들로 구성되고, 노드들은 가중치와 활성화 함수를 통해 연결됨 • **활성화 함수(Activation Function)**: 신경망의 비선형성을 도입하여 복잡한 패턴을 학습할 수 있게 해주는 함수이며 대표적으로 ReLU(Rectified Linear Unit), 시그모이드(Sigmoid), 하이퍼볼릭 탄젠트 등이 사용됨
주요 딥러닝 모델	• **합성곱 신경망(Convolutional Neural Networks, CNN)**: 이미지 인식, 분류 등에 우수한 성능을 보이며, 이미지의 공간 정보를 유지하며 특징을 추출함 • **순환 신경망(Recurrent Neural Networks, RNN)**: 시계열 데이터, 자연어처리에 적합한 모델로, 순서가 있는 데이터의 정보를 메모리에 저장하는 방식으로 작동함 • **트랜스포머(Transformer)**: 주로 자연어처리에 사용되며, 어텐션(Attention) 메커니즘을 사용하여 입력 시퀀스의 모든 요소 간의 관계를 한번에 계산함으로써 뛰어난 병렬처리 능력과 효율성을 보임
장점 및 단점	• **장점** - 데이터로부터 자동으로 특징을 추출하며, 이 과정을 통해 복잡한 데이터의 내재된 패턴을 스스로 학습 - 이미지, 오디오, 텍스트 등 고차원 데이터를 효과적으로 처리하며 다양한 분야에서 성능이 우수함 - 다양한 구조와 알고리즘으로 조합하여 광범위한 문제에 적용할 수 있음 • **단점** - 대규모 데이터와 복잡한 모델 구조로 인해 한번 학습 시 오랜 시간과 자원 필요함 - 매우 큰 모델은 과적합 위험이 있으며, 적절한 규제 및 학습 전략 필요함 - 모델의 결정 과정이 복잡하여 해석이 어려움

4. 연구내용 및 방법

1) 연구내용

(1) 산업안전 데이터 수집 및 전처리

가) 산업안전 데이터 조사 및 수집

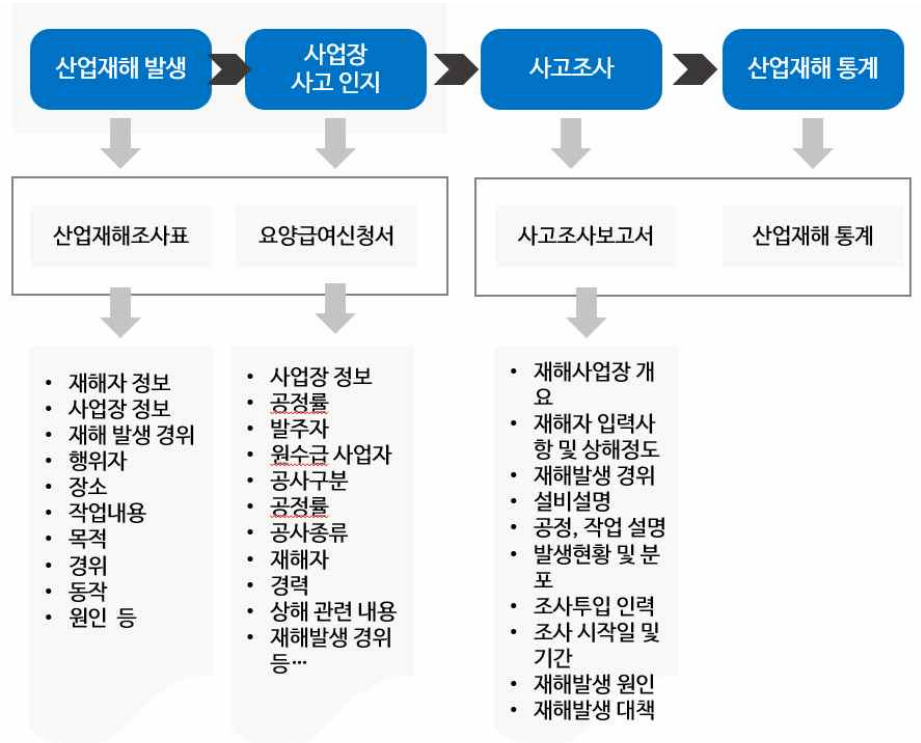
○ 산업안전 관련 데이터(위험성 평가, 산재승인 통계, 중대재해 조사통계 등)와 사고 발생 상황, 원인, 대책 등의 사례 정보 등을 포괄적으로 조사 및 수집함.

○ 산업재해를 일으키는 상황은 매우 다양하고 복잡하며, 여러 상황이 서로 연계되어 산업재해가 발생될 수 있어, 산업재해가 발생된 후 그 당시의 상황을 정확하게 기록한 산업안전 관련 데이터의 수집이 무엇보다 중요함.

○ 재해의 원인과 결과를 분석하기 위하여 수집한 산업안전 관련 데이터를 아래와 같이 산업재해 분류와 관련된 주요 메타 정보를 추출함.

산업재해 분류와 관련된 주요 메타 정보
❶ 재해 개요 : 사업주 또는 업무 담당자가 기록하는 발생 된 재해에 대한 설명
❷ 고위험 작업/상황 : 약 4,000여 건의 위험한 작업 명과 작업별 상세 작업들(공종, 작업명, 단위작업으로 구성)
❸ 기인물 : 산업재해가 발생된 장소/시설/장비 등
❹ 재해유발 요인 : 재해 발생을 야기한 요인들 (날씨, 부주의, 작업 미숙 등)
❺ 위험성 감소 대책 : 고위험 작업/상황과 재해유발 요인 등을 전문가가 분석하여 제공하는 경보
❻ 재해 발생 형태 : 재해나 사고가 어떤 형태로 발생 되었는지를 기록한 정보
❼ 작업지역 공정 : 작업을 수행하는 과정이나 공정에 대한 기록
❽ 작업내용 : 산재가 발생할 때 작업자가 처리하였던 작업내용
❾ 기타 : 작업방식, 장비의 상태정보(재질, 무게 방법 등),

○ 산업재해 관련 보고 단계별 수집할 수 있는 메타데이터는 [그림 1-13]과 같으며, 산업재해 발생 시 산업재해조사표, 사업장사고인지 시 요양급여신청서, 사고조사 시 사고조사보고서에서 그리고 산업재해통계에서 수집할 수 있음. [그림 1-14]는 요양급여신청서와 산업재해조사표 양식임.

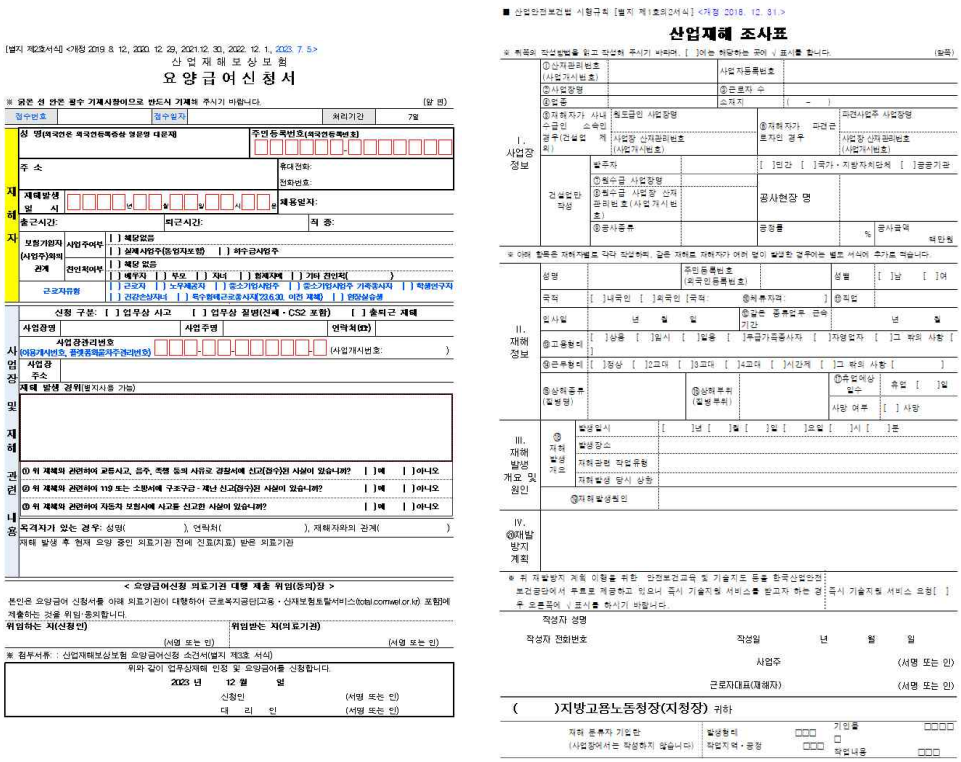


[그림 1-13] 산업재해 관련 보고 단계별 메타데이터

○ 산업재해조사표는 재해자 정보, 사업장 정보, 재해 발생 경위, 행위자, 사고 발생 장소, 작업내용, 작업목적, 작업동작, 원인 등의 메타데이터가 있음.

○ 요양급여신청서는 사업장 정보, 공정률, 발주자, 원수급 사업자, 공사구분, 공정률, 공사종류, 재해자, 경력, 상해 관련 내용, 재해발생 경위 등이 있음.

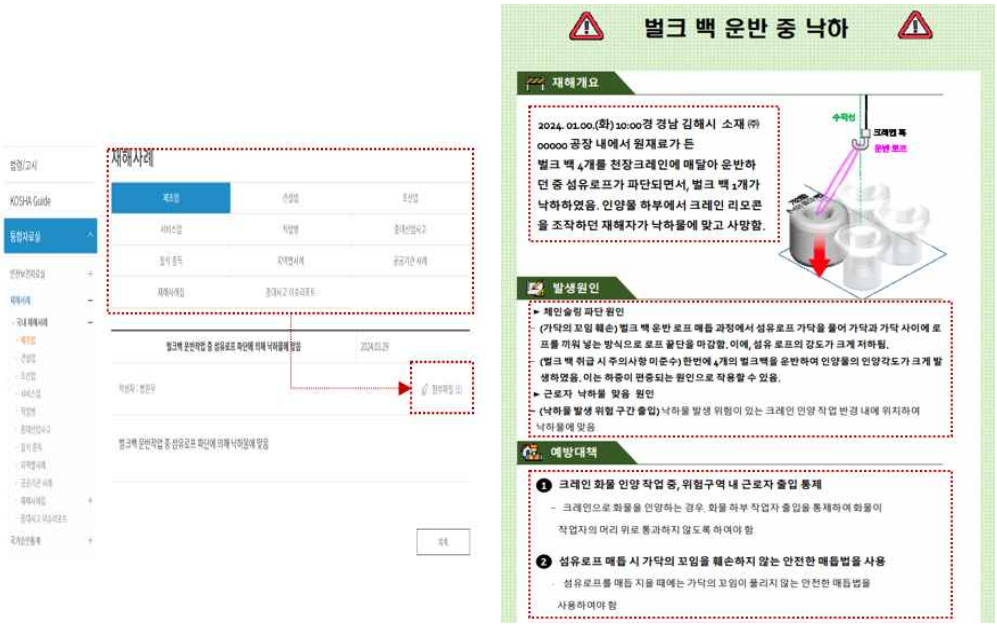
○ 사고조사보고서의 메타데이터는 재해사업장 개요, 재해자 입력사항, 재해발생 경위, 설비설명, 공정·작업 설명, 발생현황 및 분포, 조사투입 인력, 조사 시작일 및 기간, 재해발생 원인, 재해발생 대책 등이 있음.



[그림 1-14] 산업재해 관련 보고서 양식

○ 요양급여신청서, 산업재해조사표와 사고조사보고서에 작성되는 메타데이터 들의 주요 내용은 데이터 처리를 위한 기초 정보임. 이들 문서는 사업장에서 작성되는 로우데이터로서 작성자가 어떻게 작성했느냐에 따라 데이터 처리나 알고리즘의 수준이 달라질 수 있음.

○ [그림 1-15]는 “벌크 백 운반 중 낙하”로 인한 재해사례 기록으로 재해 개요, 발생원인, 예방대책 등 산업안전 정보를 수집하는 예임.



[그림 1-15] 재해사례 예시

○ 안전보건공단은 '산업재해기록·분류에 관한 지침'을 제공하고 있는데, 이 지침에서는 산업재해 특성 분석항목을 규정하여 어떤 항목들이 어떻게 기록되어야 하는지를 명시하고 있으며, 해당 내용에 따라 산업재해 관련 보고서 작성 및 기준 적용과 산업재해 관리에 참조할 수 있도록 규정하고 있어 사업장 위험요인 분석 및 산업안전 패턴식별 등에 활용함.

○ 아래는 '산업재해기록·분류에 관한 지침'에서 규정하고 있는 "산업재해 특성 분석 항목"임.

사업장 특성 분석 항목	재해자 분석 항목	재해발생 특성 분석 항목
(1) 사업자등록번호 (2) 산재관리번호 (3) 사업장명 (4) 소재지 (5) 산업(업종)	(1) 국적 (2) 성별 (3) 연령 (4) 직업 (5) 고용형태	(1) 재해발생일시 (2) 재해발생시점 (3) 재해종류 (4) 피해현황(인적, 물적, 조업정지)

사업장 특성 분석 항목	재해자 분석 항목	재해발생 특성 분석 항목
(6) 규모(근로자수) (7) 행정구역 (8) 사업장형태 (9) 공사종류 (10) 공사금액 (11) 공사기간 (12) 공정율	(6) 근무형태 (7) 동종업무 근속기간(입사근속기간)	(5) 안전방호조치 (6) 안전방호조치여부 (7) 개인보호조치 (8) 개인보호조치여부 (9) 작업형태 (10) 발생형태 (11) 기인물(가해물) (12) 작업지역·공정(평소수행, 재해당시 수행, 재해유발) (13) 작업내용(평소수행, 재해당시수행, 재해유발) (14) 불안전한 상태 (15) 불안전한 행동 (16) 추락장소 (17) 추락높이 (18) 감전전압 (19) 점화원 (20) 상병종류 (21) 상병부위 (22) 근로손실일

나) 산업안전 데이터 정제 및 전처리

○ 데이터 정제와 전처리 과정은 모델 성능을 최적화하는 데 필수적이며, 이상치 처리, 데이터 불균형 처리, 코드값 변환, 결측치 처리 등의 과정은 데이터의 품질을 높이고, 모델이 정확한 예측을 할 수 있도록 함.

○ 연구에서 적용한 전처리 목표는 다음과 같음.

1. 데이터 품질 향상: 전처리 기법들은 데이터의 결함을 보완하고 품질을 향상시켜 모델이 유의미한 패턴을 학습할 수 있도록 도움
2. 모델의 민감도 확인: 다양한 전처리 방법을 적용함으로써 모델이 데이

터 변환에 어떻게 반응하는지 확인

3. 전처리 방법 선택: 여러 전처리 기법을 비교함으로써 모델 성능에 가장 긍정적인 영향을 미치는 방법 선택

○ 〈표 1-7〉는 본 연구에서 사용된 데이터 전처리 방법을 정리한 표로, 각 방법의 특징과 장단점들을 비교하여 나타냄.

〈표 1-7〉 데이터 전처리 방법

구분	방법	설명
이상치처리	Z-Score	데이터가 평균으로부터 얼마나 떨어져 있는지를 표준편차 단위로 계산하여 이상치를 식별하는 방법
	IQR	데이터의 중앙값을 기준으로 1사분위수(Q1)와 3사분위수(Q3)의 차이를 이용하여 이상치를 식별하는 방법
	Isolation Forest	랜덤으로 선택한 피처와 분할 값을 기반으로 트리를 생성하고, 이상치가 다른 데이터보다 일찍 분리되는 원리를 이용함
	DBSCAN	밀도 기반 군집화 알고리즘으로, 밀도가 낮은 데이터 포인트를 이상치로 간주
	LOF	각 데이터 포인트의 밀도를 인근 포인트의 밀도와 비교하여 이상치를 판별하는 방법
데이터 불균형처리	SMOTE	소수 클래스의 데이터를 랜덤하게 선택하고, 해당 데이터의 K-최근접 이웃을 기반으로 새로운 샘플을 합성하는 방법
	ADASYN	SMOTE와 유사하지만, 소수 클래스 중 어려운 예제에 더 많은 가중치를 부여하여 데이터를 생성하는 방법
	Random Undersampling	다수 클래스의 데이터를 랜덤하게 제거하여 클래스 간의 균형을 맞추는 방법
	Gaussian Noise	소수 클래스의 데이터에 Gaussian Noise를 추가하여 데이터의 변형을 유도하는 방법
	Bootstraping	원본 데이터에서 중복을 허용하여 랜덤 샘플을 추출하는 방법
코드값 변환(인코딩)	Label Encoding	각 범주형 변수를 정수 값으로 변환하는 방법, 카테고리마다 고유한 숫자를 할당
	One-Hot Encoding	각 범주형 변수를 이진 벡터로 변환하는 방법, 각 범주는 독립된 열로 변환되며, 해당하는 열에만 1, 나머지는 0으로 표시됨
	Target Encoding	각 카테고리 값을 해당 범주가 가진 타겟 변수의

구분	방법	설명
		평균값이나 가중 평균으로 변환하는 방법
	Mean Encoding	Target Encoding의 한 형태로, 범주형 변수의 각 값에 대해 타겟 변수의 평균값을 사용하여 변환하는 방법
	Binary Encoding	각 카테고리 값을 정수로 변환한 후, 그 정수를 이진수로 변환하여 각 자릿수를 새로운 열로 분리하여 인코딩하는 방법
결측치처리	평균값 대치	결측치를 해당 열의 평균값으로 대체하는 방법
	중앙값 대치	결측치를 해당 열의 중앙값으로 대체하는 방법
	최빈값 대치	결측치를 해당 열에서 가장 빈번하게 발생하는 값으로 대체하는 방법
	선형회귀/로지스틱 회귀	회귀 분석을 사용하여 결측치를 예측하고 대체하는 방법
	KNN Imputer	결측치가 있는 샘플을 K-최근접 이웃(KNN)알고리즘을 사용하여 주변의 값으로 대체하는 방법
	MICE	여러 회귀 모델을 사용하여 결측치를 여러번 예측하고 평균을 구하여 대체하는 방법
특성 분포 불균형처리	로그 변환	값의 로그를 취해 데이터의 분포를 압축하는 방법으로, 주로 데이터가 양의 값이고 비대칭적인 분포일 때 사용하는 방법
	제곱근 변환	데이터의 제곱근을 취해 데이터의 스케일을 줄이는 변환 방법
	역수 변환	데이터의 역수를 취하는 변환. 데이터의 큰 값을 작게 만들고 작은 값을 상대적으로 더 크게 만드는 방법
	Box-Cox 변환	값이 양수인 경우에만 적용이 가능한 변환 함수로, 특정 파라미터(람다)를 사용하여 데이터를 정규 분포에 가깝게 변환
	Yeo-Johnson 변환	Box-Cox 변환의 확장으로, 음수나 0값을 포함한 데이터에도 적용 가능한 방법. 데이터 분포를 정규 분포에 가깝게 변환
데이터 범주화처리	Equal-Width Binning	데이터 값을 일정한 간격(폭)으로 나누어 구간을 생성하고, 해당 구간에 속하는 데이터들을 같은 범주로 분류하는 방법
	Equal-Frequency Binning	데이터를 각 구간에 동일한 수의 데이터를 배분하여 범주로 나누는 방법
	K-Means Binning	K-Means 알고리즘을 사용하여 데이터를 유사한 값끼리 클러스터로 묶은 뒤, 각 클러스터를 하나의 범주로 나누는 방법

구분	방법	설명
특성 스케일링	Standard Scaler	특성의 평균을 0, 표준편차를 1루 맞추어 데이터의 분포를 표준 정규 분포로 변환하는 스케일링 방법
	Min-Max Scaler	데이터를 최솟값과 최댓값 사이의 범위로(일반적으로 0과 1)로 변환하는 스케일링 방법
	Robust Scaler	중앙값과 IQR(사분위 범위)를 사용하여 데이터 스케일을 조정하는 방법
	Power Transformer	Box-Cox 또는 Yeo-Johnson 변환을 사용하여 정규 분포에 가까운 형태로 변환
데이터 분할 처리	Train-Test Split	데이터를 훈련 세트와 테스트 세트로 단순하게 비율로 나누는 방법. 일반적으로 70~80%는 훈련, 20~30%는 테스트 데이터로 사용
	K-Fold Cross Validation	데이터를 K개의 폴드(fold)로 나누고, K번의 훈련 및 테스트를 진행. 각 번마다 다른 폴드를 테스트 세트로 사용
	Stratified K-Fold	K-Fold와 동일하나, 각 폴드에서 클래스 비율을 유지하도록 데이터를 나눔. 주로 분류 문제에서 사용

□ 이상치 처리(Outlier Handling)

○ 이상치는 모델 성능에 부정적인 영향을 미치는 극단적인 값으로, IQR(Interquartile Range) 방법이나 Z-Score와 같은 기법을 통해 처리함.

○ XGBoost는 트리 기반 알고리즘으로 이상치에 상대적으로 강하며, 이상치가 있어도 트리가 분할을 통해 적절히 처리하므로, 직접 이상치를 제거하지 않아도 되는 경우가 있음.

〈표 1-8〉 이상치 처리 방법 비교 분석

구분	특징	장점	단점
Z-Score	• 평균과 표준편차에 의존 • Z-Score 값이 임계값을 넘어가면 이상치로 간주됨(일반적으로 ±3 사용)	• 계산이 간단하고 해석이 쉬움 • 데이터의 정규성 가정하에 효과적	• 정규 분포를 따르지 않는 데이터에서 부정확 • 극단값이 민감
IQR	• 데이터의 사분위수를	• 정규분포 가정을 하지	• 데이터 왜곡 시

구분	특징	장점	단점
	사용 • Q1 − 1.5IQR보다 작거나 Q3 + 1.5IQR보다 크면 이상치로 간주	않음 • 극단값에 덜 민감	비효율적 • 데이터 분포에 따라 임계값이 달라질 수 있음
Isolation Forest	• 이상치는 다른 데이터보다 더 쉽게 격리될 것이라는 아이디어에 기반함 • 분포에 관계없이 적용 가능	• 비정규분포 데이터에 효과적 • 다양한 차원에서 작동 가능 • 학습 기반 알고리즘으로 확장성 있음	• 설정해야 할 파라미터가 많음(트리 개수 등) • 대규모 데이터셋의 경우 속도 저하
DBSCAN	• 데이터 포인트 간의 밀도를 기준으로 군집을 형성하고, 밀도가 낮은 포인트를 이상치로 처리	• 비선형적인 분포나 복잡한 데이터에 효과적 • 군집의 크기와 모양에 관계없이 적용 가능	• 파라미터 조정 필요 • 밀집도가 균일하지 않은 경우 성능 저하
LOF	• 이웃 데이터와의 밀도를 비교하여 이상치 여부를 판단 • 국부적인 밀도 차이 이용	• 비정규 분포에 적합 • 지역적 이상치 탐지 가능	• 이웃의 수(k) 설정에 민감 • 고차원 데이터셋의 경우 성능 저하

□ 데이터 불균형 처리(Imbalanced Data Handling)

○ 데이터 불균형은 클래스 간 표본 크기가 차이가 있을 때 발생하는 문제로, 이를 해결하기 위해 대표적으로 오버샘플링과 언더샘플링, 이들을 결합한 SMOTE-Tomek 기법을 주로 사용함. 이는 소수 클래스의 데이터를 보완하는 동시에 다수 클래스의 과잉 데이터를 일부 제거하는 방법임.

〈표 1-9〉 데이터 불균형 처리 방법 비교 분석

구분	특징	장점	단점
SMOTE	• 소수 클래스의 데이터 간에 새로운 데이터를 생성하여 불균형 문제를 해결 • 이웃 간의 차이를 바탕으로 새로운	• 데이터의 다양성을 유지하면서 균형을 맞춤 • 과적합 문제를 줄이는 효과 있음	• 소수 클래스 내 데이터가 복잡한 경우 생성된 데이터가 품질 낮을 수 있음 • 경계 영역에서 과적합 발생 가능성 있음

구분	특징	장점	단점
	데이터 포인트 생성		
ADASYN	• 소수 클래스 중 어려운 예제에 더 집중하여 새로운 샘플을 생성 • 데이터 분포의 복잡성 반영	• SMOTE보다 복잡한 데이터에 더 적합 • 중요한 데이터 포인트를 우선시하여 데이터 생성	• 과적합 발생 가능성 • 지나치게 복잡한 데이터 생성할 수 있음
Random Under sampling	• 다수 클래스의 데이터를 무작위로 제거하여 클래스 비율 조정	• 간단하고 빠르게 적용 가능 • 불필요한 데이터를 줄여서 학습 속도 향상	• 중요한 데이터 손실 가능성 존재 • 데이터의 정보가 줄어들어 성능 저하
Gaussian Noise	• 소수 클래스의 데이터에 약간의 변형을 주어 새로운 샘플 생성 • 노이즈 추가를 통한 데이터 다양성 증가	• 간단한 방법으로 데이터 다양성 증대 가능 • 비선형적 분포에도 적용 가능	• 지나친 노이즈 추가 시 데이터 품질 저하 가능 • 반대로, 노이즈가 너무 적으면 효과가 미비할 수 있음
SMOTE-Tomek	• SMOTE의 오버샘플링과 Tomek Link의 경계결 데이터 제거를 결합한 방법	• SMOTE 단독 사용시 발생할 수 있는 과적합 문제를 해결할 수 있음 • 단순한 오버샘플리옵다 더 정교한 방식으로 소수 클래스 생성	• Tomek Link 제거 과정에서 다수 클래스와 소수 클래스 사이에 중요한 경계데이터가 제거될 수있음

□ 코드값 변환(Categorical Encoding)

○ 범주형 데이터는 One-Hot Encoding과 Target Encoding 등을 통해 수치형 데이터로 변환함.

○ CatBoost, LGBM 등 일부 모델에서는 별도 처리 없이 자동으로 처리할 수 있음.

〈표 1-10〉 코드값 변환 처리 비교 분석

구분	특징	장점	단점
Label Encoding	• 각 카테고리 값에 숫자 부여 • 범주의 순서가 없는 경우에도 순서가 있는	• 구현이 간단하고 빠름 • 메모리 사용량 적음	• 값에 순서가 생겨 모델이 잘못된 가정으로 학습할 수 있음 • 고유 값이 많으면

구분	특징	장점	단점
	것처럼 처리됨		숫자가 커져 모델 성능 저하 가능
One-Hot Encoding	• 카테고리의 개수만큼 새로운 열이 생성됨 • 값들 간의 순서나 관계를 나타내지 않음	• 범주 간의 순서나 관계가 없는 경우 적합 • 대부분의 모델에서 일반적으로 잘 작동함	• 카테고리 수가 많을 경우 차원이 매우 커짐(차원의 저주 문제) • 메모리 및 계산 자원 소모 큼
Target Encoding	• 범주형 데이터를 타겟값과의 관계로 인코딩 • 회귀나 분류 문제에 따라 타겟 변수에 맞춘 인코딩	• 카테고리와 타겟값과의 관계를 반영하여 성능 향상 가능 • 범주형 변수의 수가 많은 경우에도 효과적	• 교차 검증을 사용하거나 노이즈를 추가해 과적합 방지 필요
Mean Encoding	• 타겟 변수의 평균을 사용하여 범주형 변수 인코딩 • 타겟과의 관계를 반영하는 인코딩	• 타겟 값과 강하게 연관된 범주형 변수를 효과적으로 처리 가능 • 모델 성능 개선 가능	• 교차 검증을 사용하거나 노이즈를 추가해 과적합 방지 필요
Binary Encoding	• 카테고리를 숫자로 변환 후 이진수로 변환 • 차원 수를 줄이면서도 정보 손실 방지함	• One-Hot Encoding보다 효율적 • 범주형 변수의 차원을 줄이면서도 정보 손실 줄임	• 이진 변환을 사용하므로 모델에 따라 적합하지 않을 수 있음 • 데이터에 따라 복잡해질 수 있음

□ 결측치 처리(Missing Data Handling)

○ 결측치는 데이터에서 누락된 값을 의미하며, 이를 처리하지 않으면 모델의 성능에 부정적인 영향을 미칠 수 있음. 결측치 처리 방법으로는 대체 평균 또는 K-Nearest Neighbors(KNN) 대체와 같은 기법을 사용함.

○ XGBoost에서는 트리를 생성할 때, 각 노드에서 결측치가 있는 데이터를 어떻게 처리할지 스스로 결정함. 결측치가 있는 샘플을 두 갈래로 나뉘는 대신 최적의 방향으로 보내지며, 이 방향은 XGBoost가 학습과정에서 최적화하여 결측치를 가진 데이터가 어느 쪽 분기에서 더 좋은 성능을 보이는지를 기반으로 결정함.

○ 결측치를 별도 처리 없어도 스스로 모델이 처리하지만, 너무 많은 경우 모델의 성능에 부정적 영향을 미치므로 직접 처리함.

〈표 1-11〉 결측치 처리 방법 비교 분석

구분	특징	장점	단점
평균값 대치	• 연속형 변수에 적합 • 각 열의 평균값을 사용하여 결측치를 대체	• 계산이 매우 간단하고 빠름 • 극단값이 없는 경우 데이터의 변동성을 적절히 유지함	• 극단값(Outlier)에 민감하여 평균이 왜곡될 수 있음 • 데이터의 분포와 변동성이 왜곡될 수 있음
중앙값 대치	• 연속형 변수에 적합 • 극단값의 영향을 받지 않는 중앙값을 사용하여 결측치를 대체	• 극단값의 영향을 받지 않으므로 데이터의 분포가 왜곡되지 않음 • 간단하고 빠르게 적용 가능	• 평균값에 비해 정확성이 떨어질 수 있음 • 분포가 왜곡될 가능성 있음
최빈값 대치	• 범주형 변수에 적합 • 범주형 데이터에서 가장 자주나타나는 값을 결측치에 적용	• 범주형 변수의 결측치 처리에 효과적 • 계산이 간단하고 빠름	• 데이터에 왜곡을 초래할 수 있음 • 실제 데이터의 분포를 왜곡하여 편향된 결과를 초래할 수 있음
선형회귀/ 로지스틱 회귀	• 연속형(선형회귀) 또는 범주형(로지스틱 회귀) 데이터에 적용 가능 • 변수 간의 관계를 기반으로 결측치를 대체	• 변수 간 상관관계를 활용해 더 정확한 대체 가능 • 데이터의 구조와 특성을 반영	• 복잡한 모델을 필요로 하며 과적합 가능성 있음 • 회귀 모델이 잘못되면 부정확한 대체치가 생성될 수 있음
KNN Imputer	• 각 샘플의 결측치를 주변 데이터의 값에 기반해 대체 • 결측치를 다양한 변수들과의 관계를 고려하여 대체 가능	• 변수 간 상관관계를 반영하여 결측치 처리 가능 • 복잡한 데이터에서도 유효	• 계산 비용이 크고 대규모 데이터에서는 성능 저하 가능 • 최적의 K값 설정이 어려움
MICE	• 연속형 변수와 범주형 변수 모두에 적용 가능 • 다중 대체법으로 결측치 대체	• 변수 간의 관계를 고려하여 결측치 대체 가능 • 데이터의 분포와 통계적 특성을 더 잘 유지할 수 있음	• 계산 복잡도가 높고 시간이 오래 걸림 • 잘못된 모델링으로 인한 오류 가능성 존재

□ 특성 분포 불균형 처리(Feature Distribution Balancing)

○ 로그 변환(Log Transformation), Box-Cox 변환 등의 기법을 적용하여, 비대칭적인 분포를 정규분포에 가깝게 변환하였음. 이를 통해 편향된 특성은 교정하고, 모델이 더 일관성 있게 학습할 수 있도록 함.

〈표 1-12〉 특성분포 불균형 처리 방법 비교 분석

구분	특징	장점	단점
로그 변환	• 값의 크기를 줄여 비대칭 분포를 정규 분포에 가깝게 만듦 • 양수 값만 사용 가능	• 분포의 왜곡을 완화하고 데이터 범위를 줄이는 데 효과적 • 큰 값을 축소시켜 이상치 영향 감소	• 음수나 0값을 처리할 수 없음 • 분포가 정규에 가깝지 않을 경우 효과가 크지 않을 수 있음
제곱근 변환	• 양수 데이터에 적용 가능 • 이상치의 영향을 줄이며 데이터 분포 완화	• 로그 변환과 비슷한 효과로 비대칭 분포 완화 • 이상치 영향을 줄임	• 음수나 0을 처리할 수 없음 • 변환 후에도 분포가 정규 분포에 맞지 않을 가능성 존재
역수 변환	• 매우 큰 값들을 작은값으로 변환 • 데이터의 극단값을 제어하는 데 유효	• 극단값의 영향을 줄이는 데 효과적 • 특정 비대칭 분포에서 효과적	• 0을 처리할 수 없고, 0에 가까운 값이 극단값으로 변환될 수 있음 • 데이터가 정규 분포로 변환되지 않을 수 있음
Box-Cox 변환	• 양수 값에만 적용 가능 • 람다 값을 조정하여 다양한 분포에 적용 가능 • 정규성 가정이 중요한 분석에 사용	• 다양한 람다 값을 통해 최적의 분포로 변환 가능 • 데이터의 정규성을 더 정확하게 맞출 수 있음	• 양수 데이터에만 사용 가능 • 최적의 람다값을 찾아야 함 • 음수나 0값이 있으면 사용 불가
Yeo-Johnson 변환	• 음수와 0값을 포함한 데이터에서도 변환 가능 • Box-Cox와 달리 양수/음수 모두에 적용 가능	• 양수와 음수 데이터 모두에서 사용 가능 • 분포 왜곡이 심한 데이터에 효과적 • 다양한 데이터에 유연하게 적용 가능	• 최적의 변환을 위해 람다 값을 찾아야 함 • 작은 데이터셋에서는 성능이 떨어질 수 있음 • 계산 비용 큼

□ 데이터 범주화 처리(Data Binning)

○ 데이터 범주화는 연속형 데이터를 구간별로 나누어 범주형 데이터로 변환하는 과정이며, 이를 통해 데이터 간 차이를 보다 명확하게 구분함.

〈표 1-13〉 데이터 범주화 처리 비교 분석

구분	특징	장점	단점
Equal-Width Binning	• 각 구간의 너비가 동일 • 일정 범위 간격으로 구간을 나눔	• 구현이 간단하고 직관적 • 각 구간의 폭이 일정하여 해석이 쉬움	• 데이터 분포를 반영하지 않음 • 빈도가 낮은 구간이 생길 수 있음 • 이상치가 있을 경우 구간이 왜곡될 수 있음
Equal-Frequency Binning	• 각 구간에 들어가는 데이터의 개수가 동일 • 데이터 분포에 따라 구간 크기가 달라질 수 있음	• 데이터 분포를 반영하여 빈도수 차이를 줄임 • 각 구간이 비어있는 경우가 없어 구간 해석 용이	• 구간의 폭이 일정하지 않음 • 다른 구간의 값들이 지나치게 가까울 경우, 해석이 어려울 수 있음
K-Means Binning	• K-Means 알고리즘을 기반으로 함 • 군집 내 데이터들이 서로 유사하고, 다른 군집과는 다름	• 데이터 내 유사한 패턴을 반영하여 자연스러운 범주화 가능 • 각 클러스터가 데이터의 특성을 반영함	• K 값을 설정해야 함 • 계산 복잡도가 높아 대규모 데이터에서 성능 저하 가능 • 이상치가 있을 경우 결과 왜곡 가능

□ 특성 스케일링(Feature Scaling)

○ 데이터 스케일링을 통해 특성 값의 범위를 조정함. Min-Max 스케일링과 표준화(Standardization)을 적용하여, 모델이 각 특성을 공정하게 학습하도록 함.

○ XGBoost는 트리 기반 모델이기 때문에 특성 간의 스케일에 민감하지 않음. 선형 모델이나 SVM과 달리, 특성 스케일링을 하지 않아도 성능에 큰 영향을 미치지 않음.

〈표 1-14〉 특성 스케일링 처리 비교 분석

구분	특징	장점	단점
Standard Scaler	• 각 변수의 평균을 0, 표준편차를 1로 변환 • 정규 분포를 따르는 데이터에 적합	• 정규분포 가정 시, 성능 최적화 가능 • 분포가 유사한 변수 간의 영향 균등화	• 정규 분포를 따르지 않는 데이터에선 성능 저하 • 이상치에 민감하여 이상치가 있는 경우 스케일링이 왜곡될 수 있음
Min-Max Scaler	• 각 변수의 최소값을 0, 최대값을 1로 맞추어 값의 범위를 조정	• 범위가 명확하게 정해진 경우에 적합 • 이상치 없는 데이터를 효과적으로 스케일링	• 이상치에 민감하여 이상치가 있는 경우 스케일링이 왜곡될 수 있음 • 데이터의 분포 정보가 왜곡될 가능성 있음
Robust Scaler	• 중앙값을 기준으로 하고, 사분위수로 분포를 조정하여 이상치에 덜 민감	• 이상치가 많은 경우에도 안정적인 스케일링 가능 • 분포의 중앙에 더 집중하여 극단값의 영향을 줄임	• 분포 자체를 표준화하지 않으므로 정규 분포를 가정하는 모델에서 비효율적일 수 있음 • 비정규 분포에서는 최적의 결과를 보장하지 않음
Power Transformer	• 데이터의 분포를 정규분포에 가깝게 변환 • Box-Cox 및 Yeo-Johnson 처리 방식 사용	• 데이터 분포를 정규화하여 모델의 성능 개선 • 정규 분포를 가정하는 모델에 유용	• 변환 과정이 복잡하고 시간이 오래 걸림

□ 데이터 분할 처리(Data Splitting)

○ 데이터 분할 처리는 데이터를 학습용(Train), 검증용(Validation), 테스트용(Test)으로 나누어 사용함.

〈표 1-15〉 데이터 분할 처리 비교 분석

구분	특징	장점	단점
Train-Test Split	• 가장 기본적인 데이터 분할 방법 • 한 번의 분할로 훈련 및 테스트 진행 • 무작위로 나눔	• 간단하고 직관적임 • 계산 비용이 적고 빠름 • 모델의 성능을 빠르게 확인 가능	• 데이터 편향이 발생할 수 있음 • 데이터가 충분하지 않을 경우
K-Fold Cross Validation	• 데이터를 K개의 폴드로 나눠서 K번 반복 • 모델이 모든 데이터를 한 번씩 테스트 데이터로 사용 가능	• 모든 데이터를 한번씩 테스트하여 모델 성능 평가의 안정성 증가 • 데이터 손실없이 전반적인 성능 평가 가능, 편향 줄임	• K번 학습하므로 계산 비용이 증가 • 데이터셋이 클 경우 시간이 오래 걸릴 수 있음 • 데이터가 많지 않은 경우에도 계산 복잡도가 커질 수 있음
Stratified K-Fold	• 클래스 비율을 동일하게 유지하여 데이터를 나눔 • 분류 문제에서 클래스 불균형 문제 완화	• 불균형한 데이터셋에서 각 클래스의 비율을 유지하여 편향된 학습을 방지 • 소수 클래스가 충분히 반영되도록 처리	• 계산 비용 증가 • 각 폴드의 비율을 맞추기 위해 추가 계산 필요 • 비율을 맞추기 위해 데이터 개수가 양이 적은 클래스에 맞춰질 수 있음

2) 기존 고위험사업장 선정 모델 및 데이터 분석

(1) 기존 연구 및 모델 분석

○ 기존 모델의 데이터 사용 방식, 선정 기준, 성능 지표 외에도 모델의 예측 오류 유형, 성공 사례, 실패 사례 분석을 통해 모델의 장단점을 더욱 구체적으로 파악.

가) 고위험사업장 선정 모델 개요

○ 기존의 고위험 사업장 선정 모델은 상대적으로 위험 사업장을 선별하여 이를 지도 및 점검하는 것을 목표로 하며, 사업장의 안전보건 수준을 전반적으로 향상시키데 중점을 두고 있음. 이를 위해 인공지능(AI) 모델인 XGBoost(eXtreme Gradient Boosting) 알고리즘을 활용하여, 사업장의 위험도를 0~1 사이의 수치로 예측하고 있음.

- XGBoost는 Gradient Boosting을 개선한 트리 앙상블 모델로, 대규모 데이터 처리에서 효율적이며, 과적합 방지 및 병렬 처리가 가능하다는 장점을 지니고 있음.
- 모델 학습 시 손실함수의 기울기(Gradient)를 따라 최적화가 이루어지며, 각 트리가 이전 트리의 잔차를 보완하는 방식으로 성능을 점진적으로 향상시킴.
- 분류 문제에서는 주로 로그 손실(Log Loss) 함수를 사용하며, 모델이 예측한 값과 실제 값 간의 차이를 최소화하는 방향으로 학습이 진행됨. 또한, 정규화 항(Regularization Term)이 추가되어 과적합을 방지함.

○ 위험 사업장 데이터는 불균형한 특성을 가지며, 안전 사업장에 비해 위험 사업장의 비율이 낮은 편임. 이러한 데이터 불균형 문제는 클래스 가중치(Class Weight)조정이나 오버샘플링(Oversampling) 기법을 통해 처리할 수 있음. 또한, 사업장 규모, 사고 발생 이력, 근로 환경 등 다차원적인 특성을 포함하는 대규모 데이터셋을 처리할 수 있으며, XGBoost는 이를 기반으로 높은 예측 성능을 발휘함.

○ 〈표 1-16〉은 XGBoost 알고리즘으로 고위험 사업장 선정 모델 학습 시 하이퍼파라미터 최적화 과정에서 역할과 적용 범위임.

〈표 1-16〉 XGBoost 주요 하이퍼파라미터

구분	역할	적용 범위
부스팅 라운드 수 (n_estimators)	모델이 사용할 결정 트리의 수를 설정함. 많은 트리를 사용할수록 모델이 복잡해지고 과적합 위험이 커짐	100 ~ 1000
트리 최대 깊이 (max_depth)	각 결정 트리의 최대 깊이를 설정함. 깊이가 깊을수록 트리가 복잡해지고, 세부적인 패턴을 학습할 수 있지만,과적합 위험이 커짐	3~10
학습률 (learning_rate)	각 부스팅 라운드에서 가중치 업데이트의 크기를 결정하며, 학습률이 낮으면 모델이 더 천천히 학습하며, 높은 학습률은 더 빠르게 수렴하지만 불안정할 수 있음	0.01~0.1
샘플 비율 (subsample)	각 트리를 학습할 때 사용할 훈련 데이터의 비율을 설정함. 1.0이면 전체 데이터를 사용하고, 그보다 작으면 데이터를 무작위로 샘플링하여 사용함	0.5~1.0
트리별 특성 샘플 비율 (colsample_bytree)	각 트리를 학습할 때 사용할 특성의 비율을 설정함	0.3~1.0
레벨별 특성 샘플 비율 (colsample_bylevel)	트리의 각 레벨에서 사용할 특성 비율을 설정함	0.3~1.0
분할 기준 최소 손실 감소값 (gamma)	트리의 리프 노드가 분할되기 위해 필요한 최소 손실 감소값이며, 값이 클수록 덜 복잡한 트리가 생성됨	0~5
리프 노드의 최소 가중치 합 (min_child_weight)	각 리프 노드의 최소 가중치 합을 설정하며, 값이 클수록 모델이 더 보수적으로 분할함	1~10
L2 정규화 항 (lambda)	가중치에 대한 L2 정규화를 적용하여 모델 복잡도를 제어함. 큰 값일수록 가중치가 작아지며, 모델의 과적합을 방지함	0~10
L1 정규화 항 (alpha)	가중치에 대한 L1 정규화를 적용하여 가중치가 0이 되는 특성을 강제로 선택하게 만듦.	0~10
불균형 클래스에 대한 가중치 (scale_pos_weight)	클래스 불균형이 있는 경우, 양성 클래스에 대한 가중치를 설정함	1~5

나) 모델 성능 평가

○ 모델의 성능을 평가하기 위해 Confusion Matrix를 사용하였으며, 모델의 예측 정확도 뿐 아니라 정밀도, 재현율 등의 성능지표를 활용함.

1. Confusion Matrix 구조

- Confusion Matrix는 모델이 예측한 값과 실제 값 간의 관계를 시각적으로 표현하며, 정확도, 정밀도, 재현율 등의 성능 지표를 도출하는 도구이며, 이를 통해 모델의 성능을 구체적으로 평가할 수 있음.

〈표 1-17〉 Confusion Matrix 구조

	실제 Positive	실제 Negative
예측 Positive	True Positive(TP)	False Positive(FP)
예측 Negative	False Negative(FN)	True Negative(TN)

- **True Positive(TP):** 모델이 긍정 클래스(Positive, 고위험사업장)를 정확히 예측한 경우
- **False Positive(FP):** 모델이 부정 클래스(Negative, 안전사업장)를 긍정으로 잘못 예측한 경우
- **False Negative(FN):** 모델이 긍정 클래스를 부정으로 잘못 예측한 경우
- **True Negative(TN):** 모델이 부정 클래스를 정확히 예측한 경우

2. Confusion Matrix 기반 성능 지표

- **정확도(Accuracy):** 모델이 전체 데이터 중 얼마나 정확하게 예측했는지를 나타내며, 데이터가 불균형할 경우 정확도가 높더라도 잘못된 결론을 낼 수 있음

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

- **정밀도(Precision):** 모델이 고위험사업장으로 예측한 것 중 실제로 고위험사업

장 데이터의 비율을 나타냄

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **재현율(Recall):** 모델이 실제 고위험사업장 데이터 중 얼마나 많은 데이터를 올바르게 예측했는지를 나타냄

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1 Score:** 정밀도와 재현율의 조화 평균을 나타내는 지표로, 두 지표 간의 균형을 중요시할 때 사용하며, 본 연구에서 전반적인 모델의 성능을 평가할 때 활용함

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

○ 특히, 불균형 데이터의 경우에는 정확도(Accuracy)만으로는 모델 성능을 정확히 평가하기 어려우며, F1 Score는 정밀도와 재현율을 균형있게 고려하여, 모델이 고위험 사업장을 얼마나 잘 탐지하고 있는지를 종합적으로 평가할 수 있음.

(2) 개선 가능성 및 고위험사업장 선정 모델 개발 방향 설정

○ 공단 및 고용부의 고위험사업장 선정 모델 등, 관련 기관의 데이터 및 모델과의 통합 가능성 및 추후 모델 다양화 가능성 등을 고려하여 모델 고도화 방안 구성.

(3) 산업안전 데이터 기반의 고위험 사업장 분석 모델 개발

가) 고위험사업장 선정 중요 요인 및 특성 분석

○ 학습데이터의 기본적 전처리 외 키워드 분포와 전반적인 데이터의 균형

도 및 성향 분석을 수행함.

○ 특히, 서로 다른 산업분류별 위험도 분석 시 아래와 같은 방법으로 수행함.

- 범주별로 데이터를 분류하여 별도 학습
- 범주를 특성으로 포함하여 한번에 학습

○ “범주별로 데이터를 분류하여 별도 학습”시키는 방법은 각 산업 특성에 맞춘 모델을 개발할 수 있어 특화된 모델 생성에 효과적이고, 특정 산업에서 발생하는 데이터의 양이 다를 수 있어, 산업별로 분리 시 이러한 불균형 완화할 수 있으며, 산업별로 모델 분리시 각 모델의 복잡도 관리가 용이하다는 장점이 있음, 반면, 각 산업별로 별도의 모델을 개발하고 학습시켜야 하므로, 시간과 계산 자원이 더 많이 소모되고, 산업별 모델은 해당 산업에만 집중되기 때문에, 다른 산업으로 일반화가 어렵다는 단점이 있음.

○ “범주를 특성으로 포함하여 한번에 학습”시키는 방법은 하나의 모델을 사용하여 모든 산업 데이터를 학습시킬 수 있으므로, 자원 사용이 효율적이고, 모델이 산업 간 차이를 학습하고 일반화하는 능력이 향상될 수 있으며, 여러 산업의 데이터를 통합하여 분석함으로써, 특정 산업에서만 발견되지 않는 새로운 패턴이나 인사이트 발견할 수 있다는 장점이 있음. 반면, 다양한 산업의 데이터를 한꺼번에 학습시키려면 모델이 복잡해질 수 있으며, 이는 과적합 위험이 있고, 다양한 산업을 포괄하는 큰 데이터셋에는 특정 산업의 데이터가 과소 대표될 위험이 있다는 단점이 있음.

〈표 1-18〉 범주별 데이터 학습 처리 장단점

방법	장점	단점
범주별로 데이터를 분류하여 별도 학습	• **(특화된 모델 생성)** 각 산업 특성에 맞춘 모델을 개발할 수 있으며, 해당 산업에 특화된 예측 성능을 달성할 수 있음 • **(데이터 불균형 해소)** 특정 산업에서 발생하는 데이터의 양이 다를 수 있으며, 산업별로 분리 시 이러한 불균형 완화할 수 있음 • **(복잡도 관리)** 산업별로 모델 분리시 각 모델의 복잡도를 관리하는 데 더 용이할 수 있음	• **(자원 소모)** 각 산업별로 별도의 모델을 개발하고 학습시켜야 하므로, 시간과 계산 자원이 더 많이 소모됨 • **(일반화의 제한)** 산업별 모델은 해당 산업에만 집중되기 때문에, 다른 산업으로 일반화가 어려움
범주를 특성으로 포함하여 한번에 학습	• **(효율성)** 하나의 모델을 사용하여 모든 산업 데이터를 학습시킬 수 있으므로, 자원 사용 효율적 • **(일반화)** 모델이 산업 간 차이를 학습하고 일반화하는 능력이 향상될 수 있음 • **(인사이트 발견)** 여러 산업의 데이터를 통합하여 분석함으로써, 특정 산업에서만 발견되지 않는 새로운 패턴이나 인사이트 발견할 수 있음	• **(모델 복잡도)** 다양한 산업의 데이터를 한꺼번에 학습시키려면 모델이 복잡해질 수 있으며, 이는 과적합 위험이 있음 • **(데이터 불균형)** 다양한 산업을 포괄하는 큰 데이터셋에는 특정 산업의 데이터가 과소 대표될 위험 있음

○ 데이터의 양과 질, 목적에 따라 실험을 통해 두 접근법의 성능을 비교 분석하는 것은 유의미한 접근 방법이 될 수 있으며, 모델 학습에 사용된 특성의 중요도를 분석하여, 고위험사업장 선정에 핵심적인 요인 식별, 이는 모델의 해석 가능성과 정책 결정 과정에 기여할 수 있음.

나) 고위험사업장 선정 모델 설계 및 실험

○ 단일 모델뿐만 아니라 “머신러닝”, “딥러닝 모델”, “앙상블 모델” 등 다양한 아키텍처를 실험하여 고위험사업장 선정에 최적화된 모델을 선정하여 활용함.

○ “머신러닝”은 이해하기 쉬우며, 결과를 해석하는데 용이하고, 구조화된 데이터에 대해 효과적이며, 특성 중요도를 제공하여 어떤 변수가 예측에 가장 큰 영향을 미치는지 파악할 수 있다는 장점이 있어, 구조화된 데이터에 대한 고위험 사업장 선정 기준 분석 시 사용됨.

○ “딥러닝 모델”은 비구조화된 텍스트 데이터(ex. 사고보고서, 작업 환경 설명 등)의 처리에 효과적이며, BERT나 GPT와 같은 최신 NLP 모델은 복잡한 텍스트에서 유용한 정보 추출 시 우수한 성능 보여, 복잡한 데이터 구조, 비정형 데이터에서 더 나은 예측 성능을 달성하기 위해 사용됨.

○ “앙상블 모델”은 여러 학습기의 예측을 결합하여 보다 강력한 예측 모델을 만들 수 있으며, 과적합 방지하며 안정적 성능 제공하고 있어, 복잡한 데이터 구조에서 더 나은 예측 성능을 달성하기 위해 사용됨.

〈표 1-19〉 모델별 특징 및 적용 예

모델	특징	적용 예
머신러닝 (ex. 결정 트리, 랜덤포레스트 등)	• 이해하기 쉬우며, 결과를 해석하는 데 용이함. 구조화된 데이터에 대해 효과적이며, 특성 중요도를 제공하여 어떤 변수가 예측에 가장 큰 영향을 미치는지 파악할 수 있음	• 구조화된 데이터에 대한 고위험 사업장 선정 기준 분석 시 사용
딥러닝((NLP 모델 포함)	• 비구조화된 텍스트 데이터(ex. 사고보고서, 작업 환경 설명 등)의 처리에 효과적이며, BERT나 GPT와 같은 최신 NLP 모델은 복잡한 텍스트에서 유용한 정보 추출 시 우수한 성능 보임	• 복잡한 데이터 구조, 비정형 데이터에서 더 나은 예측 성능을 달성하기 위해 사용 될 수 있음
앙상블 모델과 부스팅	• 여러 학습기의 예측을 결합하여 보다 강력한 예측 모델을 만들 수 있으며, 과적합 방지하며 안정적 성능 제공	• 복잡한 데이터 구조에서 더 나은 예측 성능을 달성하기 위해 사용할 수 있음

○ 본 연구에서 사용될 데이터는 구조화된 형식(예: 사업장 크기, 근로자 수 등)과 비구조화된 형식(예: 사고보고서, 사업장 작업 환경 설명 등) 모두 포함될 수 있으며, 이를 고려할 때 연구 목적에 적합한 모델을 선택하려면 다음 기준들을 고려해야 함.

- 다양한 데이터 유형 처리 능력
- 고차원 데이터에 대한 처리 능력
- 해석 가능성

〈표 1-20〉 모델선정 기준

① 다양한 데이터 유형 처리 능력
- 구조화된 데이터와 비구조화된 데이터 모두를 처리할 수 있는 유연성 필요
- 예를 들어 ,텍스트 데이터는 자연어 처리(NLP) 기술을 사용하여 분석할 수 있어야 하며, 구조화된 데이터는 전통적인 머신러닝 기법으로도 효과적으로 처리할 수 있어야 함

② 고차원 데이터에 대한 처리 능력
- 산업재해 데이터는 다양한 소스에서 온 복잡하고 고차원 정보를 포함할 수 있음
- 따라서 모델은 고차원 데이터에서 중요한 정보를 추출하고 이를 기반으로 예측을 수행할 수 있는 능력이 중요함

③ 해석 가능성
- 고위험 사업장 선정 모델의 결정에 대해 설명할 수 있는 능력은 정책 입안자와 실무자에게 중요한 정보를 제공함
- 모델이 어떤 기준을 기반으로 사업장을 고위험으로 분류하였는지 이해하는 것이 중요

○ 〈표 1-20〉은 구조화된 데이터 모델과 비구조화된 데이터를 하이브리드 형으로 함께 결합하여 모델을 구성하고 각 모델의 예측에 대한 신뢰도를 반영하여 보다 정밀한 사업장 위험도 평가를 가능하게 하는 앙상블 모델를 적용한 방법 예임.

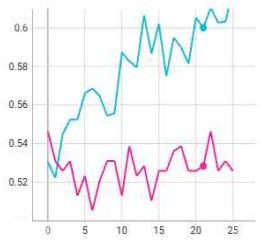
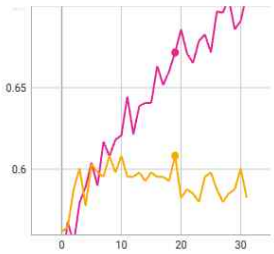
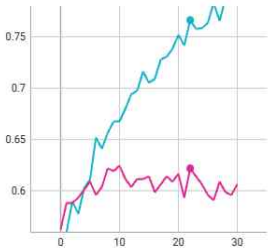
〈표 1-21〉 앙상블 모델 적용 예

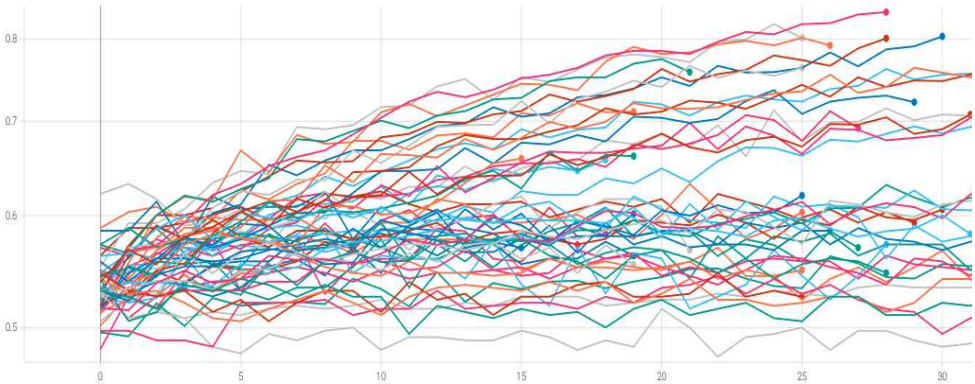
단계	내용
1. 모델 구성	• (구조화된 데이터 모델) 결정 트리, 랜덤 포레스트, 그레디언트 부스팅 등을 포함하는 전통적인 머신러닝 모델로 구조화된 데이터(근로자 수, 이전 사고 기록, 사업장 규모 등) 분석 • (비구조화된 데이터 모델) BERT, GPT 등의 NLP 모델을 사용하여 사업장의 작업 환경 설명, 사고보고서와 같은 텍스트 데이터에서 유용한 정보 추출
2. 개별 모델의 예측	• 각 모델은 독립적으로 사업장의 위험도에 대한 예측을 수행하며, 이때의 출력은 0에서 1사이의 확률 값으로, 사업장이 고위험에 속할 확률 나타냄
3. 예측 결합	• 개별 모델의 예측을 결합하는 방법으로는 간단한 평균, 가중 평균, 스태킹(다른 모델을 사용하여 개별 모델의 출력을 결합) 등이 있음 • 간단하게는 각 모델의 예측에 가중치를 부여하여 최종 점수를 계산하는 가중 평균 방식을 사용할 수 있음
4. 최종 점수 계산 예시	• 사업장 A에 대한 개별 모델 예측과 가중치가 다음과 같다고 가정 - 구조화된 데이터 모델 예측: 0.7 (가중치: 0.3) - 비구조화된 데이터 모델 예측: 0.8 (가중치: 0.7) - 최종 점수 = (0.7 x 0.3) + (0.8 x 0.7) = 0.21 0.56 = 0.77

○ 학습 과정에서 다양한 데이터의 특성을 지속적으로 추가 및 최적화하고, 모델의 하이퍼파라미터를 조정하면서 여러 모델을 지속적으로 비교 및 개선하며, 최종적으로 가장 효과적인 예측 결과를 도출하는 모델을 선정함으로써, 연구 목적에 부합하는 최적화된 해결책 제공.

○ 〈표 1-22〉는 FCN(딥러닝 모델), XGBoost(앙상블 모델), SVM(머신러닝 모델) 모델들의 학습 결과를 비교한 예시이며, [그림 1-16]은 학습 회차별 모델 Accuracy 값 변화의 예시임.

〈표 1-22〉 여러 모델 학습 결과 비교

구분	모델별 학습		
테스트 모델	FCN(딥러닝 모델)	XGBoost(앙상블 모델)	SVM(머신러닝 모델)
Epoch별 정확률			
범례 표시	train validation	train validation	train validation



[그림 1-16] 학습 회차별 모델 Accuracy 값 변화

(4) 언어모델 기반 고위험사업장 선정 모델 최적화

가) 언어모델의 산업안전 도메인 특화 개선

○ 산업안전 분야의 특수성을 고려한 맞춤형 언어모델 개발 및 적용으로, 분야 특유의 용어와 문맥을 정확히 이해하고 분석할 수 있는 모델 구성.

나) 성능 비교 분석 및 모델의 지속적 성능 개선

○ 다양한 모델들의 성능을 비교 분석하여 모델별 장단점 및 개선 가능성 제시.

2) 연구방법

○ 고위험 사업장 선정 모델을 개선하고 데이터기반 감독·점검 체계 구축을 위한 연구목적을 달성하기 위하여 [그림 1-17]와 같이 5단계 연구단계를 설정하고 각 연구단계를 달성하기 위하여 총 10개의 세부 단계을 설정하여 수행함.

[그림 1-17] 연구 흐름도

(1) 1단계: 산업안전 데이터의 탐색 및 구조화

가) 산업안전 데이터 조사 및 도메인 특성 분석

○ 산업안전 관련 데이터(위험성평가, 직업환경측정, 산재승인 통계, 중대재해 조사통계 등) 및 사고발생 상황, 원인, 대책 등 다양한 데이터를 조사·수집하고, 산업안전 분야의 도메인적 특성 분석.

나) 비정형 데이터의 정형화 및 가공

○ 언어모델에 활용할 수 있는 PDF 형식의 문서, 사고 보고서 등 비정형 데이터를 텍스트로 변환 및 데이터베이스에 저장 가능한 정형화된 형태로 변환·가공.

다) 고위험사업장 선정모델 학습데이터셋 구성

○ 학습용 데이터셋과 연구에 필요한 데이터 전반을 구성하며, 사업장의 위험도 영향 요인을 분석하고, 데이터 전처리(불용어, 결측치, 이상치, 정규화, 통합 등) 작업을 수행하여 수집된 데이터를 분석에 적합한 형태로 가공.

(2) 2단계: 기본 모델 및 데이터 분석

가) 기존 연구 및 모델 분석

○ 고위험사업장 선정에 관한 기존 연구 및 모델들을 분석하여, 성능 지표, 사용된 데이터 종류, 분석 방법론 등을 정리하고, 모델의 장단점 및 개선 가능성과 통합 가능성 등 분석.

나) 산업안전 데이터 특성 분석

○ 수집된 데이터에 대한 전처리 및 특성 분석 수행, 데이터의 균형도 등 전반적 성향을 분석하여 고위험사업장 선정에 중요한 요인 식별.

(3) 3단계: 고위험사업장 선정 모델 설계 및 개발

가) 모델 설계 및 실험

○ 다양한 머신러닝 및 딥러닝 모델을 설계하고 개발하여, 기존 모델뿐 아닌 여러 설계된 모델들을 반복 학습하여 성능 개선하고, 전통적인 머신러닝 모델부터 언어 모델까지 다양한 방법론으로 점진적 반복 학습 수행.

나) 초기 모델 비교 분석 및 평가

○ 초기 모델의 성능을 평가하여 분석 수행하고, 정확도, 정밀도, 재현율 등 다양한 평가 지표를 사용하여 각 모델의 장단점을 식별하며, 산업안전 분야에 적합한 모델 특성 파악.

(4) 4단계: 산업안전 도메인 맞춤형 언어모델 적용 및 모델 성능 비교분석

가) 산업안전 도메인 맞춤형 언어모델 적용

○ 산업안전 관련 도메인의 특수성을 고려하여 언어 모델 학습을 실시하고, 단어 수준을 넘어 문맥적 이해가 가능하도록 방안을 구성.

나) 모델 성능 비교 분석

○ 기존 모델과 개선된 모델의 성능 및 활용 방법을 비교 분석하고, 언어 모델과 결합하여 개선된 모델의 해석력을 기존 방법과 비교해 장단점 및 한계점을 도출.

(5) 5단계: 연구 결과에 대한 실용적방안 검토 및 제언

○ 사용자가 학습된 모델과 다양한 데이터 시나리오로 예측 성능 확인할 수 있는 포로토타입 구성.

○ 시스템 통합 및 활용 전략 제언.

- 연구를 통해 도출된 결과와 관찰된 패턴을 체계적으로 정리하여, 산재 발생 현황 모니터링 시스템 등에 적용 방안 등 제언

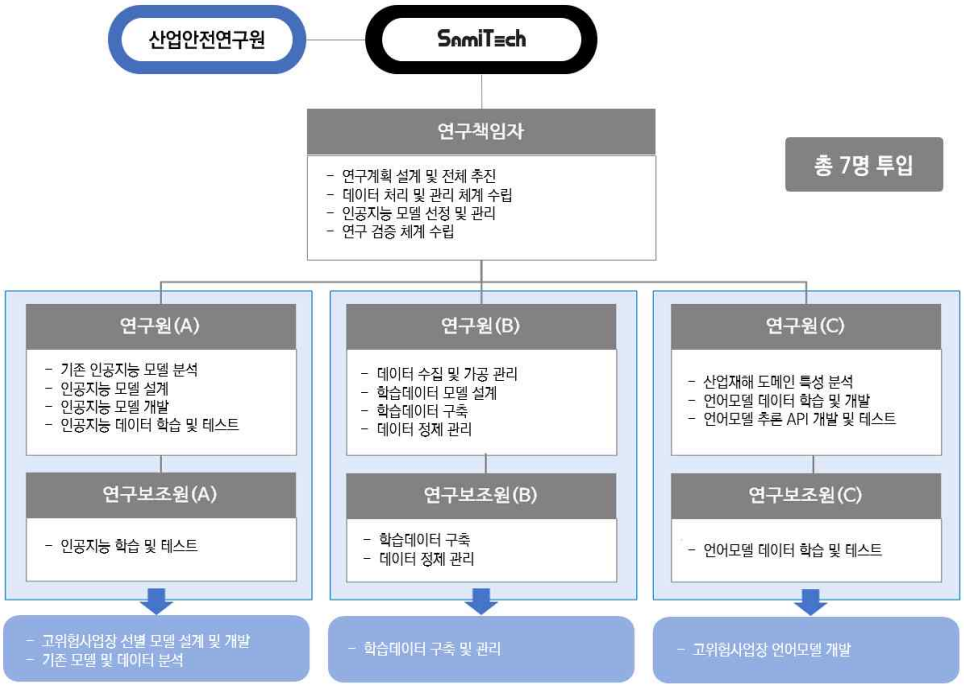
Ⅱ. 연구 수행 체계

Ⅱ. 연구 수행 체계

1. 연구 추진조직

1) 연구 추진조직

○ 고위험사업장 선정모델 개선을 통한 감독·점검 효과성 제고방안 연구는 연구책임자를 포함하여 연구원 3명, 연구보조원 3명, 총 7명이 참여하여 수행함.



[그림 2-1] 연구 추직 조직도

2) 참여 역할별 세부 수행 내용

○ 효율적인 연구수행을 위해 과업내용을 기반으로 연구책임자의 역할 외에 3개의 파트로 나누어 역할을 다음과 같이 수행함.

연구책임: 연구수행 총괄

- 연구계획 설계 및 전체 추진
- 데이터 처리 및 관리 체계 수립
- 인공지능 모델 선정 및 관리
- 연구 검증 체계 및 실증화 방안 수립

제1 파트: 고위험사업장 선별 모델 설계 및 개발, 기존 모델 및 데이터 분석

- 기존 인공지능 모델 분석
- 인공지능 모델 설계 및 개발
- 인공지능 데이터 학습 및 테스트 서비스
- 인공지능 학습 테스트

제2 파트: 학습데이터 구축 및 관리

- 데이터 수집 및 가공 관리
- 학습 데이터 모델 설계 및 구축
- 데이터 정제 관리
- 데이터 가공 및 모니터링 수행

제3 파트: 언어모델 개발 및 테스트

- 산업재해 도메인 특성 분석
- 언어모델 데이터 학습 및 개발
- 언어모델 추론 API 개발 및 테스트

2. 연구 추진 경과

1) 기존모델 및 데이터 중간 분석

○ 기존 모델은 연구개발 환경에서 재현 후, 현황 분석 및 개선점 도출.

○ 기존 모델 환경에서 해당 수준 평가(3개 문항 5점 척도)를 반영하여 효과성 확인.

○ 고위험사업장 AI 예측 모델 개발과정에서 다양한 추가 방법(데이터 가공, 학습방법 등) 적용 및 효과성 확인.

○ 데이터 편향성과 모델 복잡성의 상호작용이 학습 성능에 부정적인 영향을 미치고 있는 것으로 보이며, 모델이 특정 특성에 지나치게 집중하거나, 데이터의 특성이 고르게 반영되지 않음.

○ 모델 성능 개선을 위한 단순한 조치뿐 아니라, 데이터 품질과 특성의 편향성에 대한 분석 필요.

〈표 2-1〉 기존모델 및 데이터 분석 결과

연번	구분	내용
1	특성 중요도의 편향성	• XGBoost 모델에서 특정 특성에 과도한 중요도가 부여되어, 모델이 특정 특성에만 의존하는 예측 수행 • 다른 중요한 특성이 충분히 반영되지 않는 현상 관찰됨
2	데이터 특성의 편향성	• 일부 특성들이 다른 특성들에 비해 비정상적으로 높은 또는 낮은 분포를 보임 • 특성 편향으로 인해 모델이 학습 과정에서 과도하게 편향된 결과를 도출하게 되어, 일반화 성능을 저해하는 원인 중 하나가 될 수 있음

연번	구분	내용
3	과적합 문제	• 훈련 데이터에 지나치게 맞추어져 있어, 훈련 데이터에서는 성능이 높지만 새로운 데이터에 대해 성능이 크게 떨어지며 일반화 성능이 저하됨 • 노이즈나 특이 패턴까지 학습된 결과인지 확인 필요
4	다중공선성 및 과도한 특성 사용	• 다중공선성이 있는 특성들이 모델에 입력될 경우, 모델이 과도하게 상관된 특성에 의존할 수 있어 성능 저하 초래함 • 과도한 특성 수로 인해 모델 복잡성이 증가하고 과적합 가능성이 높아질 수 있으므로, 주요 특성을 선정해 수를 줄이는 방법 필요

○ 이러한 문제를 해결하기 위해 앞서, 다양한 모델을 적용하여 데이터 문제인지 또는 모델 문제인지 검토 수행.

○ 데이터와 모델 간의 상호작용, 모델별로 성능 차이가 발생하였는지 비교분석 수행.

2) 고위험사업장 선별 모델 설계 및 개발 점검

(1) 모델 및 데이터 전처리 영향 확인

○ 다양한 딥러닝 모델 및 머신러닝 모델을 적용한 결과, 학습이 조기에 종료되었으며 모델 성능평가 점수는 높게 나타나 보이나, 실질적인 학습 성과는 기대에 미치지 못함. 이러한 현상은 여러 모델에서 반복적으로 관찰되었으며, 다양한 데이터 전처리 기법을 적용하여도 효과는 미미함.

○ 모델의 성능 개선보다 데이터 자체의 문제 검토, 특히 학습 데이터 재설정이 필요함.

〈표 2-2〉 모델 및 데이터 전처리 영향 확인 결과

연번	구분	내용
1	과적합 및 조기 수렴	• 높은 정확도와 F1 Score는 단순히 훈련 데이터에 과적합된 결과일 가능성이 높음 • 모델이 데이터의 일반화에 실패하고, 훈련 데이터에만 지나치게 맞춰져 학습이 빠르게 수렴하는 문제 발생
2	데이터 전처리 방식의 한계	• 다양한 데이터 전처리 기법을 적용했음에도 성능 차이가 미미하며, 데이터 자체의 편향성이나 불균형 문제 등이 여전히 모델 성능이 영향을 미치고 있음을 시사함 • 전처리 기법으로 해결할 수 없는 데이터 품질 문제가 내재되어 있을 수 있음
3	학습데이터 재설정의 필요성	• 특성과 Target Label 간의 관계가 제대로 설정되지 않은 것이 문제일 수 있음 • 특정 특성들이 실제로 위험을 잘 반영하지 못하는 경우, 특성과의 관계를 재설정해야 학습 성능이 개선될 가능성이 높음
4	주요 특성 추출	• 유사한 특성들이 많을 때, 모델이 특정 특성에 과도하게 의존하게 되며, 이로 인한 특성들이 서로 비슷한 정보 제공하여 모델이 불안정해질 수 있음 • 특성 수가 많고 과적합에 영향을 미치는 문제를 확인하기 위해, 상관분석, SHAP, RFE 등의 기법을 적용해 주요 특성 추출, 특성별 모델 성능 변화 확인 필요

(2) 학습데이터 재조정 후 학습결과

○ Label 1(위험사업장)의 경우 위험성을 확실하게 확인이 가능하지만, 사고 발생 이력이 없는 사업장은 사고 발생 시, 경미한 피해가 주로 발생할지, 큰 피해가 발생할지 알 수 없으므로, 이를 Label 0(안전사업장)으로 사용 시 노이즈가 될 수 있음을 가정함.

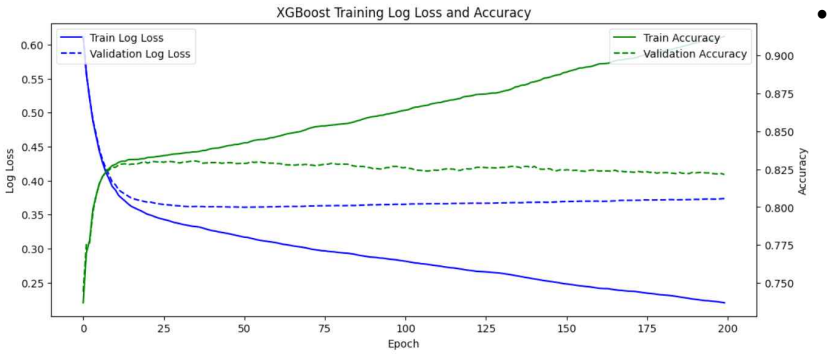
〈표 2-3〉 학습데이터 재조정 후 학습 결과

연번	구분	내용
1	노이즈 데이터 제거	• 승인통계에 기반한 경미한 사고가 발생한 사업장만을 선별하여 확실한 Label 0 데이터를 구축 • 노이즈 가능성이 있는 데이터를 배제하여 모델 학습 진행
2	근로손실일수 상관분석	• 근로손실일수와의 상관관계를 기반으로 상관점수 0.3 이상인 50개의 주요 특성을 추출하여, 근로손실일에 영향을 미치는 특성들만을 학습에 사용함
3	RFE를 통한 특성 제거	• Recursive Feature Elimination(RFE) 기법을 통해 영향력이 낮은 특성을 제거하며 모델 성능 변화를 확인하고 특성 수를 조정
4	클러스터링 및 군집분석	• 클러스터링을 적용하여 Label 0 데이터를 군집별로 분류하고, 각 군집을 개별로 학습 데이터로 활용하여 비교 분석 진행
5	다중분류학습 적용	• 사고 경중을 반영한 다중분류 모델을 적용하여, 사고의 경중에 따라 보다 정교한 예측이 가능한지 학습 진행
6	저위험 사업장 교집합 처리	• 승인통계에 등록되지 않은 사업장과 위험 수준 평가에서 3점 이하인 저위험 사업장들을 Label 0으로 선별하여 학습에 사용

(3) 승인통계 데이터, 사업장 위험 수준 현장평가 데이터 활용 학습

○ Label 0(안전사업장)으로 분류한 사업장 중에서도, 최근 3년 이내에 발생한 사고사례가 존재하며, 사망자 발생이나 근로손실일수가 큰 사고가 확인된 사례들이 포함됨.

〈표 2-4〉 승인통계, 사업장 위험 수준 현장평가 데이터 활용 학습 결과

구분	내용
처리방법	• 승인통계에서 정보가 있는 사업장만을 선별하여 Label 0 샘플을 구성함 • L1, L2 규제를 적용하였으나, 하이퍼파라미터 최적화는 진행되지 않음
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.82 2. 정밀도(Precision): - Label 0 (0.92): 모델이 안전 사업장으로 예측한 것 중 92%가 실제 안전 사업장임을 의미함 - Label 1 (0.76): 모델이 위험 사업장으로 예측한 것 중 76%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.70): 실제 안전 사업장 중 70%가 정확하게 예측함 - Label 1 (0.94): 실제 위험 사업장 중 94%가 정확하게 예측함 • 4. F1-Score: - Label 0 : 0.80 - Label 1 : 0.84
정리 의견	• Validation Log Loss: 안정적으로 감소한 후 수렴 • Validation Accuracy: 0.82 수준에서 유지 • 정밀도(Precision): 안전 사업장에 대한 예측 정확도가 높음 • 재현율(Recall): Label 1의 재현율이 0.94로 매우 높아 위험 사업장 탐지에 우수 • F1 Score: Label1의 F1 스코어는 0.84로, 위험 사업장을 잘 탐지하는 모델 • 전체적으로 준수하지만 고점이 더 높을 여지 있음

○ 승인통계 데이터가 없는 사업장 중에서도 Label 0(안전사업장)으로 분류된 데이터 중 일부는 Label 1(위험사업장)과 유사한 특성 패턴을 보일 수 있음.

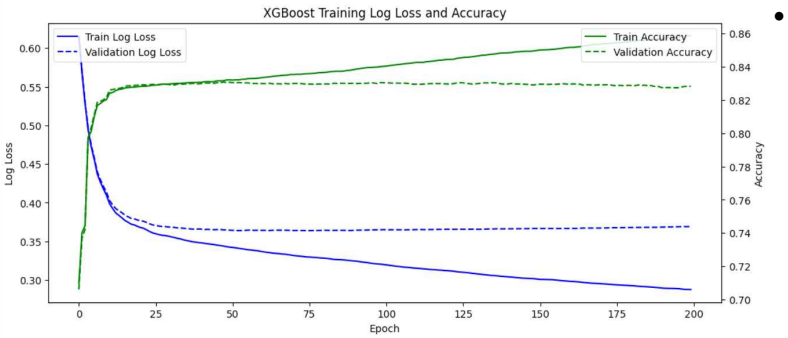
○ 제조업 410,117건 중에서, 승인통계 데이터가 있는 68,924건의 사업장만을 학습데이터로 활용하고, 이 중에서도 경미한 사고가 발생한 사업장을 안전사업장으로 간주하여 학습에 사용함.

(4) 근로손실일수 상관분석 결과 반영한 학습

○ 사고의 경중에 영향을 미치는 특성값을 확인하기 위해 연속값을 가지는 사업장 근로손실일수를 대상으로 각 특성별 상관 분석 수행.

○ 대부분 특성이 양의 상관관계를 가지며, 상관계수 0.3(약한 상관관계) 이상 특성을 학습에 반영하여 수행.

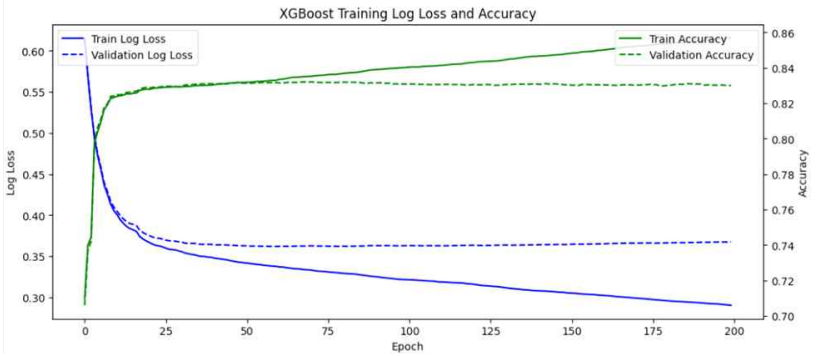
〈표 2-5〉 근로손실일수 상관분석결과 반영 학습 결과

구분	내용
처리방법	• 승인통계에서 정보가 있는 사업장만을 선별하여 Label 0 샘플을 구성함 • 상관분석 결과, 근로손실일수에 영향을 미치는 주요 특성 약 50개만 학습에 반영 • L1, L2 규제를 적용하였으나, 하이퍼파라미터 최적화는 진행되지 않음
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.83 2. 정밀도(Precision): - Label 0 (0.96): 모델이 안전 사업장으로 예측한 것 중 96%가 실제 안전 사업장임을 의미함 - Label 1 (0.76): 모델이 위험 사업장으로 예측한 것 중 76%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.68): 실제 안전 사업장 중 68%가 정확하게 예측함 - Label 1 (0.97): 실제 위험 사업장 중 97%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.80 - Label 1 : 0.85
정리 의견	• Validation Log Loss: 초기 급격한 감소 후 수렴. 과적합 신호는 크게 보이지 않음 • Validation Accuracy: 0.83으로 높고 일정함 • 정밀도(Precision): 안전 사업장에 대한 높은 예측 성능 • 재현율(Recall): 위험 사업장 예측에서 매우 좋은 성능 보임 • F1 Score: Label 1의 F1 스코어는 0.85로 우수 • 위험 사업장 예측에서 높은 성능을 보이며, 고점이 높아 데이터 추가 시 성능 향상 가능성 큼

(5) 특성중요도가 높은 특성을 반영한 학습

○ 다중공선성 문제를 확인하고, 지나치게 유사한 특성들의 경우 특성중요도가 높은 특성만 남기고 제거 후 학습 수행.

〈표 2-6〉 특성중요도가 높은 특성을 반영한 학습 결과

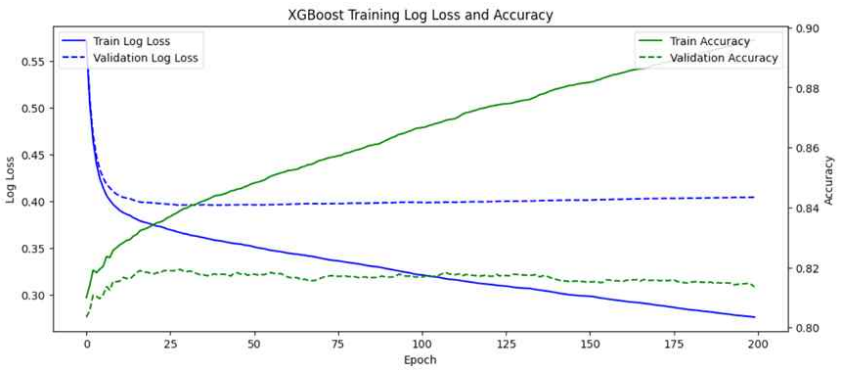
구분	내용
처리방법	• 승인통계에서 정보가 있는 사업장만을 선별하여 Label 0 샘플을 구성함 • 상관분석 결과, 근로손실일수에 영향을 미치는 주요 특성 약 50개 선택 • 특성 50개 내에서 상관계수가 높은 특성 중 특성중요도가 높은 특성만 선별하여 학습 반영 • L1, L2 규제를 적용하였으나, 하이퍼파라미터 최적화는 진행되지 않음
학습 그래프	• 
모델 평가	1. 정확도(Accuracy): 0.83 2. 정밀도(Precision): - Label 0 (0.97): 모델이 안전 사업장으로 예측한 것 중 97%가 실제 안전 사업장임을 의미함 - Label 1 (0.76): 모델이 위험 사업장으로 예측한 것 중 76%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.68): 실제 안전 사업장 중 68%가 정확하게 예측함 - Label 1 (0.98): 실제 위험 사업장 중 98%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.80 - Label 1 : 0.85
정리 의견	• Validation Log Loss: 안정적으로 감소하며 과적합 가능성 크지 않음 • Validation Accuracy: 0.83으로 높고 일정함 • 정밀도(Precision): 안전 사업장에서 높은 예측 성능 • 재현율(Recall): 위험 사업장 예측에서 매우 좋은 성능 보임 • F1 Score: Label 1의 F1 스코어는 0.85로 우수 • 두 번째 그래프와 유사한 성능 보이며, 위험 사업장 예측에서 높은 성능을 보이며, 고점이 높아 데이터 추가 시 성능 향상 가능성 큼

(6) 클러스터링 및 군집분석 결과 반영

○ Label 0(안전사업장)으로 분류한 사업장 중 특성들이 유사한 사업장을 그룹으로 묶고, 별도 학습데이터로 구성 후 모델학습 수행.

○ 잘못된 라벨링이나 이질적인 데이터를 배제하여 Label 0과 Label 1간의 차이를 더 명확하게 반영하는지 비교 수행.

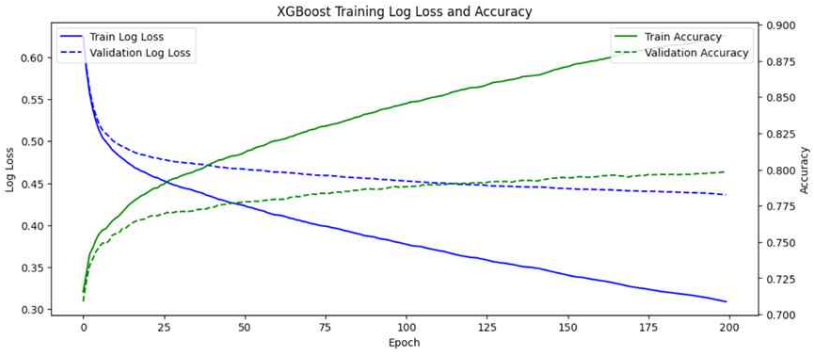
〈표 2-7〉 클러스터링 및 군집분석 결과 반영 학습 결과

구분	내용
처리방법	• K-Means 알고리즘으로 Label 0 사업장 그룹화 후, 개별로 학습에 반영 • L1, L2 규제를 적용하였으나, 하이퍼파라미터 최적화는 진행되지 않음
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.81 2. 정밀도(Precision): - Label 0 (0.83): 모델이 안전 사업장으로 예측한 것 중 83%가 실제 안전 사업장임을 의미함 - Label 1 (0.80): 모델이 위험 사업장으로 예측한 것 중 80%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.82): 실제 안전 사업장 중 82%를 정확하게 예측함 - Label 1 (0.81): 실제 위험 사업장 중 81%를 정확하게 예측함 4. F1-Score: - Label 0 : 0.82 - Label 1 : 0.80
정리 의견	• Validation Log Loss: 안정적 감소 후 수렴 • Validation Accuracy: 0.81에서 안정적 유지 • 정밀도(Precision): 균형적 • 재현율(Recall): 균형적 • F1 Score: 0.82 • 안전사업장과 위험사업장에 대해 균형은 잡히나, 최고 성능은 아님

(7) REF 적용 특성수 고려 최적모델 발굴

○ RFE를 적용하여, 특성 수를 줄이며, best 모델을 구함.

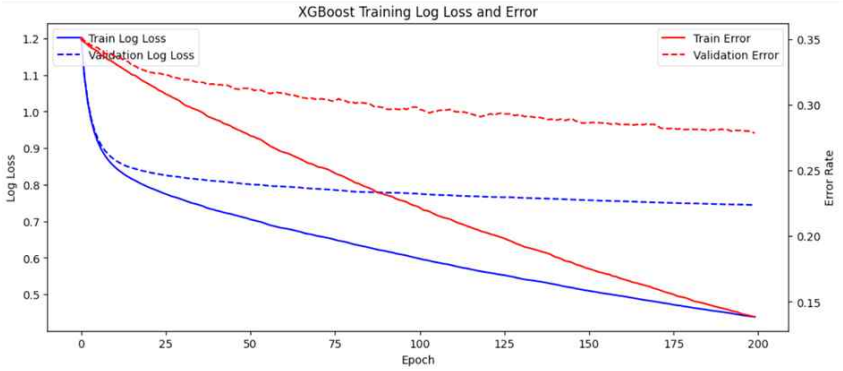
〈표 2-8〉 REF 적용 특성수 조절을 통한 학습결과

구분	내용
처리방법	• K-Means 알고리즘으로 Label 0 사업장 그룹화 후, 개별로 학습에 반영 • 특성수를 줄여나가며 특성 100개까지 중요하지 않은 특성 제거 • L1, L2 규제를 적용하였으나, 하이퍼파라미터 최적화는 진행되지 않음
학습 그래프	 •
모델 평가	1. 정확도(Accuracy): 0.80 2. 정밀도(Precision): - Label 0 (0.81): 모델이 안전 사업장으로 예측한 것 중 81%가 실제 안전 사업장임을 의미함 - Label 1 (0.79): 모델이 위험 사업장으로 예측한 것 중 79%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.78): 실제 안전 사업장 중 78%가 정확하게 예측함 - Label 1 (0.81): 실제 위험 사업장 중 81%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.79 - Label 1 : 0.81
정리 의견	• Validation Log Loss: 안정적으로 수렴하는 듯 하지만 성능 개선이 크지 않음 • Validation Accuracy: 0.8로 약간 낮음 • 정밀도(Precision): 균형 • 재현율(Recall): 균형 • F1 Score: 0.8 • 균형적이라 볼 수 있지만, 전반적으로 높은 성능 보이지 않음

(8) 다중분류학습 적용

○ 사고 경중에 따라 Label을 세분화하여 학습 시 모델 성능 변화 확인.

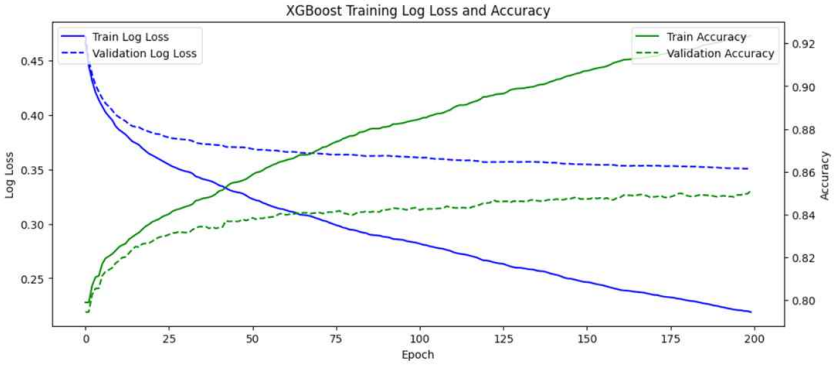
〈표 2-9〉 다중분류학습 적용 결과

구분	내용
처리방법	• 사망자 및 1년 이상, 6개월~1년, 6개월~3개월, 3개월 미만 등 분류기준 세분화 • L1, L2 규제를 적용하였으나, 하이퍼파라미터 최적화는 진행되지 않음
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.72 2. F1-Score: 70.8

(9) 저위험 사업장 교집합 처리

○ 승인통계에 등록되지 않은 사업장이면서 동시에 사업장 위험 수준 평가 결과 3개 항목에 대해 모두 3점 이하인 사업장을 교집합하여 학습 샘플 재구성.

〈표 2-10〉 저위험 사업장 교집합처리 반영 학습 결과

구분	내용
처리방법	• 승인통계 내 등록되지 않으며, 사업장 위험 수준 평가 결과 저위험에 가까운 사업장 교집합으로 Label 0 학습데이터 적용 • L1, L2 규제를 적용하였으나, 하이퍼파라미터 최적화는 진행되지 않음
학습 그래프	 •
모델 평가	1. 정확도(Accuracy): 0.85 2. 정밀도(Precision): - Label 0 (0.73): 모델이 안전 사업장으로 예측한 것 중 73%가 실제 안전 사업장임을 의미함 - Label 1 (0.87): 모델이 위험 사업장으로 예측한 것 중 87%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.44): 실제 안전 사업장 중 44%가 정확하게 예측함 - Label 1 (0.96): 실제 위험 사업장 중 96%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.55 - Label 1 : 0.91
정리 의견	• Validation Log Loss: 수렴하며 낮은 수준에서 안정적 • Validation Accuracy: 0.85로 전체 모델 중 가장 높음 • 정밀도(Precision): 위험 사업장 예측 성능이 더 뛰어남 • 재현율(Recall): Label 0이 낮음 • F1 Score: Label0은 낮지만 Label1은 0.91로 매우 높음 • 위험 사업장에 대한 매우 높은 예측 성능을 보이며, 데이터가 추가될 경우 고위험 사업장 예측에서 우수한 성능을 보일 가능성이 큼

(10) 고위험사업장 선별 모델 점검 결과 적용 방안

○ 학습데이터 재조정 후 학습 결과 검토 후, 방법 및 효과가 유사한 경우, 과업 내용에서 벗어나거나 데이터 재구성이 학습 의도, 해석 가능성을 떨어뜨리는 경우는 제외하였음. (위의 6,7,8번 참고)

- RFE를 통한 특성 제거
- 클러스터링 및 군집 분석
- 다중분류학습 적용

○ 모델이 재조정된 학습데이터를 받아들이는 경향을 확인한 후, 하이퍼파라미터 최적화를 통해 모델의 일반화 성능을 향상 시키는데 집중하였음.

3) 고위험사업장 예측 언어모델 제시

○ 산업재해조사표, 사업수행 경과 보고서 등 비정형데이터는 사업장 정보, 개인정보가 포함되어있어서 데이터 보안을 위해 전처리가 필요하므로 공단 내부검토 후 제공 범위를 정함.

○ 고위험사업장 예측 언어모델은 아래와 같이 1안, 2안을 수행사에서 제시하여, 1안과 2안 모두 수행 후, sLLM(BERT 모델) 파인튜닝 적용 후 모델 성능평가 결과 앙상블 모델 적용.

- (1안) 산업재해조사표 및 승인통계에서 제공되는 사고 개요, 원인, 기인물, 사고 규모 등의 데이터를 활용하여 고위험 사업장 예측 모델을 위한 학습데이터 구성.
- (2안) 각 특성값을 텍스트 형식으로 변환하여, 고위험 사업장을 위한 모델 학습 데이터를 생성함.

4) 고위험사업장 예측 언어모델 활용방안

○ 고위험사업장 예측 언어모델 개발 과정에 앞서, 1안, 2안은 기존 고위험

사업장 선정 모델 내 반영하거나 함께 활용하기 어려우므로, 사업장 위험성 평가 결과 내 평가자 코멘트 정보를 활용함.

- 데이터 검토 결과, 제조업과 서비스업의 사업장 관련 평가 내용은 작성된 데이터의 양이 충분치 않아 모델 학습에 한계가 있으며, 작성 내용과 평가 내용 결과가 상이한 점이 있어서 활용이 어려움.
- 이에, 고위험 사업장 선정 모델과 독립적으로 동작할 수 있는 생성형 언어모델을 개발하기로 하고, 비전문가도 쉽게 모델 결과를 해석하고 활용할 수 있도록 학습데이터를 직접 생성하고, 연구환경 내에서 효율적인 학습을 위해 LoRA(Low-Rank Adaption) 기법을 적용하여 학습을 진행함.

Ⅲ. 연구추진 및 실적

Ⅲ. 연구추진 및 실적

1. 산업안전 데이터 수집 및 전처리

1) 산업안전 데이터 조사 및 수집

(1) 고위험 사업장 선정 모델 학습에 사용된 데이터

○ 본 연구에서는 안전보건공단이 제공한 고위험 사업장 선정 모델에 활용된 데이터를 사용하였음. 이 모델은 2019년 ISP(Information Strategy Plan)을 수행하여 방향성을 설정하고, 2021년 시범 모델을 개발하였으며, 2023년에는 제조업 분야에서 상용화를 이루고 이후 서비스업과 건설업으로도 확장하고 있음.

- 제조업 데이터는 23개 사업, 410,117건의 사업장 데이터로 구성되어 있으며, 총 419개 특성이 모델 학습에 사용됨.
- 서비스업 데이터는 21개 사업, 1,127,652건의 데이터를 포함하며, 총 473개 특성이 모델 학습에 사용됨.

○ 제공받은 데이터는 사업장을 식별할 수 없도록 암호화된 키 형태로 비식별화하여 제공되었으며, 수치형으로 가공된 데이터와 Raw 데이터 두 가지 형태로 제공되었으며, 모델 학습에는 주로 수치형 데이터를 사용하였고, Raw 데이터는 학습 데이터의 개선이 필요한 경우 일부 활용하였음.

- Label 0: 해당 사업장의 근로 손실일이 동일 업종 및 규모의 평균 이하인 경우
- Label 1: 해당 사업장의 근로 손실일이 동일 업종 및 규모의 평균 초과인 경우

〈표 3-1〉 업종별 '고위험사업장 선정 모델'에 활용된 사업리스트

제조업			서비스업		
번호	사업명	특성 합계	번호	사업명	특성 합계
1	제조업 사업장 리스트	11	1	서비스 사업장 리스트	5
2	패트롤 수행 결과	32	2	패트롤 수행 결과	28
3	고용보험_근로자정보	25	3	고용보험_근로자정보	25
4	안전보건관계자	14	4	안전보건관계자	13
5	재정지원	9	5	재정지원	13
6	유해위험기계기구 (안전검사+자율)	23	6	유해위험기계기구 (안전검사+자율)	38
7	소방청 위험물 제조소 등 정보	26	7	소방청 위험물 제조소 등 정보	26
8	지게차 실태조사	11	8	지게차 실태조사	11
9	KOSHA_MS_18001	1	9	KOSHA_MS_18001	1
10	위험성평가	8	10	위험성평가	7
11	민간위탁기술지도	101	11	민간위탁기술지도	140
12	공단교육(안전보건, 인터넷, 직무)	15	12	공단교육(안전보건, 인터넷, 직무)	14
13	유해위험방지계획서	13	13	유해위험방지계획서	14
14	작업환경측정	23	14	작업환경측정	22
15	고용보험ERP근로자수	6	15	고용보험ERP근로자수	4
16	특수건강진단	23	16	특수건강진단	23
17	산업안전보건실태조사	21	17	산업안전보건실태조사	23
18	산업재해조사표	12	18	산업재해조사표	12
19	작업환경실태조사	39	19	작업환경실태조사	49
20	민간위탁기관평가데이터	1	20		-
21		-	21	공공기관등급데이터	1
22	재해데이터	1	22		-
23		-	23	사업장수준조사평가 +재해율	4
24	감성평가	3	24		-
25	PSM	1	25		-
소계		419	소계		473

○ 데이터의 특성을 분석하여 특성별 데이터 타입은 텍스트, 수치형, 범주형, 이진형, 실수형, 상수형, 숫자형 7개로 분류하였으며, 데이터가 수치형으로 명시된 데이터라도 0과 1로 분류되는 경우는 이진형으로 정리하여 처리하였음. 또한, 단일 값으로만 구성된 특성은 상수형으로 구분하였으며, 나머지는 수치형으로 타입을 지정하였음.

(2) 제조업 데이터 특성

○ 제조업의 데이터 특성은 419개로 특성별 데이터 타입은 〈표 3-2〉와 같음. 이 중 이진형이 159개로 가장 많으며, 수치형 137개, 실수형 57개, 범주형 53개, 상수형 10개, 텍스트 2개, 숫자형 1개 순으로 되어 있음.

〈표 3-2〉 제조업 특성명별 데이터 타입

번호	사업명	특성명	데이터 타입
1	제조업 사업장 리스트	사업장관리번호	텍스트
		사업장개시번호	텍스트
		일선기관	범주형
		노동지청	범주형
		중업종명	범주형
		소업종명	범주형
		표준산업분류	범주형
		근로자수	수치형
		규모1	범주형
		행정구역	범주형
		행정구역_세부	범주형
2	패트롤 수행 결과 (현장점검 정보)	동행한타기관	이진형
		선정기준_공단 자체	이진형
		선정기준_재해예방기관 등 기타	이진형
		점검결과조치_개선확인 후 종결(미개선시 감독연계)	이진형
		점검결과조치_사업장 자체개선 후 종결/점검 종결	이진형
		경영자마인드	이진형
		안전보건관리및개선노력	실수형
		안전관리수준평가사업장위험도_현장위험관리수준	실수형
		안전보건수준평가종합	실수형
		점검차수	범주형

번호	사업명	특성명	데이터 타입
	패트롤 수행 결과 (시정부적합정보)	사고유발요인_갯수	수치형
		위험기인물_그 밖의 위험	이진형
		위험기인물_끼임	이진형
		위험기인물_떨어짐	이진형
		위험기인물_부딪힘	이진형
		위험기인물_질식	이진형
		위험기인물_화재	이진형
		안전관리수준평가사업장위험도_현장위험관리수준	실수형
		점검차수	범주형
	패트롤 수행 결과 (위험설비보유정보)	보유건수	수치형
		위험설비_분쇄,파쇄기	이진형
		위험설비_사출성형기	이진형
		위험설비_산업용로봇	이진형
		위험설비_승강기(리프트 포함)	이진형
		위험설비_식품가공용기계	이진형
		위험설비_지게차	이진형
		위험설비_컨베이어	이진형
		위험설비_크레인(천장,갠트리)	이진형
		위험설비_타워크레인	이진형
		위험설비_프레스	이진형
		위험설비_혼합기	이진형
		점검차수	범주형
3	고용보험_ 근로자정보	피보험자_합계	수치형
		만나이_평균	실수형
		근속기간_년수_평균	실수형
		성별_남_합계	수치형
		성별_여_합계	수치형
		연령대_10대_20대_합계	수치형
		연령대_30대_합계	수치형
		연령대_40대_합계	수치형
		연령대_50대_합계	수치형
		연령대_60대이상_합계	수치형
		근속기간범주_1년이하_합계	상수형
		근속기간범주_1년초과3년이하_합계	상수형
		근속기간범주_3년초과5년이하_합계	상수형
		근속기간범주_5년초과10년이하_합계	상수형
		근속기간범주_10년초과20년이하_합계	상수형
		근속기간범주_20년초과_합계	상수형
		직종_대분류_건설·채굴직_합계	수치형
		직종_대분류_교육·법률·사회복지·경찰·소방직및 군인_합계	수치형
		직종_대분류_농림어업직_합계	수치형
		직종_대분류_미용·여행·숙박·음식·경비·청소직_	수치형

번호	사업명	특성명	데이터 타입
		합계	
		직종_대분류_보건·의료직_합계	수치형
		직종_대분류_설치·정비·생산직_합계	수치형
		직종_대분류_연구직및공학기술직_합계	수치형
		직종_대분류_영업·판매·운전·운송직_합계	수치형
		직종_대분류_예술·디자인·방송·스포츠직_합계	수치형
4	안전보건관계자	전담유무	범주형
		건설안전관리자	이진형
		명예산업안전감독관	이진형
		보건관리자	이진형
		사업장담당자	범주형
		산업보건의	이진형
		안전관리자	이진형
		안전보건관리책임자	이진형
		안전보건총괄책임자	이진형
		관리자	범주형
		보건담당자	범주형
		선임자총수	수치형
		선임자종류수	수치형
		안전보건관계자_데이터존재여부	이진형
5	재정지원 (안전투자혁신사업)	지원금액	수치형
		사업구분_고소작업대	이진형
		사업구분_고위험 TOP3 업종	이진형
		사업구분_노후 위험기계기구(30년이상)	이진형
		사업구분_리프트	이진형
		사업구분_뿌리공정	이진형
		사업구분_이동식크레인	이진형
	재정지원 (융자지원)	대하금액(천원)	수치형
	재정지원 (클린사업장)	교부금액	수치형
6	유해위험기계기구 (안전검사+자율)	컨베이어종류_밸트	수치형
		컨베이어종류_체인	수치형
		컨베이어종류_롤러	수치형
		컨베이어종류_트롤리	수치형
		컨베이어종류_버킷	수치형
		컨베이어종류_나사	수치형
		고소작업대	수치형
		곤돌라	수치형
		국소배기장치	이진형

번호	사업명	특성명	데이터 타입
		롤러기	이진형
		리프트	이진형
		사출성형기	이진형
		산업용로봇	이진형
		압력용기	이진형
		원심기	이진형
		전단기	이진형
		컨베이어	이진형
		크레인	이진형
		프레스	수치형
		심사결과_적합	수치형
		심사결과_부적합	수치형
		안전검사_사업수행여부	이진형
		자율검사_사업수행여부	이진형
7	소방청 위험물 제조소	예방규정제출대상여부	이진형
		대량위험물제조소등여부	이진형
		석유화학단지내사업장여부	이진형
		소화난이도등급_1등급	수치형
		소화난이도등급_2등급	수치형
		소화난이도등급_3등급	수치형
		설치기간_10년이상	수치형
		설치기간_20년이상	수치형
		설치기간_5년미만	수치형
		설치기간_5년이상	수치형
		위험물제조소_총합	수치형
		위험물제조소_종류	수치형
		위험물허가_유별_제1류	수치형
		위험물허가_유별_제2류	수치형
		위험물허가_유별_제3류	수치형
		위험물허가_유별_제4류	수치형
		위험물허가_유별_제5류	수치형
		위험물허가_유별_제6류	수치형
		위험물허가_유별_합계	수치형
		위험물허가_유별_여부	상수형
		탱크총합	수치형
		탱크여부	이진형
		이송취급소수	수치형
		이송취급소여부	이진형
		이동탱크수	수치형
		이동탱크여부	이진형
8	지게차 실태조사	지게차대수_자가	수치형
		지게차대수_그외	수치형
		지게차보유대수	수치형
		지게차용량	실수형

번호	사업명	특성명	데이터 타입
		충돌방지장치점수	실수형
		안전띠점수	실수형
		법정방호장치점수	실수형
		운전자격점수	실수형
		관리자이해점수	실수형
		운전자관찰점수	실수형
		점수총합	실수형
9	KOSHA_MS_18001	KOSHA_MS18001_사업수행여부	이진형
10	위험성평가 (컨설팅)	사업주의관심도	실수형
		위험성평가실행수준	실수형
		구성원의참여및이해수준	실수형
		재해발생수준	실수형
	위험성평가 (인정/불인정)	사업주의 관심도	실수형
		위험성평가 실행수준	실수형
		구성원의 참여 및 이해수준	실수형
		심사결과_불인정	범주형
11	민간위탁기술지도 (안전)	선정기준코드_물질관련	이진형
		선정기준명_신규사업장	이진형
		선정기준명_유해물질존재사업장	이진형
		선정기준명_재해발생사업장	이진형
		선정기준명_특별대책사업장	이진형
		지원횟수	범주형
		전담	이진형
		겸직	이진형
		등급_중급	이진형
		등급_초급	이진형
		교육인원합	수치형
		자료제공_총합	수치형
		검사실시_종류수	범주형
		검사미실시_종류수	범주형
		검사비대상_종류수	범주형
		검사실시_합	수치형
		검사미실시_합	수치형
		검사비대상_합	수치형
		전체_개선	수치형
		전체_조치의뢰	범주형
	민간위탁기술지도 (보건)	처리사유_개선완료	이진형
		처리사유_조치의뢰	이진형
		선정기준코드_산재관련	이진형
		선정기준코드_실비관련	이진형
		선정기준코드_물질관련	이진형
		선정기준명_신규사업장	이진형
		선정기준명_유해물질존재사업장	이진형
		선정기준명_재해발생사업장	이진형

번호	사업명	특성명	데이터 타입
		선정기준명_특별대책사업장	이진형
		지원횟수	범주형
		전담	이진형
		겸직	이진형
		등급_중급	이진형
		등급_초급	이진형
		교육인원_총수	수치형
		자료제공_총합	수치형
		밀폐실태밀폐공간작업유무	이진형
		밀폐실태_밀폐공간장소_통의내부등	이진형
		밀폐실태_밀폐공간장소_정화조등	이진형
		밀폐실태_밀폐공간장소_기타밀폐공간	이진형
		밀폐실태_밀폐공간장소_반응기등내부	이진형
		밀폐실태_밀폐공간장소_콘크리트양생	이진형
		밀폐실태_밀폐공간장소_강재등시설	이진형
		밀폐실태_밀폐공간장소_불활성기체설비	이진형
		밀폐실태-질식사고 인지도	범주형
		밀폐실태-질식사고 위험관리 인지도	범주형
		밀폐실태-질식사고 교육이수	범주형
		밀폐실태-가스농도측정기 보유	수치형
		급기팬 보유	수치형
		밀폐실태-위험도평가 총점	수치형
		개선필요_합	수치형
		개선_합	수치형
		최종실태평가결과	범주형
		실태평가결과_사업주의지	범주형
	민간위탁기술지도 (화학)	처리사유내용_종결	이진형
		처리사유내용_조치의뢰	이진형
		선정기준코드_산재관련	이진형
		선정기준코드_실비관련	이진형
		선정기준코드_물질관련	이진형
		선정기준명_신규사업장	이진형
		선정기준명_유해물질존재사업장	이진형
		선정기준명_재해발생사업장	이진형
		선정기준명_특별대책사업장	이진형
		지원횟수	범주형
		전담	이진형
		겸직	이진형
		등급_중급	이진형
		등급_초급	이진형
		교육인원_총수	수치형
		자료제공_총합	수치형
		기술지원_사고성재해예방	이진형
		기술지원_화학사고예방	이진형

번호	사업명	특성명	데이터 타입
		유해위험물질수	수치형
		화학설비건수	수치형
		위험기계기구_보유건수	범주형
		개선_총계	수치형
		조치의뢰_총계	이진형
		국소배기장치	이진형
		롤러기	이진형
		리프트	이진형
		분쇄파쇄기	이진형
		사출성형기	이진형
		산업용로봇	이진형
		식품제조용설비	이진형
		원심기	이진형
		전단기	이진형
		지게차	이진형
		컨베이어	이진형
		크레인	이진형
		프레스	숫자형
		혼합기	이진형
		압력용기	이진형
		위험기계기구총합	수치형
		밀폐_정화조등	이진형
		밀폐_통의내부등	이진형
		밀폐_불활성기체설비	이진형
		밀폐_강재등시설	이진형
		밀폐_반응기등내부	이진형
		밀폐_콘크리트양생	이진형
		밀폐_중독위험장소	이진형
		밀폐_기타밀폐공간	이진형
12	공단교육 (안전보건교육)	교육분야코드_관리자	실수형
		교육분야코드_안전보건관계자	실수형
		교육분야코드_일반근로자	실수형
		교육분야코드_취약계층	실수형
	공단교육 (인터넷교육센터)	교육대상_근로자	실수형
		교육대상_책임자	실수형
		교육대상_특수형태근로자	실수형
		수료여부_미수료	실수형
		수료비율	실수형
	공단교육 (직무교육센터)	과정구분_온라인	실수형
		과정구분_집체	실수형
		교육대상_전문기관종사자	실수형
		수료여부_미수료	실수형
		교육대상_안전보건관계자	실수형
		교육수료비율	실수형

번호	사업명	특성명	데이터 타입
13	유해위험방지계획서	대상설비합계	수치형
		대상규모명_over_2000	수치형
		대상규모명_under_2000	범주형
		대상규모명_under_500	범주형
		전기계약용량변경	범주형
		사업구분_변경	수치형
		사업구분_설치	범주형
		사업구분_이전	범주형
		최종확인회차	범주형
		심사결과_반려부적정	수치형
		심사결과_적정조건부적정	수치형
		고용부조치통보	범주형
		유해위험방지계획서_사업수행여부	이진형
14	작업환경측정 (측정)	물질군명_노출기준제정물질	수치형
		물질군명_허가대상 유해물질	범주형
		지원대상구분_대상	수치형
		초과율_평균	실수형
		취급인원	수치형
		물질군명_물리적인자_합	수치형
		물질군명_화학적인자_합	수치형
		물질군명_분진_합	수치형
		취급구분_사용	수치형
		취급구분_제조	수치형
		취급용도_기타	수치형
	작업화경측정 (화학물질취급현황)	취급용도_세척	수치형
		취급용도_시약	수치형
		취급용도_실험	수치형
		취급용도_용접	수치형
		취급용도_원료	수치형
		취급물질군명_기타유해물질	수치형
		취급물질군명_노출기준제정물질	수치형
		취급물질군명_제조금지 유해물질	수치형
		취급물질군명_허가대상 유해물질	범주형
		취급물질군명_물리적인자_통합	수치형
		취급물질군명_분진인자_통합	수치형
		취급물질군명_화학적인자_통합	수치형
15	고용보험ERP 근로자수	고용상시인원수	수치형
		남성근로자수	수치형
		여성근로자수	수치형
		장년근로자수	상수형
		외국인근로자수	수치형
		장애인근로자수	상수형
16	특수건강진단 (특검)	총검진자수(명)	수치형
		유해물질군명_노출기준제정물질	이진형

번호	사업명	특성명	데이터 타입
		유해물질군명_야간작업	수치형
		유해물질군명_제조금지 유해물질	상수형
		유해물질군명_허가대상 유해물질	이진형
		A판정비율	실수형
		C1판정비율	실수형
		C2판정비율	실수형
		D1판정비율	실수형
		D2판정비율	실수형
		CN판정비율	실수형
		DN판정비율	실수형
		U판정비율	실수형
		유해물질군명_물리적인자_합	범주형
		유해물질군명_화학적인자_합	수치형
		유해물질군명_분진_합	범주형
	특수건강진단 (사업장별검진내역)	검진종목별 수진자수(일반)	수치형
		검진종목별 수진자수(특수)	수치형
		검진종목별 수진자수(배치전)	수치형
		검진종목별 수진자수(수시)	수치형
		검진종목별 수진자수(임시)	수치형
		검진종목별 수진자수(수첩)	수치형
		검진종목별_총합	수치형
17	산업안전보건실태조사	교대근무제여부	이진형
		노동조합여부	이진형
		산업안전보건위원회여부	이진형
		작업환경관련위험요인	실수형
		신체적부담관련위험요인	실수형
		생화학물질관련위험요인	실수형
		기계전기기타위험요인	실수형
		위험성평가	이진형
		스트레스심각도	실수형
		스트레스관리노력정도	실수형
		경영진안전보건의지	실수형
		사업장내안전문화	실수형
		근로자안전보건의지	실수형
		일반건강진단사후관리	이진형
		특수건강진단사후관리	이진형
		유해인자축소노력여부	이진형
		상주협력업체	이진형
		상주협력업체수	수치형
		상주협력업체근로자수	수치형
		원청회사여부	이진형
		법인지여부	이진형
18	산업재해조사표	재해자동종경력년수_10~20	수치형
		재해자동종경력년수_1~3	수치형

번호	사업명	특성명	데이터 타입
		재해자동종경력년수_20~	수치형
		재해자동종경력년수_3~5	수치형
		재해자동종경력년수_5~10	수치형
		상해부위_기타	수치형
		상해부위_다리	수치형
		상해부위_다발성	수치형
		상해부위_머리	수치형
		상해부위_몸통	수치형
		상해부위_전신	수치형
		상해부위_팔	수치형
19	작업환경실태조사 (일반현황)	전기계약용량	범주형
		야간작업유무	이진형
		정비_보수여부	이진형
		하청사업장수	수치형
		하청근로자수	수치형
		교대근무여부	이진형
		근골격계부담작업대상여부	이진형
		유해요인조사실시여부	이진형
		복지시설_개수	범주형
		원청	이진형
		하청	이진형
	작업환경실태조사 (화학물질취급)	취급	이진형
		생산	이진형
		허용대상물질여부	이진형
		허용기준물질여부	이진형
		관리대상물질여부	이진형
		안전검사물질여부	이진형
		안전관리물질여부	이진형
		기타물질여부	이진형
		특검대상물질여부	이진형
		측정대상물질여부	이진형
		PSM대상물질여부	이진형
		건강관리수첩대상물질여부	이진형
		사고대상물질여부	이진형
		금지대상물질여부	이진형
		근로자_월_취급시간	실수형
	작업환경실태조사 (기계기구설비현황)	기계설비_제조_보유갯수	수치형
		기계설비_제조_보유종류	범주형
		기계설비_비제조_보유갯수	수치형
		기계설비_비제조_보유종류	범주형
	작업환경실태조사 (작업환경)	소음발생공정수	수치형
		밀폐공간수	수치형
		작업환경_고열/한랭/다습및방사선취급작업	범주형
		작업환경_밀폐공간(산소결핍위험장소)현황	수치형

번호	사업명	특성명	데이터 타입
		작업환경_분진/흄발생작업	범주형
		작업환경_사내도급작업	범주형
		작업환경_소음작업	범주형
		작업환경_제조나노물질의제조및취급작업	범주형
		작업환경_진동발생작업	범주형
20	민간위탁기관평가 데이터	위탁기관평가_점수	범주형
22	재해데이터	재해율3년평균	실수형
24	감성평가	사업주관리자마인드	범주형
		근로자안전보건행동수준	실수형
		작업장 및 근로자환경수준	범주형
25	PSM	PSM_사업수행여부	이진형

(3) 서비스업 데이터 특성

○ 서비업의 데이터 특성은 473개로 특성별 데이터 타입은 〈표 3-3〉와 같음. 이 중 수치형이 222개로 가장 많으며, 이진형이 138개, 실수형 54개, 상수형 46개, 범주형 11개, 텍스트 2개 순으로 되어 있음.

〈표 3-3〉 서비스업 특성명별 데이터 타입

번호	사업명	특성명	데이터 타입
1	서비스업 사업장 리스트	사업장관리번호	텍스트
		사업장개시번호	텍스트
		일선기관	수치형
		소업종명	수치형
		규모1	범주형
2	패트롤 수행 결과 (서비스)	동행한타기관	상수형
		선정기준_재해예방기관 등 기타	이진형
		경영자마인드	범주형
		안전보건관리및개선노력	범주형
		안전관리수준평가사업장위험도_현장위험관리수준	범주형
		안전보건수준평가종합	범주형
		점검차수	범주형
	패트롤 수행 결과 (시정 부적합 정보)	사고유발요인_갯수	수치형
		위험기인물_그 밖의 위험	수치형
		위험기인물_끼임	수치형
		위험기인물_떨어짐	수치형
		위험기인물_부딪힘	수치형
		위험기인물_질식	이진형

번호	사업명	특성명	데이터 타입
		위험기인물_화재_폭발	수치형
		안전관리수준평가사업장위험도_현장위험관리수준	수치형
		점검차수	범주형
	패트롤 수행 결과 (위험설비보유정보)	보유건수	수치형
		위험설비_분쇄_파쇄기	이진형
		위험설비_산업용로봇	이진형
		위험설비_승강기(리프트 포함)	이진형
		위험설비_식품가공용기계	이진형
		위험설비_지게차	이진형
		위험설비_컨베이어	이진형
		위험설비_크레인_천장_갠트리	이진형
		위험설비_타워크레인	이진형
		위험설비_프레스	이진형
		위험설비_혼합기	이진형
		점검차수	범주형
3	고용보험 근로자정보	피보험자_합계	수치형
		만나이_평균	실수형
		근속기간_년수_평균	실수형
		성별_남_합계	수치형
		성별_여_합계	수치형
		연령대_10대_20대_합계	수치형
		연령대_30대_합계	수치형
		연령대_40대_합계	수치형
		연령대_50대_합계	수치형
		연령대_60대이상_합계	수치형
		근속기간범주_1년이하_합계	상수형
		근속기간범주_1년초과3년이하_합계	상수형
		근속기간범주_3년초과5년이하_합계	상수형
		근속기간범주_5년초과10년이하_합계	상수형
		근속기간범주_10년초과20년이하_합계	상수형
		근속기간범주_20년초과_합계	상수형
		직종_대분류_건설·채굴직_합계	수치형
		직종_대분류_교육·법률·사회복지·경찰·소방직및 군인_합계	수치형
		직종_대분류_농림어업직_합계	수치형
		직종_대분류_미용·여행·숙박·음식·경비·청소직_ 합계	수치형
		직종_대분류_보건·의료직_합계	수치형
		직종_대분류_설치·정비·생산직_합계	수치형
		직종_대분류_연구직및공학기술직_합계	수치형
		직종_대분류_영업·판매·운전·운송직_합계	수치형
		직종_대분류_예술·디자인·방송·스포츠직_합계	수치형
4	안전보건관계자	전담유무	수치형
		건설안전관리자	수치형

번호	사업명	특성명	데이터 타입
		명예산업안전감독관	수치형
		보건관리자	수치형
		사업장담당자	상수형
		산업보건의	수치형
		안전관리자	수치형
		안전보건관리책임자	수치형
		안전보건총괄책임자	이진형
		관리자	수치형
		보건담당자	수치형
		선임자총수	수치형
		선임자종류수	수치형
5	재정지원 (안전투자혁신사업)	지원금액	수치형
		사업구분_고소작업대	이진형
		사업구분_고위험 TOP3 업종	상수형
		사업구분_노후 위험기계기구(30년이상)	이진형
		사업구분_리프트	이진형
		사업구분_뿌리공정	상수형
		사업구분_이동식크레인	이진형
		안투지원여부	이진형
		사업수행여부	이진형
	재정지원 (융자지원)	대하금액(천원)	수치형
		융자지원여부	이진형
	재정지원 (클린사업장)	교부금액	수치형
		클린지원여부	이진형
6	유해위험기계기구 (안전검사)	컨베이어(구간내컨베이어종류)_밸트	수치형
		컨베이어(구간내컨베이어종류)_체인	수치형
		컨베이어(구간내컨베이어종류)_롤러	수치형
		컨베이어(구간내컨베이어종류)_트롤리	수치형
		컨베이어(구간내컨베이어종류)_버킷	수치형
		컨베이어(구간내컨베이어종류)_나사	수치형
		검사대상품_대상품_대_고소작업대	수치형
		검사대상품_대상품_대_곤돌라	수치형
		검사대상품_대상품_대_국소배기장치	수치형
		검사대상품_대상품_대_롤러기	수치형
		검사대상품_대상품_대_리프트	수치형
		검사대상품_대상품_대_사출성형기	수치형
		검사대상품_대상품_대_산업용로봇	수치형
		검사대상품_대상품_대_압력용기	수치형
		검사대상품_대상품_대_원심기	이진형
		검사대상품_대상품_대_전단기	수치형
		검사대상품_대상품_대_컨베이어	수치형
		검사대상품_대상품_대_크레인	수치형
		검사대상품_대상품_대_프레스	수치형
		심사결과_심사결과_반려	수치형

번호	사업명	특성명	데이터 타입
		심사결과_심사결과_부적합	수치형
		심사결과_심사결과_적합	수치형
		심사결과_심사결과_진행	상수형
		자진신고여부_유	이진형
		사업수행여부	이진형
	유해위험기계기구 (자율안전검사)	인증대상품_대상품_대_곤돌라	수치형
		인증대상품_대상품_대_국소배기장치	이진형
		인증대상품_대상품_대_롤러기	상수형
		인증대상품_대상품_대_리프트	상수형
		인증대상품_대상품_대_사출성형기	상수형
		인증대상품_대상품_대_산업용로봇	상수형
		인증대상품_대상품_대_압력용기	수치형
		인증대상품_대상품_대_원심기	상수형
		인증대상품_대상품_대_전단기	상수형
		인증대상품_대상품_대_컨베이어	수치형
		인증대상품_대상품_대_크레인	수치형
		인증대상품_대상품_대_프레스	상수형
		사업수행여부	이진형
7	소방청 위험물 제조소 등 정보	예방규정제출대상여부	이진형
		대량위험물제조소등여부	이진형
		석유화학단지내사업장여부	이진형
		소화난이도등급_1등급	수치형
		소화난이도등급_2등급	수치형
		소화난이도등급_3등급	수치형
		설치기간_10년이상	수치형
		설치기간_20년이상	수치형
		설치기간_5년미만	수치형
		설치기간_5년이상	수치형
		위험물제조소_총합	수치형
		위험물제조소_종류	수치형
		위험물허가_유별_제1류	수치형
		위험물허가_유별_제2류	수치형
		위험물허가_유별_제3류	수치형
		위험물허가_유별_제4류	수치형
		위험물허가_유별_제5류	수치형
		위험물허가_유별_제6류	수치형
		위험물허가_유별_합계	수치형
		위험물허가_유별_여부	이진형
		탱크총합	수치형
		탱크여부	이진형
		이송취급소수	수치형
		이송취급소여부	이진형
		이동탱크수	수치형
		이동탱크여부	이진형

번호	사업명	특성명	데이터 타입
8	지게차 실태조사	지게차대수_자가	수치형
		지게차대수_그외	수치형
	지게차 실태조사 (안전관리체계화 수행 결과)	지게차보유대수	수치형
		지게차용량	실수형
		충돌방지장치점수	실수형
		안전띠점수	실수형
		법정방호장치점수	실수형
		운전자격점수	실수형
		관리자이해점수	실수형
		운전자관찰점수	실수형
		점수총합	실수형
9	KOSHA_MS_18001	사업대상여부	이진형
10	위험성평가 (컨설팅)	사업주의관심도	실수형
		위험성평가실행수준	실수형
		구성원의참여및이해수준	실수형
	위험성평가 (인정/불인정)	사업주의 관심도	실수형
		위험성평가 실행수준	실수형
		구성원의 참여 및 이해수준	실수형
		재해발생수준	수치형
11	민간위탁기술지도 (안전)	선정기준코드_물질관련	이진형
		선정기준명_기타	이진형
		선정기준명_신규사업장	이진형
		선정기준명_유해물질존재사업장	이진형
		선정기준명_재해발생사업장	이진형
		선정기준명_특별대책사업장	이진형
		지원횟수	수치형
		전담	이진형
		겸직	이진형
		등급_중급	이진형
		등급_초급	이진형
		교육인원합	수치형
		자료제공_총합	수치형
		검사실시_종류수	수치형
		검사미실시_종류수	수치형
		검사비대상_종류수	수치형
		검사실시_합	수치형
		검사미실시_합	수치형
		검사비대상_합	수치형
		전체_개선	수치형
		전체_조치의뢰	수치형
		위탁기관평가	범주형
	민간위탁기술지도 (보건)	처리사유_개선완료	상수형
		처리사유_조치의뢰	이진형
		선정기준코드_산재관련	이진형

번호	사업명	특성명	데이터 타입
		선정기준코드_실비관련	이진형
		선정기준코드_물질관련	이진형
		선정기준명_기타	이진형
		선정기준명_신규사업장	이진형
		선정기준명_유해물질존재사업장	이진형
		선정기준명_재해발생사업장	이진형
		선정기준명_특별대책사업장	이진형
		지원횟수	수치형
		전담	이진형
		겸직	이진형
		등급_중급	이진형
		등급_초급	이진형
		교육인원_총수	수치형
		자료제공_총합	수치형
		밀폐실태_밀폐공간장소_통의내부등	이진형
		밀폐실태_밀폐공간장소_정화조등	이진형
		밀폐실태_밀폐공간장소_기타밀폐공간	이진형
		밀폐실태_밀폐공간장소_반응기등내부	이진형
		밀폐실태_밀폐공간장소_강재등시설	이진형
		밀폐실태_밀폐공간장소_불활성기체설비	이진형
		밀폐실태-질식사고 인지도	수치형
		밀폐실태-질식사고 위험관리 인지도	수치형
		밀폐실태-질식사고 교육이수	수치형
		밀폐실태-가스농도측정기 보유	수치형
		급기팬 보유	수치형
		밀폐실태-위험도평가 총점	수치형
		개선_합	수치형
		최종위험성평가수준평가결과	수치형
		실태평가결과_사업주의지	수치형
	민간위탁기술지도 (화학)	처리사유내용_종결	이진형
		처리사유내용_조치의뢰	이진형
		선정기준코드_산재관련	이진형
		선정기준코드_실비관련	이진형
		선정기준코드_물질관련	이진형
		선정기준명_기타	이진형
		선정기준명_신규사업장	이진형
		선정기준명_유해물질존재사업장	이진형
		선정기준명_재해발생사업장	이진형
		선정기준명_특별대책사업장	이진형
		지원횟수	수치형
		전담	이진형
		겸직	이진형
		등급_중급	이진형
		등급_초급	이진형

번호	사업명	특성명	데이터 타입
		교육인원_총수	수치형
		자료제공_총합	수치형
		기술지원_사고성재해예방	이진형
		기술지원_화학사고예방	이진형
		유해위험물질수	수치형
		화학설비건수	수치형
		위험기계기구_보유건수	수치형
		개선_총계	수치형
		조치의뢰_총계	이진형
		국소배기장치	수치형
		롤러기	상수형
		리프트	수치형
		분쇄파쇄기	이진형
		사출성형기	이진형
		산업용로봇	이진형
		식품제조용설비	이진형
		원심기	상수형
		전단기	이진형
		지게차	수치형
		컨베이어	이진형
		크레인	수치형
		프레스	상수형
		혼합기	이진형
		압력용기	이진형
		위험기계기구총합	수치형
		밀폐_정화조등	수치형
		밀폐_통의내부등	상수형
		밀폐_불활성기체설비	상수형
		밀폐_강재등시설	수치형
		밀폐_반응기등내부	수치형
		밀폐_콘크리트양생	상수형
		밀폐_중독위험장소	이진형
		밀폐_기타밀폐공간	이진형
		위탁기관평가	범주형
		사업수행여부	이진형
	민간위탁기술지도 (서비스)	처리사유_개선완료	상수형
		처리사유_조치의뢰	이진형
		선정기준코드_산재관련	이진형
		선정기준코드_설비관련	이진형
		선정기준코드_물질관련	이진형
		선정기준명_기타	이진형
		선정기준명_신규사업장	이진형
		선정기준명_유해물질존재사업장	이진형
		선정기준명_재해발생사업장	이진형

번호	사업명	특성명	데이터 타입
		선정기준명_특별대책	이진형
		전담	이진형
		겸직	이진형
		등급_초급	이진형
		등급_중급	이진형
		교육인원_총수	수치형
		자료제공_총합	수치형
		지적건수_총합	수치형
		개선건수_총합	수치형
		조치건수_총합	수치형
		최종위험성평가수준평가결과	범주형
	민간위탁기술지도 (사고사망예방 서비스)	지게차사용수량(대)	수치형
		지게차조종면허자격보유(명)	수치형
		지게차운행속도제한표지판설치여부	이진형
		지게차작업안전수칙제정및게시여부	이진형
		지게차사람공동사용출입구(개소)	수치형
		지게차운행경사로수(개소)	수치형
		지게차모니터링CCTV(대)	수치형
		컨베이어설치수량(대)	수치형
		전용상하역장확보여부	수치형
		상하역공간조명등설치여부	수치형
		이동식사다리보유수량(대)	수치형
		사다리사용시안전모착용여부	수치형
		일평균차량출입(대)	이진형
		챠량운행모니터링CCTV(대)	수치형
		연평균사다리사용일수(일)	수치형
		사업수행여부	이진형
12	공단 교육 (안전보건교육)	교육분야코드_관리자	실수형
		교육분야코드_안전보건관계자	실수형
		교육분야코드_일반근로자	실수형
		교육분야코드_취약계층	실수형
	공단 교육 (인터넷교육센터)	교육대상_재분류_근로자	실수형
		교육대상_재분류_책임자	실수형
		교육대상_재분류_특수형태근로자	실수형
		수료여부_미수료	실수형
		수료비율	실수형
	공단 교육 (직무교육센터)	과정구분_new_온라인	실수형
		과정구분_new_집체	실수형
		교육대상_전문기관종사자	실수형
		수료여부_미수료	실수형
		안전보건관계자	실수형
13	유해위험방지계획서	대상업종_합계	수치형
		대상규모명_over_2000	수치형
		대상규모명_under_2000	수치형

번호	사업명	특성명	데이터 타입
		대상규모명_under_500	수치형
		사업구분_변경	수치형
		사업구분_설치	수치형
		사업구분_이전	수치형
		심사결과_반려	수치형
		심사결과_부적정	이진형
		심사결과_적정	수치형
		심사결과_조건부적정	수치형
		고용부조치통보_미제출	이진형
		고용부조치통보_지연	이진형
		최종확인회차	수치형
14	작업환경측정 (측정)	물질군명_노출기준제정물질	수치형
		물질군명_허가대상 유해물질	상수형
		지상대상구분_대상	수치형
		초과율	실수형
		취급인원	수치형
		물질군명_물리적인자_합	수치형
		물질군명_화학적인자_합	수치형
		물질군명_분진_합	수치형
	작업환경측정 (화학물질취급현황)	취급구분_사용	수치형
		취급구분_제조	수치형
		취급용도_기타	수치형
		취급용도_세척	이진형
		취급용도_실험	수치형
		취급용도_용접	수치형
		취급용도_원료	수치형
		취급물질군명_기타유해물질	수치형
		취급물질군명_노출기준제정물질	수치형
		취급물질군명_제조금지 유해물질	상수형
		취급물질군명_허가대상 유해물질	수치형
		취급물질군명_물리적인자_통합	상수형
		취급물질군명_분진인자_통합	수치형
		취급물질군명_화학적인자_통합	수치형
15	고용보험ERP근로자수	고용상시인원수	수치형
		남성근로자수	수치형
		여성근로자수	수치형
		외국인근로자수	수치형
16	특수건강진단 (특검)	총수진자수(명)	수치형
		유해물질군명_노출기준제정물질	수치형
		유해물질군명_야간작업	수치형
		유해물질군명_제조금지 유해물질	상수형
		유해물질군명_허가대상 유해물질	수치형
		A판정비율	실수형
		C1판정비율	실수형

번호	사업명	특성명	데이터 타입
		C2판정비율	실수형
		D1판정비율	실수형
		D2판정비율	실수형
		CN판정비율	실수형
		DN판정비율	실수형
		유해물질군명_물리적인자_합	수치형
		유해물질군명_화학적인자_합	수치형
		유해물질군명_분진_합	수치형
		사업수행여부	이진형
	특수건강진단 (사업장별 검진내역)	검진종목별 수진자수(일반)	수치형
		검진종목별 수진자수(특수)	수치형
		검진종목별 수진자수(배치전)	수치형
		검진종목별 수진자수(수시)	이진형
		검진종목별 수진자수(임시)	이진형
		검진종목별 수진자수(수첩)	이진형
		사업수행여부	이진형
17	산업안전보건실태조사	종사자수	수치형
		교대 근무제 시행	이진형
		노동조함 유무(예, 아니오)	이진형
		현장 산업안전보건위원회 구성 및 운영	이진형
		1년간 유해, 위험 요인에 대한 위험성 평가 및 필요한 조치 문서작업	이진형
		2020년 일반건강진단 결과 사후관리조치대상자 조치 이행 여부 확인	상수형
		2020년 특수건강진단 결과 사후관리조치대상자 조치 이행 여부 확인	상수형
		2020년 작업 환경 측정 결과를 바탕으로 유해인자의 노출량을 최소화하기 위한 구체적 노력	이진형
		사업체 내에 상주하며 연간 계약을 하는 협력 업체	이진형
		사내 상주하며 연간 계약 협력 업체 수	상수형
		사내 상주하며 연간 계약 협력 업체의 총 근로자 수	상수형
		거래하는 원청 회사 여부, 사업체가 원청 회사의 사업체 내에 위치하는지 여부	이진형
		교대근무_종사자_비율	실수형
		야간근무_종사자_비율	실수형
		작업 환경 관련 위험 요인 평균	실수형
		신체적 부담 관련 위험 요인 평균	실수형
		생/화학 물질 관련 위험 요인 평균	실수형
		기계, 전기, 기타 위험 요인 평균	실수형
		스트레스 심각도 평균	실수형
		스트레스 관리 노력 정도 평균	실수형
		경영진 안전보건의지 평균	실수형
		사업장내 안전문화 평균	실수형
		근로자 안전보건의지 평균	실수형

번호	사업명	특성명	데이터 타입
18	산업재해조사표	재해자동종경력년수_1~3	수치형
		재해자동종경력년수_3~5	수치형
		재해자동종경력년수_5~10	수치형
		재해자동종경력년수_10~20	수치형
		재해자동종경력년수_20~	수치형
		상해부위_머리	수치형
		상해부위_몸통	수치형
		상해부위_팔	수치형
		상해부위_전신	수치형
		상해부위_다발성	수치형
		상해부위_기타	수치형
		재해발생요일_주말	수치형
19	작업환경실태조사 (일반현황)	전기계약용량	수치형
		야간작업유무	이진형
		정비_보수여부	이진형
		안전관리자	수치형
		안전관리자_유형	이진형
		보건관리자	수치형
		보건관리자_유형	이진형
		안전보건담당자	수치형
		안전보건담당자수	수치형
		원청_하청여부	이진형
		하청사업장수	수치형
		하청근로자수	수치형
		근골격계부담작업대상여부	이진형
		유해요인조사실시여부	수치형
		복지시설_개수	수치형
	작업환경실태조사 (화학물질취급)	취급/생산	상수형
		허용대상물질여부	상수형
		허용기준물질여부	상수형
		관리대상물질여부	이진형
		안전검사물질여부	이진형
		안전관리물질여부	이진형
		기타물질여부	이진형
		특검대상물질여부	이진형
		측정대상물질여부	이진형
		PSM대상물질여부	이진형
		건강관리수첩대상물질여부	상수형
		사고대상물질여부	이진형
		금지대상물질여부	상수형
		근로자_월_취급시간	실수형
	작업환경실태조사 (기계기구설비현황)	제조_보유수량총개수	상수형
		제조_총종류수량	상수형
		비제조_보유수량총개수	수치형

번호	사업명	특성명	데이터 타입
		비제조_총종류수량	수치형
	작업환경실태조사 (작업환경)	소음발생공정수_총합	수치형
		밀폐공간수_총합	수치형
		작업환경구분_고열_한랭_다습_및_방사선_취급_작업_총합	수치형
		작업환경구분_밀폐공간(산소결핍_위험장소)_현황_총합	수치형
		작업환경구분_분진_흄_발생작업_총합	수치형
		작업환경구분_사내도급작업_총합	상수형
		작업환경구분_소음작업_총합	수치형
		작업환경구분_제조나노물질의_제조_및_취급_작업_총합	상수형
		작업환경구분_진동발생작업_총합	수치형
		고열_한랭_다습_및_방사선_취급_작업_종사근로자수합	수치형
		밀폐공간(산소결핍_위험장소)_현황_종사근로자수합	상수형
		분진_흄_발생작업_종사근로자수합	수치형
		사내도급작업_종사근로자수합	상수형
		소음작업_종사근로자수합	수치형
		제조나노물질의_제조_및_취급_작업_종사근로자수합	상수형
		진동발생작업_종사근로자수합	수치형
21	공공기관등급 데이터	등급	수치형
23	사업장수준조사평가 +재해율	사업주 · 관리자 마인드	실수형
		근로자 안전보건 행동 수준	실수형
		작업장 및 근로환경 수준	실수형
		3년 평균 재해율	실수형

2) 산업안전 데이터 정제 및 전처리

○ 산업안전 데이터의 전처리는 수치형, 범주형, 상수형, 이진형으로 분류된 특성을 대상으로 진행하였으며, 상수형과 텍스트 데이터는 전처리 및 모델 학습에서 제외하였음.

○ 텍스트 데이터인 '사업장관리번호'와 '사업장개시번호'를 제외한 제조업 분야의 417개 특성과 서비스업 분야의 471개 특성을 해당 데이터 타입별로 분류하고, 각 데이터 타입에 따라 다음과 같은 순서로 전처리를 수행하였음.

1. 결측치 처리: 데이터에서 누락된 값을 식별하고, 적절한 대체 방법이나 제거를 통해 결측치를 처리함.
2. 이상치 처리: 통계적 방법을 중심으로 이상치를 탐지하고 수정하거나 제거하였음.
3. 코드값 변환: 범주형 변수의 문자 데이터를 모델이 처리할 수 있는 수치형 또는 원-핫 인코딩 형태로 변환하였음.
4. 데이터 범주화: 연속형 변수를 분석 목적에 맞게 구간화하여 범주형 변수로 변환하였음.
5. 특성 분포 불균형 처리: 특정 값이나 범주에 데이터가 편중된 경우, 로그 변환 기법 등을 활용하여 분포의 불균형을 완화함.
6. 특성 스케일링: 변수들 간의 규모 차이를 조정하기 위해 정규화나 표준화 등의 스케일링 기법을 적용함.
7. 데이터 분할: 모델의 일반화 성능을 평가하기 위해 데이터를 학습용, 검증용, 테스트용으로 분할함.
8. 데이터 불균형 처리: 타깃 변수의 클래스 불균형을 해결하기 위해, 오버샘플링, 언더샘플링 등의 기법을 적용함.

○ 이러한 전처리 과정을 통해 사업장 데이터의 품질을 향상시키고, 모델 학습에 최적화된 형태로 데이터를 구성하였음. 〈표 3-4〉는 데이터 전처리 대상 데이터임.

〈표 3-4〉 데이터 전처리 대상 특성 데이터

번호	특성코드	특성명	데이터 타입
1	m_02_1_c010	경영자마인드	이진형
2	m_02_1_c011	안전보건관리및개선노력	실수형
3	m_02_2_c010	안전관리수준평가사업장위험도_현장위험관리수준	실수형
4	m_03_c004	만나이_평균	실수형

번호	특성코드	특성명	데이터 타입
5	m_09_c007	충돌방지장치점수	실수형
6	m_09_c008	안전띠점수	실수형
7	m_09_c009	법정방호장치점수	실수형
8	m_09_c010	운전자격점수	실수형
9	m_09_c011	관리자이해점수	실수형
10	m_09_c012	운전자관찰점수	실수형
11	m_09_c013	점수총합	실수형
12	m_16_c003	사업주의 관심도	실수형
13	m_16_c004	위험성평가 실행수준	실수형
14	m_16_c005	구성원의 참여 및 이해수준	실수형
15	m_13_c003	사업주의관심도	실수형
16	m_13_c004	위험성평가실행수준	실수형
17	m_13_c005	구성원의참여및이해수준	실수형
18	m_13_c006	재해발생수준	실수형
19	m_27_c004	근로자안전보건행동수준	실수형
20	m_01_c005	일선기관	범주형
21	m_01_c006	노동지청	범주형
22	m_01_c008	중업종명	범주형
23	m_01_c009	소업종명	범주형
24	m_01_c010	표준산업분류	범주형
25	m_01_c014	행정구역	범주형
26	m_01_c015	행정구역_세부	범주형
27	m_02_2_c011	점검차수	범주형
28	m_02_3_c015	점검차수	범주형
29	m_16_c006	심사결과_불인정	범주형
30	m_14_2_c013	지원횟수	범주형
31	m_14_4_c026	개선_총계	수치형
32	m_23_3_c006	기계설비_비제조_보유종류	범주형
33	m_25_c007	위탁기관평가_점수	범주형
34	m_27_c003	사업주관리자마인드	범주형
35	m_27_c005	작업장및근로환경수준	범주형
36	m_14_2_c020	자료제공_총합	수치형
37	m_14_2_c035	개선필요_합	수치형
38	m_14_2_c036	개선_합	수치형

(1) 이상치 처리(Outlier Handling)

○ 수치형과 실수형 데이터를 중심으로 이상치 처리 기법을 적용하였음. 각 특성별로 극단적인 값들을 탐지하여 하나의 특성 이상에서 이상치 발생 시, 학습 데이터셋에서 제외하였음.

○ 이진형 데이터의 경우, 가능한 값은 0과 1로 제한되므로, 이 외의 값이 존재하는 것은 데이터 오류로 간주하였으며, 범주형 데이터에서는 매우 낮은 빈도로 발생하는 범주를 확인하였음. 이러한 희귀한 범주는 모델 학습에 유의미한 정보를 제공하지 못하고, 오히려 노이즈를 발생시킬 우려가 있음.

○ 〈표 3-5〉는 이상치 처리 결과로 이상치 처리는 결과적으로 처리 기법별로 이상치 제거를 위해 과도하게 샘플 수를 제거할 경우, 모델 성능에 다소 영향을 미치는 것을 확인하였음.

〈표 3-5〉 이상치 처리 결과, 제거된 데이터

사업장 연번	경영자 마인드	만나이 평균	충돌방지 장치점수	안전띠 점수	법정방호 장치점수	…	사업주의 관심도
307162	3	0	1	2	3	…	2
13447	5	46.666	3	2	3	…	2
169901	2	51.111	3	3	3	…	2.5
29055	3	0	2	4	4	…	0
35125	3	0	1	2	2	…	3
271875	1	0	2	1	3	…	0
15420	1	54	1	3	3	…	0
346875	3	0	1	2	2	…	0
151183	2.5	51.166	2	2.5	2.5	…	1
256653	1	57.2	2	3	3	…	0
93940	3	51	1	3	2	…	4
170123	3	58.666	3	2	3	…	0
195940	2	45.2	3	3	2	…	2
(생략)							
200631	2	50.25	1	3	3	…	0
345348	3.5	38.111	1	1	2	…	1

사업장 연번	경영자 마인드	만나이 평균	충돌방지 장치점수	안전띠 점수	법정방호 장치점수	…	사업주의 관심도
130420	4	43.4	3	3	3	…	1
300052	5	46.8	2	3	3	…	3
352747	3	27.8	2	3	2	…	3
206105	2	0	1	2	3	…	0
308708	5	42.666	2	2	2	…	0

(2) 데이터 불균형 처리(Imbalanced Data Handling)

○ 사업장 데이터를 확인한 결과, 동일 업종 및 규모의 사업장 내에서 사고가 발생하지 않은 안전 사업장이 사고가 발생한 위험 사업장보다 훨씬 높은 비율을 차지하고 있음을 확인함. 학습데이터셋이 안전 사업장에 편향되어 있어, 모델 학습 시 불균형한 데이터 분포가 영향을 미칠 수 있음.

○ 이를 해결하기 위해 위험 사업장의 데이터를 오버샘플링하여 학습 데이터의 균형을 맞추고자 하였음. 반면, 언더샘플링을 적용할 경우 안전 사업장의 데이터 수가 현저히 줄어들어 모델의 성능이 저하되는 문제가 발생하여 이를 피함.

○ SMOTE-Tomek 기법을 활용하여 부족한 위험 사업장은 오버샘플링하고, 안전 사업장과 위험 사업장 경계에 있는 학습에 불필요한 데이터는 언더샘플링으로 제거하였음. 그러나 언더샘플링으로 인해 제거된 데이터 수가 매우 적어, 학습 데이터의 전체적인 정보량에는 큰 영향을 미치지 않았음.

○ 〈표 3-6〉는 분할 처리 된 데이터 중, 학습 데이터를 대상으로 증강하여 생성된 위험사업장 데이터임.

〈표 3-6〉 데이터 불균형 처리 결과, 증강된 학습데이터

증강 연번	경영자 마인드	만나이 평균	충돌방지 장치점수	안전띠 점수	법정방호 장치점수	…	사업주의 관심도
1	-0.30462	2.126907	-1.01458	0.827756	1.485781	…	1.199699
2	-0.345	0.462947	2.330972	0.752003	1.587558	…	-1.18854
3	-0.21539	0.571071	0.142392	0.658723	0.145087	…	1.346496
4	-0.20216	0.877915	-1.04481	-1.6657	0.373536	…	-0.48524
5	1.512094	0.714386	1.255335	0.686623	0.196118	…	-1.1155
6	-1.338	-1.30471	-1.07056	-1.33097	0.194783	…	0.162007
7	-0.32335	-1.42729	0.167932	-0.3547	-1.18557	…	1.121965
8	-0.296	0.515568	0.130298	0.671303	0.216059	…	1.553874
9	-0.28512	0.388579	-1.18528	-1.33649	-1.14814	…	0.857766
10	-0.16759	0.724026	-1.019	-1.38197	-0.00778	…	-0.48713
11	-1.26461	1.659099	-0.98306	-1.45799	-1.34121	…	1.849555
12	-0.2408	0.33238	0.117455	1.974675	0.07051	…	-1.10368
13	-0.18475	0.441337	-1.33125	-1.42161	-0.96815	…	0.751501
14	0.648049	0.721607	0.002278	-1.53749	0.014584	…	0.161555
15	-0.19396	1.664513	-1.01687	0.723163	1.422639	…	1.135029
16	-1.28749	1.105159	1.193727	-0.26894	0.157925	…	-1.10374
17	-1.18072	1.637439	-1.24121	-1.53939	-1.07674	…	1.767526
18	-0.3169	-1.34624	-1.12932	-0.40055	0.004398	…	0.15229
19	-0.29613	0.375368	0.119597	0.784846	0.061259	…	1.302648
20	-1.24352	1.697322	-0.83103	-1.57156	-1.2114	…	1.609408
(생략)							

(3) 코드값 변환(Categorical Encoding)

○ 사업장 정보에서 ‘행정구역’, ‘표준산업분류’, ‘업종’과 같은 정보는 중요한 특성으로 사용됨. 이러한 정보들은 일반적으로 명칭과 함께 범주형 코드값으로 관리되고 있음.

○ 그러나 머신러닝 모델이 이러한 숫자 코드로 표현된 범주형 데이터를 입력받으면, 값들 간의 존재하지 않는 순서나 크기 관계가 있다고 잘못 해석할 수 있으므로 인코딩 기법을 적용하였음. 〈표 3-7〉는 사업장 정보

중 '행정구역', '표준사업분류', '업종' 등을 코드값 변환 처리한 결과임.

〈표 3-7〉 코드값 변환(타겟 인코딩 기법) 처리 결과

사업장 연번	경영자 마인드	…	알선기관	소업종명	표준산업 분류	행정구역	…	사업주의 관심도
1	2	…	0.2	0.35	0.2	0.433333	…	3
2	5	…	0.5	0.35	0.465517	0.433333	…	0
3	3	…	0.75	0.35	0.465517	0.666667	…	1
4	5	…	0.5	0.35	0.465517	0.433333	…	5
5	3	…	0.4	0.35	0.465517	0.444444	…	3.5
6	4	…	0.6	0.35	0.465517	0.6	…	2
7	3	…	0.25	0.35	0.465517	0.166667	…	3
8	2	…	0.2	0.35	0.133333	0	…	2
9	3	…	0.2	0.35	0.2	0.433333	…	4
10	5	…	0	0.35	0.133333	0.5	…	3
11	5	…	0.4	0.35	0.465517	0.444444	…	0
12	3	…	0.25	0.35	0.465517	0	…	3.5
13	3	…	1	0.35	0.465517	0.433333	…	2
14	2	…	0.125	0.35	0.465517	0.111111	…	0
15	2	…	0.125	0.35	0.2	0.111111	…	4.5
16	4	…	0.25	0.35	0.133333	0	…	1
17	3	…	0.333333	0.35	0.465517	0	…	2
18	3	…	0.5	0.35	0.465517	0.433333	…	0
19	4	…	0.5	0.35	0.465517	0.444444	…	0
20	3	…	0.25	0.35	0.2	0.166667	…	0
(생략)								

(4) 결측치 처리(Missing Data Handling)

○ 안전보건공단에서 제공받은 초기 데이터는 이미 기본적인 처리가 완료되어 결측치가 존재하지 않았음. 그러나 기존의 사업장 데이터 외에 사업장 위험 수준 현장평가 데이터를 추가로 제공받아 새로운 특성으로 반영하는 과정에서 결측치 처리가 필요하게 되었음. 이 현장평가 데이터는 5점 척도의 3개 문항으로 구성되어 있으며, 제조업 및 서비스업 데이터와 데이터 규모 차이가 존재함.

○ 수치형 데이터의 특성을 고려하여, 결측치를 효율적으로 처리하기 위해 KNN Imputation, MICE와 같은 기법을 활용하여 결측치를 채움.

○ 결과적으로, 데이터 결합하여 결측치를 처리하였으나, 사업장 위험 수준 현장평가 데이터의 규모가 적어, 두 개의 데이터셋 간의 규모 차이로 인해 결측치가 크게 발생하였음.

○ 결측치 처리는 가끔 발생하는 소수의 데이터를 대상으로 하므로, 크게 영향을 이 경우는 크게 학습에 영향을 미치지 못하였음.

○ 〈표 3-8〉는 사업장 데이터와 사업장 위험 수준 현장평가 데이터를 결합한 후 발생한 결측치를 처리한 결과임.

〈표 3-8〉 사업장 위험 수준 현장평가 결합 후, 결측치 처리 결과

사업장 연번	경영자 마인드	만나이 평균	…	근로자 안전 보건 행동수준	작업장 및 근로환경	사업주·관리자 안전의식	사업주의 관심도
1	2	42	…	2.802309	2.819073	2.681659	3
2	5	49.66667	…	2.802309	2.819073	2.681659	0
3	3	51	…	2.802309	2.819073	2.681659	1
4	5	45.14286	…	2.802309	2.819073	2.681659	5
5	3	59.42857	…	2.802309	2.819073	2.681659	3.5
6	4	48.46154	…	2.802309	2.819073	2.681659	2
7	3	41.42857	…	2.802309	2.819073	2.681659	3
8	2	0	…	2.802309	2.819073	2.681659	2
9	3	44.33333	…	2.802309	2.819073	2.681659	4
10	5	0	…	2.802309	2.819073	2.681659	3
11	5	48.18182	…	2.802309	2.819073	2.681659	0
12	3	0	…	2.802309	2.819073	2.681659	3.5
13	3	55	…	2.802309	2.819073	2.681659	2
14	2	57.33333	…	2.802309	2.819073	2.681659	0
15	2	71	…	2.802309	2.819073	2.681659	4.5
16	4	46	…	2.802309	2.819073	2.681659	1
17	3	0	…	2.802309	2.819073	2.681659	2
18	3	44.46667	…	2.802309	2.819073	2.681659	0

사업장 연번	경영자 마인드	만나이 평균	…	근로자 안전 보건 행동수준	작업장 및 근로환경	사업주·관리자 안전의식	사업주의 관심도
19	4	46.78125	…	2.802309	2.819073	2.681659	0
20	3	39.14286	…	2.802309	2.819073	2.681659	0
(생략)							

(5) 특성 분포 불균형 처리(Feature Distribution Balancing)

○ 사업장 정보로부터 다양한 특성들을 생성하였으나, 생성된 특성에서 과도한 분포 불균형이 확인되었음. 특정 값에 데이터가 지나치게 집중되면 모델이 이를 편향되게 학습하여 과적합될 위험이 있으므로, 전체적인 값의 균형을 맞추는 작업을 수행함.

○ 특성 분포를 최대한 정규분포에 가깝도록 조정하고 특성으로 인한 과적합을 방지하고 모델 일반화 성능을 높일 수 있도록 처리함.

〈표 3-9〉 특성분포 불균형 처리 결과

사업장 연번	경영자 마인드	만나이 평균	충돌방지 장치점수	안전띠 점수	법정방호 장치점수	…	사업주의 관심도
1	-1.19772	0.536478	0.274978	-0.34311	-1.17011	…	0.855
2	1.709794	0.749951	1.22131	0.716875	0.153174	…	-1.25163
3	-0.32022	0.785549	-1.22069	-1.50127	0.153174	…	-0.30481
4	1.709794	0.625908	0.274978	1.709364	1.543434	…	1.650905
5	-0.32022	1.001688	-1.22069	0.716875	1.543434	…	1.076549
6	0.652928	0.717412	0.274978	-1.50127	0.153174	…	0.344927
7	-0.32022	0.519906	-1.22069	-1.50127	-1.17011	…	0.855
8	-1.19772	-1.43931	-1.22069	-1.50127	0.153174	…	0.344927
9	-0.32022	0.603143	0.274978	0.716875	0.153174	…	1.281309
10	1.709794	-1.43931	1.22131	0.716875	-1.17011	…	0.855
11	1.709794	0.709808	1.22131	0.716875	0.153174	…	-1.25163
12	-0.32022	-1.43931	0.274978	-0.34311	-1.17011	…	1.076549
13	-0.32022	0.889948	-1.22069	-1.50127	-1.17011	…	0.344927

사업장 연번	경영자 마인드	만나이 평균	충돌방지 장치점수	안전띠 점수	법정방호 장치점수	…	사업주의 관심도
14	-1.19772	0.949298	1.22131	-0.34311	0.153174	…	-1.25163
15	-1.19772	1.27739	-1.22069	-1.50127	-1.17011	…	1.472057
16	0.652928	0.649818	-1.22069	0.716875	0.153174	…	-0.30481
17	-0.32022	-1.43931	-1.22069	-0.34311	0.153174	…	0.344927
18	-0.32022	0.606905	1.896001	0.716875	1.543434	…	-1.25163
19	0.652928	0.67144	0.274978	0.716875	0.153174	…	-1.25163
20	-0.32022	0.452587	0.274978	1.709364	0.153174	…	-1.25163
(생략)							

(6) 데이터 범주화 처리(Data Binning)

○ 트리 기반 모델에서는 데이터가 범주화되어 있을 때 더 직관적이고 빠른 분할이 가능하여 학습 효율이 향상되며, 연속 데이터의 미세한 변화에 덜 민감해져, 과적합 문제를 완화할 수 있음.

○ 모든 실수형 특성이 범주화에 적합한 것은 아니므로, 도메인 지식을 충분히 고려해야 하며, 범주화를 통해 생성된 범주의 균형을 적절히 관리하지 않으면 모델의 복잡성이 증가함.

○ 또한 특성값에 대한 명확한 임계값이나 기준점, 순서 등에 대한 정보가 부족하여, 성능에 유의미한 영향이 부족하였으며, 향후 도메인 지식이 반영된 특성에 대한 적절한 기준을 설정하여 유의미한 범주를 두는 것이 필요함.

○ 〈표 3-10〉은 만나이 평균, 근로자수, 피보험자 합계, 근속기간 년수 평균 등의 특성에 대해 데이터 범주화 처리 결과임.

〈표 3-10〉 데이터 범주화 처리 결과

사업장 연번	경영자 마인드	…	만나이 평균	근로자수	피보험자 합계	근속기간 년수 평균	…	사업주의 관심도
1	2	…	40	0	0	0	…	3
2	5	…	40	5	5	3	…	0
3	3	…	50	0	0	5	…	1
4	5	…	40	10	10	1	…	5
5	3	…	50	5	5	8	…	3.5
6	4	…	40	15	10	5	…	2
7	3	…	40	5	5	0	…	3
8	2	…	0	0	0	0	…	2
9	3	…	40	10	5	1	…	4
10	5	…	0	0	0	0	…	3
11	5	…	40	10	10	0	…	0
12	3	…	0	0	0	0	…	3.5
13	3	…	50	0	1	3	…	2
14	2	…	50	0	5	1	…	0
15	2	…	70	0	0	0	…	4.5
16	4	…	40	0	0	1	…	1
17	3	…	0	0	0	0	…	2
18	3	…	40	15	30	10	…	0
19	4	…	40	25	30	1	…	0
20	3	…	30	10	10	5	…	0
(생략)								

(7) 특성 스케일링(Feature Scaling)

○ 사업장 특성 중 유무 여부, 합계, 점수 등 다양한 유형의 값들이 존재하고, 각 특성 간 값의 범위가 상이하여 머신러닝 모델이 특성의 중요도를 다르게 인식함. 이를 방지하기 위해 스케일링 처리를 수행하여 모든 특성이 균일한 범위 내에 있도록 조정하여 학습의 안정성을 높이도록 하고 모델이 데이터의 본질적인 패턴을 더욱 효과적으로 파악할 수 있게 함.

○ 〈표 3-11〉은 범주형, 이진형 특성을 제외한 전 특성에 대해 스케일링 처리 결과임.

〈표 3-11〉 특성 스케일링 처리 결과

사업장 연번	경영자 마인드	만나이 평균	충돌방지 장치점수	안전띠 점수	법정방호 장치점수	…	사업주의 관심도
1	-1.18454	0.428627	0.085641	-0.37769	-1.17274	…	0.791476
2	1.590046	0.76067	1.263301	0.701884	0.171047	…	-1.13661
3	-0.25968	0.818417	-1.09202	-1.45727	0.171047	…	-0.49391
4	1.590046	0.564744	0.085641	1.781463	1.514836	…	2.076864
5	-0.25968	1.83459	-1.09202	0.701884	1.514836	…	1.112823
6	0.665185	0.708476	0.085641	-1.45727	0.171047	…	0.148782
7	-0.25968	0.403878	-1.09202	-1.45727	-1.17274	…	0.791476
8	-1.18454	-1.3904	-1.09202	-1.45727	0.171047	…	0.148782
9	-0.25968	0.529684	0.085641	0.701884	0.171047	…	1.43417
10	1.590046	-1.3904	1.263301	0.701884	-1.17274	…	0.791476
11	1.590046	0.696362	1.263301	0.701884	0.171047	…	-1.13661
12	-0.25968	-1.3904	0.085641	-0.37769	-1.17274	…	1.112823
13	-0.25968	0.991657	-1.09202	-1.45727	-1.17274	…	0.148782
14	-1.18454	1.092714	1.263301	-0.37769	0.171047	…	-1.13661
15	-1.18454	1.684618	-1.09202	-1.45727	-1.17274	…	1.755517
16	0.665185	0.601867	-1.09202	0.701884	0.171047	…	-0.49391
17	-0.25968	-1.3904	-1.09202	-0.37769	0.171047	…	0.148782
18	-0.25968	0.535458	2.440962	0.701884	1.514836	…	-1.13661
19	0.665185	0.635703	0.085641	0.701884	0.171047	…	-1.13661
20	-0.25968	0.304884	0.085641	1.781463	0.171047	…	-1.13661
(생략)							

(8) 데이터 분할 처리(Data Splitting)

○ 모델이 학습하는 과정을 검증하고 학습된 모델의 성능을 평가하기 위해,

모델 학습에 사용되지 않은 독립적인 테스트 데이터를 별도로 분할하였음. 전체 데이터를 학습데이터, 검증 데이터, 테스트 데이터로 분할하였으며, 각각 8:1:1 비율로 구분하였음.

○ 이와 같은 데이터 분할을 통해

- 학습 데이터는 모델이 패턴을 학습하는 데 사용하고, 검증 데이터는 모델 학습 중 성능을 평가하고 하이퍼 파라미터를 조정하는 데 활용하며, 테스트 데이터는 최종적으로 학습된 모델의 일반화 성능을 검증하는데 사용됨.

○ 〈표 3-12〉는 분할된 데이터별 안전사업장, 위험사업장 비율 및 샘플수를 나타내며, 적용된 이상치 처리 기법에 따라 전체 샘플수는 변동이 있을 수 있음.

〈표 3-12〉 데이터 분할 처리 현황

업종 구분	훈련데이터		검증데이터		테스트데이터	
	샘플수	안전/위험 사업장 비율	샘플수	안전/위험 사업장 비율	샘플수	안전/위험 사업장 비율
제조업	205,479	0.864/ 0.135	25,685	0.864/ 0.135	25,685	0.864/ 0.135
서비스업	902,121	0.906/ 0.094	112,765	0.906/ 0.094	112,766	0.906/ 0.094

(9) 산업안전 데이터 정제 및 전처리 결과

○ 데이터 학습을 최적화하기 위해 다양한 전처리 방법을 적용하고 데이터를 구성하였으나, 각 전처리 기법에 따른 모델의 성능은 크게 유의미하지 않았으며, 이상치 처리와 데이터 불균형 처리 방법에 따른 학습에 필요한 샘플 수가 적어질 경우 모델 성능이 낮아지는 요인이 됨.

○ 전처리 기법 변경에도 성능 변화가 적은 이러한 특성은 기존 고위험사업장 선정 모델의 학습 알고리즘인 XGBoost의 결정 트리에 기반한 특성으로 해석될 수 있음.

1. **결측치에 대한 내재적 처리 능력**: XGBoost는 결측치를 자동으로 처리하는 기능을 갖추고 있어, 별도의 결측치 대체 방법을 적용하지 않아도 모델 학습에 큰 영향을 주지 않음. 따라서 결측치 처리 여부에 따른 성능 변화가 미미하게 나타남.

2. **이상치에 대한 강인성**: XGBoost는 결정 트리 기반의 알고리즘으로, 이상치의 영향이 적음. 이상치가 존재하더라도 트리 분할 기준에 큰 영향을 미치지 않으므로, 이상치 처리의 효과가 제한적일 수 있음.

3. **데이터 스케일링의 영향 미미**: 결정 트리 알고리즘은 변수의 스케일링이나 분포에 크게 민감하지 않기 때문에, 정규화나 표준화와 같은 스케일 조정 전처리가 모델 성능에 큰 변화를 주지 않음. 따라서 코드값 변환이나 스케일 조정의 효과가 미미하게 나타날 수 있음.

4. **불균형 데이터에 대한 내재적 대응**: XGBoost는 모델에서 가중치 조정을 통해 불균형 데이터를 효과적으로 처리할 수 있음. 기본 설정으로도 어느 정도 불균형 데이터를 다룰 수 있어, 추가적인 불균형 처리 전후의 성능 차이가 크지 않을 수 있음.

○ 이러한 이유로, 다양한 전처리 기법을 변경하여 적용하였음에도 모델의 성능에는 큰 변화가 없었던 것으로 판단됨. 이는 XGBoost 알고리즘이 데이터의 결함이나 변환에 대해 내재적으로 강건한 특성을 가지고 있어, 전처리의 영향이 상대적으로 감소되었기 때문임.

○ 결론적으로, 데이터 전처리보다는 모델의 선택이나 하이퍼파라미터 튜닝이 성능 향상에 더 큰 영향을 미칠 수 있음. 따라서 다음 절에서는 다른 모델들을 적용하여 성능을 비교 분석함으로써, 모델 선택이 결과에 미치는 영향을 확인함.

2. 기존 고위험사업장 선정 모델 및 데이터 분석

1) 기존 연구 및 모델 분석

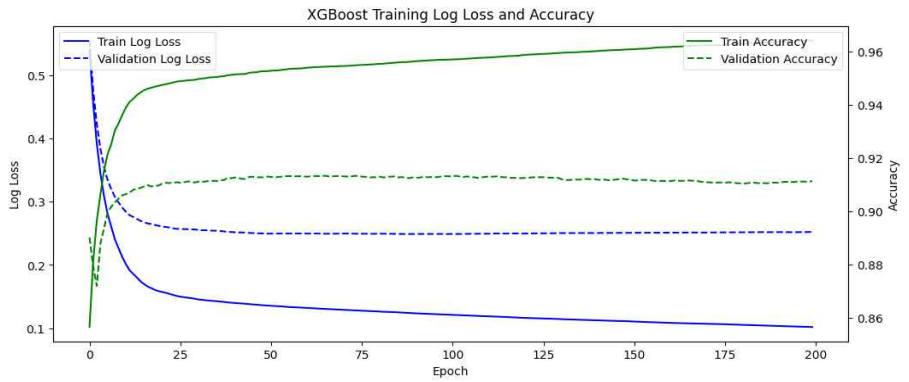
○ 기존의 고위험 사업장 선정 모델을 분석하기 위하여 인공지능(AI) 모델인 XGBoost(eXtreme Gradient Boosting) 알고리즘을 활용하여, 사업장의 위험도를 0~1 사이의 수치로 분석하였음.

○ 〈표 3-13〉은 XGBoost 알고리즘의 고위험 사업장 선정 모델 학습 시 하이퍼파라미터 최적화 과정에 적용한 범위임.

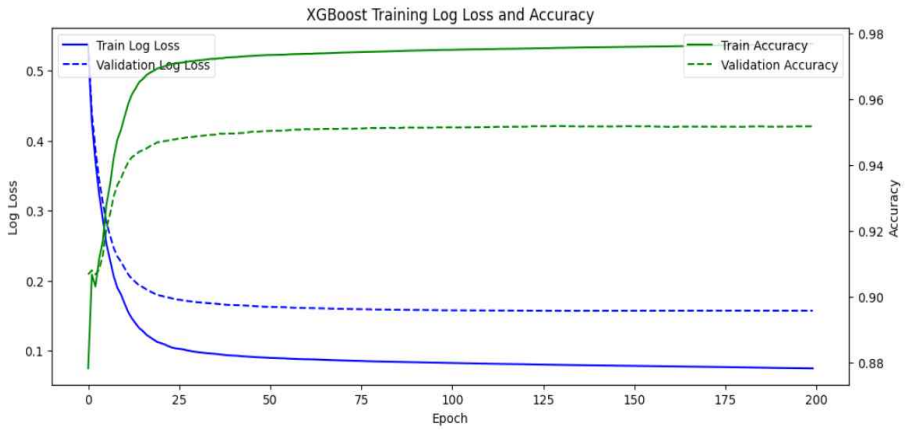
〈표 3-13〉 XGBoost 하이퍼파라미터 적용 범위

구분	적용 범위
부스팅 라운드 수(n_estimators)	100 ~ 1000
트리 최대 깊이(max_depth)	3~10
학습률(learning_rate)	0.01~0.1
샘플 비율(subsample)	0.5~1.0
트리별 특성 샘플 비율(colsample_bytree)	0.3~1.0
레벨별 특성 샘플 비율(colsample_bylevel)	0.3~1.0
분할 기준 최소 손실 감소값(gamma)	0~5
리프 노드의 최소 가중치 합(min_child_weight)	1~10
L2 정규화 항(lambda)	0~10
L1 정규화 항(alpha)	0~10
불균형 클래스에 대한 가중치(scale_pos_weight)	1~5

○ XGBoost 모델을 통해 기본적인 데이터 전처리를 적용하고 학습한 후, 초기 모델의 성능을 아래와 같이 확인하였음.

구분	내용
작업명	제조업_XGBoost_기본설정
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.91 2. 정밀도(Precision): - Label 0 (0.92): 모델이 안전 사업장으로 예측한 것 중 92%가 실제 안전 사업장임을 의미함 - Label 1 (0.80): 모델이 위험 사업장으로 예측한 것 중 80%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.98): 실제 안전 사업장 중 98%가 정확하게 예측함 - Label 1 (0.46): 실제 위험 사업장 중 46%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.95 - Label 1 : 0.58

○ Confusion Matrix와 Classification Report를 통해 제조업 분야 모델의 성능을 분석한 결과, 정확도 0.91로 Label 0(저위험 사업장)에서 모델이 높은 성능을 보이는 반면, Label 1(고위험 사업장)에서는 성능이 상대적으로 낮았음. 특히 재현율(Recall)이 0.46으로 낮아, 위험 사업장을 많이 놓치고 있는 상황임. 이는 모델이 저위험 사업장에 과도하게 학습되고 있음을 나타냄.

구분	내용
작업명	서비스업_XGBoost_기본설정
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.95 2. 정밀도(Precision): - Label 0 (0.96): 모델이 안전 사업장으로 예측한 것 중 96%가 실제 안전 사업장임을 의미함 - Label 1 (0.87): 모델이 위험 사업장으로 예측한 것 중 87%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.99): 실제 안전 사업장 중 99%가 정확하게 예측함 - Label 1 (0.57): 실제 위험 사업장 중 57%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.97 - Label 1 : 0.69

○ 서비스업 모델도 마찬가지로 Label 0(저위험 사업장)에서 모델이 높은 성능을 보이는 반면, Label 1(고위험 사업장)에서는 성능이 상대적으로 낮았음.

○ 고위험사업장 선정 모델을 전체 학습데이터셋을 기초적인 전처리만 한 후 하이퍼파라미터 튜닝을 별도 진행하지 않고 학습한 결과, 제조업, 서비스업 모두 정확도는 전반적으로 높은 수준으로 측정되었음. 그러나 재

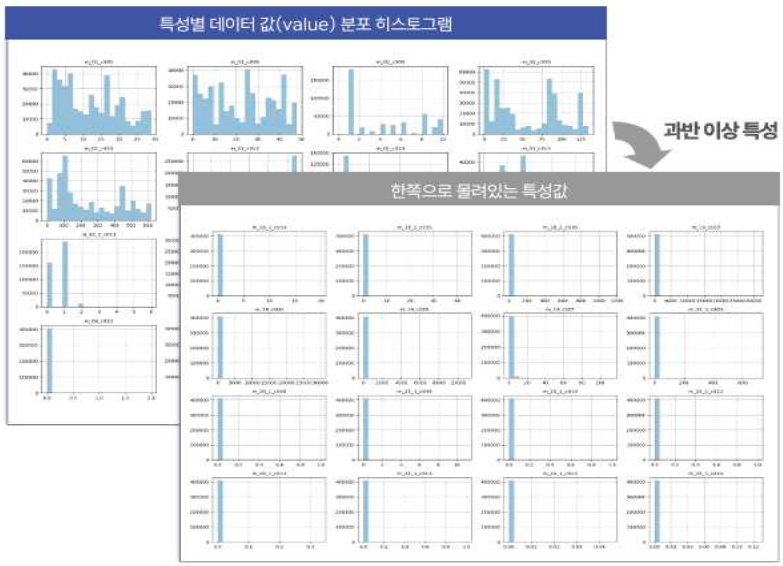
현율이 양쪽 클래스 모두에서 낮게 나타나는 현상이 확인되었음.

○ 고위험사업장을 선정하는 모델의 경우, 실제 고위험사업장을 정확하게 예측하는 것이 중요한 목표이므로, 재현율이 낮다는 것은 모델이 고위험 사업장 예측 성능에서 약점을 가지고 있음을 나타냄.

○ 고위험사업장의 특성상 재현율이 더 중요한 이유는, 실제로 위험한 사업장을 놓치는 일이 발생할 경우 큰 사고나 산재로 이어질 가능성이 있기 때문임. 따라서, 재현율을 높이면서 정밀도와 재현율의 균형을 맞추어 나가는 것을 모델 개선의 과제로 설정하였음. 모델과 데이터 양측을 모두 검토 및 보완하기 위해 여러 조건으로 테스트하고 성능을 평가해 보는 것이 필요하고 성능 개선 가능성 확인에 필요함.

2) 데이터 특성 및 모델 분석

(1) 특성 불균형(Feature Imbalance)



[그림 3-1] 특성 불균형 분석 결과

○ 특성 불균형은 데이터의 값이 특정 범위에 치우치거나 특정 값으로 집중되는 현상을 말하며, 이는 모델 학습과 분석에 있어 중요한 문제로 작용할 수 있음. 본 연구에서는 각 데이터 특성의 분포를 분석하고, 특성 불균형 문제를 해결하기 위한 전처리 방법을 적용함.

1. 특성 값의 분포 확인

- 데이터 내 여러 특성에서 균일하지 않은 값의 분포가 확인되었으며, 특히 일부 특성에서는 한가지 값으로 집중되는 현상도 확인되었음. 이러한 불균형된 분포는 모델 학습 시 편향된 학습 결과를 초래할 수 있음.

2. 특성 분포 분석 및 편향 정도 측정

- 각 특성의 Skewness를 측정하여 데이터의 비대칭성을 확인하였음. Skewness는 데이터 분포의 비대칭 정도를 나타내며, 이 값이 0에 가까우면 정규분포에 가깝고, 1 이상이거나 -1 이하이면 데이터가 강하게 치우쳐 있다고 판단됨. 이를 기준으로 특성별 불균형 여부 파악함.

3. 불균형 문제 해결을 위한 전처리 방법

- 로그 변환(Log Transformation), Box-Cox 변환 등의 전처리 방법을 적용하여 특성 불균형 문제를 해결하고, 데이터 분포를 정규 분포에 가깝게 변환함. 이를 통해 모델이 특성 간 관계를 학습할 수 있도록 함.

○ 〈표 3-14〉과 〈표 3-15〉은 각각 제조업과 서비스업의 특성별 편향 정도를 분석한 결과로, 두 산업 모두에서 대부분 특성이 편향된 분포를 보임.

○ 제조업의 경우 식별특성('사업장관리번호', '사업장개시번호')을 제외한 전체 특성 417개 중 우측 편향 344개, 좌측 편향 10개로 전체 354개(84.9%)가 편향된 분포를 나타내고 있으며, 대칭분포는 63개(15.1%)로 나타났음. 편향정도(Skewness)는 상위 10개의 특성을 보면, 'U판정비율'(640.404)과 '유해물질군명 허가대상 유해물질'(640.404)이 가장 크게 나타났으며, '취급인원'(639.799), '검진종목별 수진자수(수

첩)'(557.545), '직종 대분류 보건·의료직 합계'(552.449), 곤돌라(509.289), 'C2판정비율'(500.503), '물질군명 노출기준제정물질'(481.753), 'D1판정비율'(492.455), '물질군명 분진합'(481.753) 순으로 나타남.

○ 서비스업의 경우 전체 특성 471개 중 우측 편향 363개, 좌측 편향 15개로 전체 378개(80.3%)가 편향된 분포를 나타내고 있으며, 대칭분포는 93개(19.7%)로 나타났음. 편향정도(Skewness)는 상위 10개의 특성을 보면, '위험설비 타워크레인'(792.245), '위험율허가 유별 여부'(792.245), '밀폐 강재등시설'(792.245), '밀폐 중독위험장소'(792.245), 검진종목별 수진자수(임시)(792.245), 검진종목별 수진자수(수첩)(792.245)이 가장 크게 나타났으며, '밀폐공간수 총합'(791.469), '진동발생작업 종사근로자수합'(750.6), '고열 한랭 다습 및 방사선 취급 작업 종사근로자수합'(710.32), '직종 대분류 설치·장비·생산직 합계(561.697), 순으로 나타남.

○ 이를 통해, 중요 특성을 효과적으로 활용하고 하이퍼파라미터 튜닝을 통해 정규화 기법을 적용하거나, 중요 특성을 선별하여 모델 성능을 개선할 필요성을 확인하였음.

〈표 3-14〉 제조업 특성 불균형 목록

번호	사업명	특성명	Skewness	불균형 여부
1	제조업 사업장 리스트	사업장관리번호	-	키 값
		사업장개시번호	-	키 값
		일선기관	0.394	대칭 분포
		노동지청	0.109	대칭 분포
		중업종명	0.47	대칭 분포
		소업종명	0.223	대칭 분포
		표준산업분류	0.499	대칭 분포
		근로자수	176.618	우측 편향
		규모1	-1.904	좌측 편향
		행정구역	0.778	대칭 분포
		행정구역_세부	0.236	대칭 분포

번호	사업명	특성명	Skewness	불균형 여부
2	패트롤 수행 결과 (현장점검 정보)	동행한타기관	5.432	우측 편향
		선정기준_공단 자체	0.125	대칭 분포
		선정기준_재해예방기관 등 기타	4.71	우측 편향
		점검결과조치_개선확인 후 종결(미개선시 감독연계)	1.882	우측 편향
		점검결과조치_사업장 자체개선 후 종결/점검 종결	0.625	대칭 분포
		경영자마인드	-0.431	대칭 분포
		안전보건관리및개선노력	-0.588	대칭 분포
		안전관리수준평가사업장위험도_현장위험관리수준	-1.659	좌측 편향
		안전보건수준평가종합	-1.15	좌측 편향
		점검차수	4.343	우측 편향
	패트롤 수행 결과 (시정부적합정보)	사고유발요인_갯수	1.436	우측 편향
		위험기인물_그 밖의 위험	1.246	우측 편향
		위험기인물_끼임	0.87	대칭 분포
		위험기인물_떨어짐	1.49	우측 편향
		위험기인물_부딪힘	1.714	우측 편향
		위험기인물_질식	12.777	우측 편향
		위험기인물_화재	2.735	우측 편향
		안전관리수준평가사업장위험도_현장위험관리수준	-0.239	대칭 분포
		점검차수	0.07	대칭 분포
	패트롤 수행 결과 (위험설비보유정보)	보유건수	13.25	우측 편향
		위험설비_분쇄,파쇄기	4.977	우측 편향
		위험설비_사출성형기	4.729	우측 편향
		위험설비_산업용로봇	7.514	우측 편향
		위험설비_승강기(리프트 포함)	3.169	우측 편향
		위험설비_식품가공용기계	6.636	우측 편향
		위험설비_지게차	0.209	대칭 분포
		위험설비_컨베이어	3.081	우측 편향
		위험설비_크레인(천장,갠트리)	0.442	대칭 분포
		위험설비_타워크레인	19.446	우측 편향
		위험설비_프레스	2.313	우측 편향
		위험설비_혼합기	5.544	우측 편향
		점검차수	4.343	우측 편향
3	고용보험_ 근로자정보	피보험자_합계	198.057	우측 편향
		만나이_평균	-0.454	대칭 분포
		근속기간_년수_평균	2.111	우측 편향
		성별_남_합계	195.348	우측 편향
		성별_여_합계	289.136	우측 편향

번호	사업명	특성명	Skewness	불균형 여부
		연령대_10대_20대_합계	240.002	우측 편향
		연령대_30대_합계	274.104	우측 편향
		연령대_40대_합계	184.878	우측 편향
		연령대_50대_합계	247.065	우측 편향
		연령대_60대이상_합계	314.596	우측 편향
		근속기간범주_1년이하_합계	0	대칭 분포
		근속기간범주_1년초과3년이하_합계	0	대칭 분포
		근속기간범주_3년초과5년이하_합계	0	대칭 분포
		근속기간범주_5년초과10년이하_합계	0	대칭 분포
		근속기간범주_10년초과20년이하_합계	0	대칭 분포
		근속기간범주_20년초과_합계	0	대칭 분포
		직종_대분류_건설·채굴직_합계	309.078	우측 편향
		직종_대분류_교육·법률·사회복지·경찰·소방직및군인_합계	158.042	우측 편향
		직종_대분류_농림어업직_합계	103.881	우측 편향
		직종_대분류_미용·여행·숙박·음식·경비·청소직_합계	153.094	우측 편향
		직종_대분류_보건·의료직_합계	552.449	우측 편향
		직종_대분류_설치·정비·생산직_합계	236.418	우측 편향
		직종_대분류_연구직및공학기술직_합계	322.577	우측 편향
		직종_대분류_영업·판매·운전·운송직_합계	367.796	우측 편향
		직종_대분류_예술·디자인·방송·스포츠직_합계	189.244	우측 편향
4	안전보건관계자	전담유무	18.145	우측 편향
		건설안전관리자	369.735	우측 편향
		명예산업안전감독관	30.139	우측 편향
		보건관리자	6.864	우측 편향
		사업장담당자	3.735	우측 편향
		산업보건의	30.551	우측 편향
		안전관리자	6.339	우측 편향
		안전보건관리책임자	11.839	우측 편향
		안전보건총괄책임자	116.909	우측 편향
		관리자	11.972	우측 편향
		보건담당자	7.578	우측 편향

번호	사업명	특성명	Skewness	불균형 여부
		선임자총수	4.956	우측 편향
		선임자종류수	4.956	우측 편향
		안전보건관계자_데이터존재여부	2.95	우측 편향
5	재정지원 (안전투자혁신사업)	지원금액	10.313	우측 편향
		사업구분_고소작업대	11.892	우측 편향
		사업구분_고위험 TOP3 업종	19.383	우측 편향
		사업구분_노후 위험기계기구(30년이상)	21.089	우측 편향
		사업구분_리프트	-4.735	좌측 편향
		사업구분_뿌리공정	12.508	우측 편향
		사업구분_이동식크레인	38.301	우측 편향
	재정지원 (융자지원)	대하금액(천원)	7.316	우측 편향
	재정지원 (클린사업장)	교부금액	10.435	우측 편향
6	유해위험기계기구 (안전검사+자율)	컨베이어종류_밸트	99.371	우측 편향
		컨베이어종류_체인	314.558	우측 편향
		컨베이어종류_롤러	234.528	우측 편향
		컨베이어종류_트롤리	263.323	우측 편향
		컨베이어종류_버킷	123.368	우측 편향
		컨베이어종류_나사	337.316	우측 편향
		고소작업대	56.052	우측 편향
		곤돌라	509.289	우측 편향
		국소배기장치	206.797	우측 편향
		롤러기	94.039	우측 편향
		리프트	286.158	우측 편향
		사출성형기	81.632	우측 편향
		산업용로봇	217.479	우측 편향
		압력용기	256.989	우측 편향
		원심기	90.455	우측 편향
		전단기	70.335	우측 편향
		컨베이어	256.907	우측 편향
		크레인	313.712	우측 편향
		프레스	28.235	우측 편향
		심사결과_적합	274.421	우측 편향
		심사결과_부적합	185.948	우측 편향
		안전검사_사업수행여부	1.89	우측 편향
		자율검사_사업수행여부	28.559	우측 편향
7	소방청 위험물 제조소	예방규정제출대상여부	2.137	우측 편향
		대량위험물제조소등여부	50.919	우측 편향
		석유화학단지내사업장여부	19.687	우측 편향
		소화난이도등급_1등급	28.006	우측 편향
		소화난이도등급_2등급	10.234	우측 편향
		소화난이도등급_3등급	20.16	우측 편향

번호	사업명	특성명	Skewness	불균형 여부
		설치기간_10년이상	17.764	우측 편향
		설치기간_20년이상	26.086	우측 편향
		설치기간_5년미만	24.438	우측 편향
		설치기간_5년이상	13.504	우측 편향
		위험물제조소_총합	23.81	우측 편향
		위험물제조소_종류	4.436	우측 편향
		위험물허가_유별_제1류	29.576	우측 편향
		위험물허가_유별_제2류	13.221	우측 편향
		위험물허가_유별_제3류	23.287	우측 편향
		위험물허가_유별_제4류	12.456	우측 편향
		위험물허가_유별_제5류	42.466	우측 편향
		위험물허가_유별_제6류	32.234	우측 편향
		위험물허가_유별_합계	13.332	우측 편향
		위험물허가_유별_여부	0	대칭 분포
		탱크총합	15.715	우측 편향
		탱크여부	-0.769	대칭 분포
		이송취급소수	26.791	우측 편향
		이송취급소여부	8.637	우측 편향
		이동탱크수	16.878	우측 편향
		이동탱크여부	4.222	우측 편향
8	지게차 실태조사	지게차대수_자가	8.836	우측 편향
		지게차대수_그외	12.378	우측 편향
		지게차보유대수	8.985	우측 편향
		지게차용량	241.433	우측 편향
		충돌방지장치점수	0.604	대칭 분포
		안전띠점수	0.053	대칭 분포
		법정방호장치점수	-0.127	대칭 분포
		운전자격점수	0.325	대칭 분포
		관리자이해점수	0.538	대칭 분포
		운전자관찰점수	0.641	대칭 분포
		점수총합	0.761	대칭 분포
9	KOSHA_MS_18001	KOSHA_MS18001_사업수행여부	21.923	우측 편향
10	위험성평가 (컨설팅)	사업주의관심도	0.298	대칭 분포
		위험성평가실행수준	0.502	대칭 분포
		구성원의참여및이해수준	0.443	대칭 분포
		재해발생수준	-0.302	대칭 분포
	위험성평가 (인정/불인정)	사업주의 관심도	0.584	대칭 분포
		위험성평가 실행수준	0.943	대칭 분포
		구성원의 참여 및 이해수준	0.601	대칭 분포
		심사결과_불인정	-0.278	대칭 분포
11	민간위탁기술지도 (안전)	선정기준코드_물질관련	8.943	우측 편향
		선정기준명_신규사업장	2.695	우측 편향
		선정기준명_유해물질존재사업장	20.076	우측 편향

번호	사업명	특성명	Skewness	불균형 여부
		선정기준명_재해발생사업장	3.893	우측 편향
		선정기준명_특별대책사업장	0.959	대칭 분포
		지원횟수	-1.001	좌측 편향
		전담	0.54	대칭 분포
		겸직	2.935	우측 편향
		등급_중급	1.268	우측 편향
		등급_초급	1	대칭 분포
		교육인원합	3.01	우측 편향
		자료제공_총합	233.605	우측 편향
		검사실시_종류수	1.07	우측 편향
		검사미실시_종류수	3.959	우측 편향
		검사비대상_종류수	1.233	우측 편향
		검사실시_합	6.571	우측 편향
		검사미실시_합	17.585	우측 편향
		검사비대상_합	7.107	우측 편향
		전체_개선	1.668	우측 편향
		전체_조치의뢰	18.466	우측 편향
	민간위탁기술지도 (보건)	처리사유_개선완료	13.648	우측 편향
		처리사유_조치의뢰	34.535	우측 편향
		선정기준코드_산재관련	1.932	우측 편향
		선정기준코드_실비관련	8.092	우측 편향
		선정기준코드_물질관련	8.943	우측 편향
		선정기준명_신규사업장	2.695	우측 편향
		선정기준명_유해물질존재사업장	20.076	우측 편향
		선정기준명_재해발생사업장	3.893	우측 편향
		선정기준명_특별대책사업장	0.959	대칭 분포
		지원횟수	-1.001	좌측 편향
		전담	0.54	대칭 분포
		겸직	2.935	우측 편향
		등급_중급	1.268	우측 편향
		등급_초급	1	대칭 분포
		교육인원_총수	2.677	우측 편향
		자료제공_총합	233.605	우측 편향
		밀폐실태밀폐공간작업유무	8.631	우측 편향
		밀폐실태_밀폐공간장소_통의내부등	53.155	우측 편향
		밀폐실태_밀폐공간장소_정화조등	1.061	우측 편향
		밀폐실태_밀폐공간장소_기타밀폐공간	-0.567	대칭 분포
		밀폐실태_밀폐공간장소_반응기등내부	10.254	우측 편향
		밀폐실태_밀폐공간장소_콘크	32.34	우측 편향

번호	사업명	특성명	Skewness	불균형 여부
		리트양생		
		밀폐실태_밀폐공간장소_강재 등시설	70.7	우측 편향
		밀폐실태_밀폐공간장소_불활성기체설비	18.398	우측 편향
		밀폐실태-질식사고 인지도	2.531	우측 편향
		밀폐실태-질식사고 위험관리 인지도	2.836	우측 편향
		밀폐실태-질식사고 교육이수	1.767	우측 편향
		밀폐실태-가스농도측정기 보유	1.692	우측 편향
		급기팬 보유	1.87	우측 편향
		밀폐실태-위험도평가 총점	1.807	우측 편향
		개선필요_합	0.449	대칭 분포
		개선_합	0.683	대칭 분포
		최종실태평가결과	-1.281	좌측 편향
		실태평가결과_사업주의지	-2.289	좌측 편향
	민간위탁기술지도 (화학)	처리사유내용_종결	-1.744	좌측 편향
		처리사유내용_조치의뢰	10.754	우측 편향
		선정기준코드_산재관련	1.932	우측 편향
		선정기준코드_실비관련	8.092	우측 편향
		선정기준코드_물질관련	8.943	우측 편향
		선정기준명_신규사업장	2.695	우측 편향
		선정기준명_유해물질존재사업장	20.076	우측 편향
		선정기준명_재해발생사업장	3.893	우측 편향
		선정기준명_특별대책사업장	0.959	대칭 분포
		지원횟수	-1.001	좌측 편향
		전담	0.54	대칭 분포
		겸직	2.935	우측 편향
		등급_중급	1.268	우측 편향
		등급_초급	1	대칭 분포
		교육인원_총수	2.677	우측 편향
		자료제공_총합	233.605	우측 편향
		기술지원_사고성재해예방	1.066	우측 편향
		기술지원_화학사고예방	-0.451	대칭 분포
		유해위험물질수	3.041	우측 편향
		화학설비건수	17.738	우측 편향
		위험기계기구_보유건수	206.72	우측 편향
		개선_총계	0.816	대칭 분포
		조치의뢰_총계	26.83	우측 편향
		국소배기장치	206.797	우측 편향
		롤러기	94.039	우측 편향
		리프트	286.158	우측 편향

번호	사업명	특성명	Skewness	불균형 여부
		분쇄파쇄기	15.313	우측 편향
		사출성형기	81.632	우측 편향
		산업용로봇	217.479	우측 편향
		식품제조용설비	13.215	우측 편향
		원심기	90.455	우측 편향
		전단기	70.335	우측 편향
		지게차	1.029	우측 편향
		컨베이어	256.907	우측 편향
		크레인	313.712	우측 편향
		프레스	28.235	우측 편향
		혼합기	14.489	우측 편향
		압력용기	256.989	우측 편향
		위험기계기구총합	347.398	우측 편향
		밀폐_정화조등	9.213	우측 편향
		밀폐_통의내부등	60.76	우측 편향
		밀폐_불활성기체설비	68.246	우측 편향
		밀폐_강재등시설	80.032	우측 편향
		밀폐_반응기등내부	23.32	우측 편향
		밀폐_콘크리트양생	261.439	우측 편향
		밀폐_중독위험장소	143.189	우측 편향
		밀폐_기타밀폐공간	11.963	우측 편향
12	공단교육 (안전보건교육)	교육분야코드_관리자	79.163	우측 편향
		교육분야코드_안전보건관계자	50.092	우측 편향
		교육분야코드_일반근로자	89.714	우측 편향
		교육분야코드_취약계층	133.646	우측 편향
	공단교육 (인터넷교육센터)	교육대상_근로자	168.385	우측 편향
		교육대상_책임자	120.833	우측 편향
		교육대상_특수형태근로자	78.941	우측 편향
		수료여부_미수료	39.693	우측 편향
		수료비율	3.371	우측 편향
	공단교육 (직무교육센터)	과정구분_온라인	22.803	우측 편향
		과정구분_집체	18.979	우측 편향
		교육대상_전문기관종사자	199.678	우측 편향
		수료여부_미수료	39.693	우측 편향
		교육대상_안전보건관계자	17.489	우측 편향
		교육수료비율	5.052	우측 편향
13	유해위험방지계획서	대상설비합계	55.824	우측 편향
		대상규모명_over_2000	192.096	우측 편향
		대상규모명_under_2000	13.999	우측 편향
		대상규모명_under_500	30.731	우측 편향
		전기계약용량변경	33.874	우측 편향
		사업구분_변경	126.453	우측 편향
		사업구분_설치	13.557	우측 편향
		사업구분_이전	27.813	우측 편향

번호	사업명	특성명	Skewness	불균형 여부
		최종확인회차	8.765	우측 편향
		심사결과_반려부적정	79.888	우측 편향
		심사결과_적정조건부적정	69.571	우측 편향
		고용부조치통보	23.548	우측 편향
		유해위험방지계획서_사업수행 여부	7.259	우측 편향
14	작업환경측정 (측정)	물질군명_노출기준제정물질	494.607	우측 편향
		물질군명_허가대상 유해물질	369.735	우측 편향
		지원대상구분_대상	58.017	우측 편향
		초과율_평균	26.046	우측 편향
		취급인원	639.799	우측 편향
		물질군명_물리적인자_합	327.491	우측 편향
		물질군명_화학적인자_합	382.966	우측 편향
		물질군명_분진_합	481.753	우측 편향
		취급구분_사용	113.449	우측 편향
		취급구분_제조	371.112	우측 편향
		취급용도_기타	153.723	우측 편향
	작업화경측정 (화학물질취급현황)	취급용도_세척	94.234	우측 편향
		취급용도_시약	232.423	우측 편향
		취급용도_실험	132.38	우측 편향
		취급용도_용접	119.087	우측 편향
		취급용도_원료	419.906	우측 편향
		취급물질군명_기타유해물질	161.399	우측 편향
		취급물질군명_노출기준제정물질	188.754	우측 편향
		취급물질군명_제조금지 유해물질	458.662	우측 편향
		취급물질군명_허가대상 유해물질	242.858	우측 편향
		취급물질군명_물리적인자_통합	475.294	우측 편향
		취급물질군명_분진인자_통합	139.064	우측 편향
		취급물질군명_화학적인자_통합	181.368	우측 편향
15	고용보험ERP 근로자수	고용상시인원수	102.65	우측 편향
		남성근로자수	131.594	우측 편향
		여성근로자수	81.86	우측 편향
		장년근로자수	0	대칭 분포
		외국인근로자수	11.534	우측 편향
		장애인근로자수	0	대칭 분포
16	특수건강진단 (특검)	총검진자수(명)	315.919	우측 편향
		유해물질군명_노출기준제정물질	452.832	우측 편향
		유해물질군명_야간작업	188.076	우측 편향
		유해물질군명_제조금지 유해물질	0	대칭 분포
		유해물질군명_허가대상	640.404	우측 편향

번호	사업명	특성명	Skewness	불균형 여부
		유해물질		
		A판정비율	107.903	우측 편향
		C1판정비율	331.313	우측 편향
		C2판정비율	500.503	우측 편향
		D1판정비율	492.455	우측 편향
		D2판정비율	328.4	우측 편향
		CN판정비율	362.136	우측 편향
		DN판정비율	287.8	우측 편향
		U판정비율	640.404	우측 편향
		유해물질군명_물리적인자_합	169.371	우측 편향
		유해물질군명_화학적인자_합	274.687	우측 편향
		유해물질군명_분진_합	243.058	우측 편향
	특수건강진단 (사업장별검진내역)	검진종목별 수진자수(일반)	163.3	우측 편향
		검진종목별 수진자수(특수)	145.27	우측 편향
		검진종목별 수진자수(배치전)	207.607	우측 편향
		검진종목별 수진자수(수시)	409.718	우측 편향
		검진종목별 수진자수(임시)	449.945	우측 편향
		검진종목별 수진자수(수첩)	557.545	우측 편향
		검진종목별_총합	155.066	우측 편향
17	산업안전보건실태조사	교대근무제여부	20.096	우측 편향
		노동조합여부	12.042	우측 편향
		산업안전보건위원회여부	12.748	우측 편향
		작업환경관련위험요인	13.473	우측 편향
		신체적부담관련위험요인	13.423	우측 편향
		생화학물질관련위험요인	14.779	우측 편향
		기계전기기타위험요인	13.808	우측 편향
		위험성평가	12.782	우측 편향
		스트레스심각도	13.167	우측 편향
		스트레스관리노력정도	12.101	우측 편향
		경영진안전보건의지	11.658	우측 편향
		사업장내안전문화	11.701	우측 편향
		근로자안전보건의지	11.793	우측 편향
		일반건강진단사후관리	3.509	우측 편향
		특수건강진단사후관리	3.587	우측 편향
		유해인자축소노력여부	11.718	우측 편향
		상주협력업체	11.186	우측 편향
		상주협력업체수	6.305	우측 편향
		상주협력업체근로자수	28.323	우측 편향
		원청회사여부	24.213	우측 편향
		법인지여부	11.928	우측 편향
18	산업재해조사표	재해자동종경력년수_10~20	194.794	우측 편향
		재해자동종경력년수_1~3	236.996	우측 편향
		재해자동종경력년수_20~	233.777	우측 편향
		재해자동종경력년수_3~5	310.242	우측 편향

번호	사업명	특성명	Skewness	불균형 여부
		재해자동종경력년수_5~10	240.165	우측 편향
		상해부위_기타	82.105	우측 편향
		상해부위_다리	229.023	우측 편향
		상해부위_다발성	267.969	우측 편향
		상해부위_머리	213.472	우측 편향
		상해부위_몸통	216.41	우측 편향
		상해부위_전신	221.607	우측 편향
		상해부위_팔	212.921	우측 편향
19	작업환경실태조사 (일반현황)	전기계약용량	2.473	우측 편향
		야간작업유무	3.127	우측 편향
		정비_보수여부	6.436	우측 편향
		하청사업장수	60.319	우측 편향
		하청근로자수	263.224	우측 편향
		교대근무여부	3.539	우측 편향
		근골격계부담작업대상여부	1.783	우측 편향
		유해요인조사실시여부	2.48	우측 편향
		복지시설_개수	1.864	우측 편향
		원청	1.037	우측 편향
		하청	9.524	우측 편향
	작업환경실태조사 (화학물질취급)	취급	1.906	우측 편향
		생산	133.523	우측 편향
		허용대상물질여부	242.045	우측 편향
		허용기준물질여부	5.575	우측 편향
		관리대상물질여부	1.14	우측 편향
		안전검사물질여부	2.543	우측 편향
		안전관리물질여부	3.076	우측 편향
		기타물질여부	2.152	우측 편향
		특검대상물질여부	1.392	우측 편향
		측정대상물질여부	0.983	대칭 분포
		PSM대상물질여부	5.602	우측 편향
		건강관리수첩대상물질여부	5.318	우측 편향
		사고대상물질여부	2.993	우측 편향
		금지대상물질여부	58.929	우측 편향
		근로자_월_취급시간	420.696	우측 편향
	작업환경실태조사 (기계기구설비현황)	기계설비_제조_보유갯수	17.482	우측 편향
		기계설비_제조_보유종류	1.599	우측 편향
		기계설비_비제조_보유갯수	100.321	우측 편향
		기계설비_비제조_보유종류	0.815	대칭 분포
	작업환경실태조사 (작업환경)	소음발생공정수	173.153	우측 편향
		밀폐공간수	434.941	우측 편향
		작업환경_고열/한랭/다습및방사선취급작업	6.476	우측 편향
		작업환경_밀폐공간(산소결핍위험장소)현황	5.244	우측 편향

번호	사업명	특성명	Skewness	불균형 여부
		작업환경_분진/흄발생작업	3.555	우측 편향
		작업환경_사내도급작업	207.238	우측 편향
		작업환경_소음작업	1.438	우측 편향
		작업환경_제조나노물질의제조 및취급작업	20.271	우측 편향
		작업환경_진동발생작업	4.934	우측 편향
20	민간위탁기관평가 데이터	위탁기관평가_점수	0.599	대칭 분포
22	재해데이터	재해율3년평균	6.373	우측 편향
24	감성평가	사업주관리자마인드	-0.118	대칭 분포
		근로자안전보건행동수준	-0.279	대칭 분포
		작업장 및 근로자환경수준	-0.422	대칭 분포
25	PSM	PSM_사업수행여부	21.593	우측 편향

※ 불균형 여부의 '대칭 분포'의 기준은 Skewness값이 0~1인 경우임

〈표 3-15〉 서비스업 특성 불균형 목록

번호	사업명	특성명	Skewness	불균형 여부
1	서비스업 사업장 리스트	사업장관리번호	-	키 값
		사업장개시번호	-	키 값
		일선기관	0.736	대칭 분포
		소업종명	2.987	우측 편향
		규모1	2.093	우측 편향
2	패트롤 수행 결과 (서비스)	동행한타기관	0	대칭 분포
		선정기준_재해예방기관 등 기타	37.142	우측 편향
		경영자마인드	5.97	우측 편향
		안전보건관리및개선노력	5.956	우측 편향
		안전관리수준평가사업장위험도_현장위험관리수준	6.001	우측 편향
		안전보건수준평가종합	5.916	우측 편향
		점검차수	23.083	우측 편향
	패트롤 수행 결과 (시정 부적합 정보)	사고유발요인_갯수	5.199	우측 편향
		위험기인물_그 밖의 위험	5.567	우측 편향
		위험기인물_끼임	12.731	우측 편향
		위험기인물_떨어짐	22.307	우측 편향
		위험기인물_부딪힘	28.825	우측 편향
		위험기인물_질식	30.399	우측 편향
		위험기인물_화재_폭발	19.291	우측 편향
		안전관리수준평가사업장위험도_현장위험관리수준	6.001	우측 편향
		점검차수	23.083	우측 편향
	패트롤 수행 결과	보유건수	10.14	우측 편향

번호	사업명	특성명	Skewness	불균형 여부
	(위험설비보유정보)	위험설비_분쇄_파쇄기	14.471	우측 편향
		위험설비_산업용로봇	32.515	우측 편향
		위험설비_승강기(리프트 포함)	31.77	우측 편향
		위험설비_식품가공용기계	89.688	우측 편향
		위험설비_지게차	5.783	우측 편향
		위험설비_컨베이어	9.285	우측 편향
		위험설비_크레인_천장_갠트리	6.566	우측 편향
		위험설비_타워크레인	792.245	우측 편향
		위험설비_프레스	70.004	우측 편향
		위험설비_혼합기	23.328	우측 편향
		점검차수	23.083	우측 편향
3	고용보험 근로자정보	피보험자_합계	182.958	우측 편향
		만나이_평균	0.422	대칭 분포
		근속기간_년수_평균	3.23	우측 편향
		성별_남_합계	182.179	우측 편향
		성별_여_합계	210.32	우측 편향
		연령대_10대_20대_합계	436.126	우측 편향
		연령대_30대_합계	233.91	우측 편향
		연령대_40대_합계	204.494	우측 편향
		연령대_50대_합계	212.698	우측 편향
		연령대_60대이상_합계	194.325	우측 편향
		근속기간범주_1년이하_합계	0	대칭 분포
		근속기간범주_1년초과3년이하_합계	0	대칭 분포
		근속기간범주_3년초과5년이하_합계	0	대칭 분포
		근속기간범주_5년초과10년이하_합계	0	대칭 분포
		근속기간범주_10년초과20년이하_합계	0	대칭 분포
		근속기간범주_20년초과_합계	0	대칭 분포
		직종_대분류_건설·채굴직_합계	168.588	우측 편향
		직종_대분류_교육·법률·사회복지·경찰·소방직및군인_합계	265.775	우측 편향
		직종_대분류_농림어업직_합계	318.563	우측 편향
		직종_대분류_미용·여행·숙박·음식·경비·청소직_합계	375.731	우측 편향
		직종_대분류_보건·의료직_합계	561.654	우측 편향
		직종_대분류_설치·정비·생산직_합계	561.697	우측 편향
		직종_대분류_연구직및공학기술직_합계	147.343	우측 편향
		직종_대분류_영업·판매·운전·운송직_합계	316.546	우측 편향

번호	사업명	특성명	Skewness	불균형 여부
		직종_대분류_예술·디자인·방송·스포츠직_합계	314.997	우측 편향
4	안전보건관계자	전담유무	50.036	우측 편향
		건설안전관리자	354.299	우측 편향
		명예산업안전감독관	128.528	우측 편향
		보건관리자	2.908	우측 편향
		사업장담당자	0	대칭 분포
		산업보건의	76.427	우측 편향
		안전관리자	3.046	우측 편향
		안전보건관리책임자	40.42	우측 편향
		안전보건총괄책임자	211.73	우측 편향
		관리자	41.536	우측 편향
		보건담당자	20.519	우측 편향
		선임자총수	20.199	우측 편향
		선임자종류수	15.851	우측 편향
5	재정지원 (안전투자혁신사업)	지원금액	84.698	우측 편향
		사업구분_고소작업대	125.253	우측 편향
		사업구분_고위험 TOP3 업종	0	대칭 분포
		사업구분_노후 위험기계기구(30년이상)	560.2	우측 편향
		사업구분_리프트	250.524	우측 편향
		사업구분_뿌리공정	0	대칭 분포
		사업구분_이동식크레인	95.36	우측 편향
		안투지원여부	72.002	우측 편향
		사업수행여부	7.858	우측 편향
	재정지원 (융자지원)	대하금액(천원)	26.519	우측 편향
		융자지원여부	10.313	우측 편향
	재정지원 (클린사업장)	교부금액	28.953	우측 편향
		클린지원여부	10.915	우측 편향
6	유해위험기계기구 (안전검사)	컨베이어(구간내컨베이어종류)_밸트	7.789	우측 편향
		컨베이어(구간내컨베이어종류)_체인	8.41	우측 편향
		컨베이어(구간내컨베이어종류)_롤러	19.625	우측 편향
		컨베이어(구간내컨베이어종류)_트롤리	43.051	우측 편향
		컨베이어(구간내컨베이어종류)_버킷	11.002	우측 편향
		컨베이어(구간내컨베이어종류)_나사	9.01	우측 편향
		검사대상품_대상품_대_고소작업대	11.895	우측 편향
		검사대상품_대상품_대_곤돌라	62.966	우측 편향

번호	사업명	특성명	Skewness	불균형 여부
		검사대상품_대상품_대_국소배기장치	29.582	우측 편향
		검사대상품_대상품_대_롤러기	62.826	우측 편향
		검사대상품_대상품_대_리프트	11.633	우측 편향
		검사대상품_대상품_대_사출성형기	30.539	우측 편향
		검사대상품_대상품_대_산업용로봇	61.069	우측 편향
		검사대상품_대상품_대_압력용기	4.211	우측 편향
		검사대상품_대상품_대_원심기	109.851	우측 편향
		검사대상품_대상품_대_전단기	8.82	우측 편향
		검사대상품_대상품_대_컨베이어	8.47	우측 편향
		검사대상품_대상품_대_크레인	2.587	우측 편향
		검사대상품_대상품_대_프레스	35.296	우측 편향
		심사결과_심사결과_반려	34.467	우측 편향
		심사결과_심사결과_부적합	3.347	우측 편향
		심사결과_심사결과_적합	4.054	우측 편향
		심사결과_심사결과_진행	0	대칭 분포
		자진신고여부_유	7.665	우측 편향
		사업수행여부	7.858	우측 편향
	유해위험기계기구(자율안전검사)	인증대상품_대상품_대_곤돌라	4.755	우측 편향
		인증대상품_대상품_대_국소배기장치	161.707	우측 편향
		인증대상품_대상품_대_롤러기	0	대칭 분포
		인증대상품_대상품_대_리프트	0	대칭 분포
		인증대상품_대상품_대_사출성형기	0	대칭 분포
		인증대상품_대상품_대_산업용로봇	0	대칭 분포
		인증대상품_대상품_대_압력용기	18.079	우측 편향
		인증대상품_대상품_대_원심기	0	대칭 분포
		인증대상품_대상품_대_전단기	0	대칭 분포
		인증대상품_대상품_대_컨베이어	-0.003	대칭 분포
		인증대상품_대상품_대_크레인	16.974	우측 편향
		인증대상품_대상품_대_프레스	0	대칭 분포
		사업수행여부	7.858	우측 편향
7	소방청 위험물 제조소 등 정보	예방규정제출대상여부	3.526	우측 편향
		대량위험물제조소등여부	9.544	우측 편향
		석유화학단지내사업장여부	15.827	우측 편향
		소화난이도등급_1등급	18.852	우측 편향
		소화난이도등급_2등급	5.579	우측 편향
		소화난이도등급_3등급	4.077	우측 편향
		설치기간_10년이상	10.279	우측 편향
		설치기간_20년이상	13.857	우측 편향

번호	사업명	특성명	Skewness	불균형 여부
		설치기간_5년미만	4.678	우측 편향
		설치기간_5년이상	6.303	우측 편향
		위험물제조소_총합	12.338	우측 편향
		위험물제조소_종류	3.743	우측 편향
		위험물허가_유별_제1류	24.088	우측 편향
		위험물허가_유별_제2류	15.916	우측 편향
		위험물허가_유별_제3류	21.992	우측 편향
		위험물허가_유별_제4류	20.162	우측 편향
		위험물허가_유별_제5류	24.675	우측 편향
		위험물허가_유별_제6류	36.57	우측 편향
		위험물허가_유별_합계	19.863	우측 편향
		위험물허가_유별_여부	-792.245	좌측 편향
		탱크총합	7.77	우측 편향
		탱크여부	-2.346	좌측 편향
		이송취급소수	60.578	우측 편향
		이송취급소여부	9.087	우측 편향
		이동탱크수	2.759	우측 편향
		이동탱크여부	0.854	대칭 분포
8	지게차 실태조사	지게차대수_자가	0.963	대칭 분포
		지게차대수_그외	2.26	우측 편향
	지게차 실태조사 (안전관리체계화 수행 결과)	지게차보유대수	8.884	우측 편향
		지게차용량	6.154	우측 편향
		충돌방지장치점수	0.918	대칭 분포
		안전띠점수	0.393	대칭 분포
		법정방호장치점수	-0.506	대칭 분포
		운전자격점수	0.421	대칭 분포
		관리자이해점수	0.958	대칭 분포
		운전자관찰점수	1.198	우측 편향
		점수총합	0.872	대칭 분포
9	KOSHA_MS_18001	사업대상여부	11.637	우측 편향
10	위험성평가 (컨설팅)	사업주의관심도	0.951	대칭 분포
		위험성평가실행수준	0.599	대칭 분포
		구성원의참여및이해수준	0.605	대칭 분포
	위험성평가 (인정/불인정)	사업주의 관심도	-1.093	좌측 편향
		위험성평가 실행수준	-1.457	좌측 편향
		구성원의 참여 및 이해수준	-1.475	좌측 편향
		재해발생수준	-3.328	좌측 편향
11	민간위탁기술지도 (안전)	선정기준코드_물질관련	2.975	우측 편향
		선정기준명_기타	3.784	우측 편향
		선정기준명_신규사업장	9.729	우측 편향
		선정기준명_유해물질존재사업장	61.281	우측 편향
		선정기준명_재해발생사업장	5.647	우측 편향
		선정기준명_특별대책사업장	0.55	대칭 분포
		지원횟수	0.001	대칭 분포

번호	사업명	특성명	Skewness	불균형 여부
		전담	3.727	우측 편향
		겸직	2.47	우측 편향
		등급_중급	3.24	우측 편향
		등급_초급	4.815	우측 편향
		교육인원합	11.385	우측 편향
		자료제공_총합	54.87	우측 편향
		검사실시_종류수	5.927	우측 편향
		검사미실시_종류수	18.325	우측 편향
		검사비대상_종류수	5.301	우측 편향
		검사실시_합	8.543	우측 편향
		검사미실시_합	25.307	우측 편향
		검사비대상_합	15.429	우측 편향
		전체_개선	3.307	우측 편향
		전체_조치의뢰	24.401	우측 편향
		위탁기관평가	1.752	우측 편향
	민간위탁기술지도 (보건)	처리사유_개선완료	0	대칭 분포
		처리사유_조치의뢰	31.367	우측 편향
		선정기준코드_산재관련	2.152	우측 편향
		선정기준코드_실비관련	2.2	우측 편향
		선정기준코드_물질관련	2.975	우측 편향
		선정기준명_기타	3.784	우측 편향
		선정기준명_신규사업장	9.729	우측 편향
		선정기준명_유해물질존재사업장	61.281	우측 편향
		선정기준명_재해발생사업장	5.647	우측 편향
		선정기준명_특별대책사업장	0.55	대칭 분포
		지원횟수	0.001	대칭 분포
		전담	3.727	우측 편향
		겸직	2.47	우측 편향
		등급_중급	3.24	우측 편향
		등급_초급	4.815	우측 편향
		교육인원_총수	11.36	우측 편향
		자료제공_총합	54.87	우측 편향
		밀폐실태_밀폐공간장소_통의내부등	17.34	우측 편향
		밀폐실태_밀폐공간장소_정화조등	0.955	대칭 분포
		밀폐실태_밀폐공간장소_기타밀폐공간	13.101	우측 편향
		밀폐실태_밀폐공간장소_반응기등내부	10.531	우측 편향
		밀폐실태_밀폐공간장소_강재등시설	44.323	우측 편향
		밀폐실태_밀폐공간장소_불활성기체설비	13.933	우측 편향

번호	사업명	특성명	Skewness	불균형 여부
		밀폐실태-질식사고 인지도	2.971	우측 편향
		밀폐실태-질식사고 위험관리 인지도	3.275	우측 편향
		밀폐실태-질식사고 교육이수	1.944	우측 편향
		밀폐실태-가스농도측정기 보유	1.242	우측 편향
		급기팬 보유	1.487	우측 편향
		밀폐실태-위험도평가 총점	2.866	우측 편향
		개선_합	0.306	대칭 분포
		최종위험성평가수준평가결과	6.967	우측 편향
		실태평가결과_사업주의지	6.078	우측 편향
	민간위탁기술지도 (화학)	처리사유내용_종결	2.155	우측 편향
		처리사유내용_조치의뢰	36.007	우측 편향
		선정기준코드_산재관련	2.152	우측 편향
		선정기준코드_실비관련	2.2	우측 편향
		선정기준코드_물질관련	2.975	우측 편향
		선정기준명_기타	3.784	우측 편향
		선정기준명_신규사업장	9.729	우측 편향
		선정기준명_유해물질존재사업장	61.281	우측 편향
		선정기준명_재해발생사업장	5.647	우측 편향
		선정기준명_특별대책사업장	0.55	대칭 분포
		지원횟수	0.001	대칭 분포
		전담	3.727	우측 편향
		겸직	2.47	우측 편향
		등급_중급	3.24	우측 편향
		등급_초급	4.815	우측 편향
		교육인원_총수	11.36	우측 편향
		자료제공_총합	54.87	우측 편향
		기술지원_사고성재해예방	2.421	우측 편향
		기술지원_화학사고예방	4.604	우측 편향
		유해위험물질수	9.126	우측 편향
		화학설비건수	21.742	우측 편향
		위험기계기구_보유건수	10.406	우측 편향
		개선_총계	3.339	우측 편향
		조치의뢰_총계	82.579	우측 편향
		국소배기장치	20.081	우측 편향
		롤러기	0	대칭 분포
		리프트	38.689	우측 편향
		분쇄파쇄기	67.418	우측 편향
		사출성형기	128.508	우측 편향
		산업용로봇	105.854	우측 편향
		식품제조용설비	60.209	우측 편향
		원심기	0	대칭 분포
		전단기	140.04	우측 편향

번호	사업명	특성명	Skewness	불균형 여부
		지게차	8.72	우측 편향
		컨베이어	14.752	우측 편향
		크레인	12.895	우측 편향
		프레스	0	대칭 분포
		혼합기	22.514	우측 편향
		압력용기	12.72	우측 편향
		위험기계기구총합	12.394	우측 편향
		밀폐_정화조등	45.885	우측 편향
		밀폐_통의내부등	0	대칭 분포
		밀폐_불활성기체설비	0	대칭 분포
		밀폐_강재등시설	792.245	우측 편향
		밀폐_반응기등내부	61.513	우측 편향
		밀폐_콘크리트양생	0	대칭 분포
		밀폐_중독위험장소	792.245	우측 편향
		밀폐_기타밀폐공간	133.903	우측 편향
		위탁기관평가	1.752	우측 편향
		사업수행여부	7.858	우측 편향
	민간위탁기술지도(서비스)	처리사유_개선완료	0	대칭 분포
		처리사유_조치의뢰	31.367	우측 편향
		선정기준코드_산재관련	2.152	우측 편향
		선정기준코드_설비관련	2.837	우측 편향
		선정기준코드_물질관련	2.975	우측 편향
		선정기준명_기타	3.784	우측 편향
		선정기준명_신규사업장	9.729	우측 편향
		선정기준명_유해물질존재사업장	61.281	우측 편향
		선정기준명_재해발생사업장	5.647	우측 편향
		선정기준명_특별대책	2.604	우측 편향
		전담	3.727	우측 편향
		겸직	2.47	우측 편향
		등급_초급	4.815	우측 편향
		등급_중급	3.24	우측 편향
		교육인원_총수	11.36	우측 편향
		자료제공_총합	54.87	우측 편향
		지적건수_총합	2.974	우측 편향
		개선건수_총합	4.061	우측 편향
		조치건수_총합	4.78	우측 편향
		최종위험성평가수준평가결과	6.967	우측 편향
	민간위탁기술지도(사고사망예방 서비스)	지게차사용수량(대)	7.615	우측 편향
		지게차조종면허자격보유(명)	18.001	우측 편향
		지게차운행속도제한표지판설치여부	12.676	우측 편향
		지게차작업안전수칙제정및게시여부	13.195	우측 편향
		지게차사람공동사용출입구(개소)	9.198	우측 편향

번호	사업명	특성명	Skewness	불균형 여부
		지게차운행경사로수(개소)	12.024	우측 편향
		지게차모니터링CCTV(대)	25.049	우측 편향
		컨베이어설치수량(대)	37.049	우측 편향
		전용상하역장확보여부	11.621	우측 편향
		상하역공간조명등설치여부	7.363	우측 편향
		이동식사다리보유수량(대)	6.556	우측 편향
		사다리사용시안전모착용여부	5.281	우측 편향
		일평균차량출입(대)	3.677	우측 편향
		챠량운행모니터링CCTV(대)	32.651	우측 편향
		연평균사다리사용일수(일)	10.536	우측 편향
		사업수행여부	7.858	우측 편향
12	공단 교육 (안전보건교육)	교육분야코드_관리자	38.331	우측 편향
		교육분야코드_안전보건관계자	38.896	우측 편향
		교육분야코드_일반근로자	47.252	우측 편향
		교육분야코드_취약계층	46.313	우측 편향
	공단 교육 (인터넷교육센터)	교육대상_재분류_근로자	12.301	우측 편향
		교육대상_재분류_책임자	17.677	우측 편향
		교육대상_재분류_특수형태근로자	13.552	우측 편향
		수료여부_미수료	16.408	우측 편향
		수료비율	-0.697	대칭 분포
	공단 교육 (직무교육센터)	과정구분_new_온라인	2.241	우측 편향
		과정구분_new_집체	1.561	우측 편향
		교육대상_전문기관종사자	7.998	우측 편향
		수료여부_미수료	16.408	우측 편향
		안전보건관계자	2.196	우측 편향
13	유해위험방지계획서	대상업종_합계	5.379	우측 편향
		대상규모명_over_2000	0.87	대칭 분포
		대상규모명_under_2000	1.383	우측 편향
		대상규모명_under_500	0.584	대칭 분포
		사업구분_변경	1.326	우측 편향
		사업구분_설치	-0.823	대칭 분포
		사업구분_이전	4.231	우측 편향
		심사결과_반려	13.154	우측 편향
		심사결과_부적정	0	대칭 분포
		심사결과_적정	-0.4	대칭 분포
		심사결과_조건부적정	1.08	우측 편향
		고용부조치통보_미제출	0	대칭 분포
		고용부조치통보_지연	4.612	우측 편향
		최종확인회차	-0.418	대칭 분포
14	작업환경측정 (측정)	물질군명_노출기준제정물질	10.386	우측 편향
		물질군명_허가대상 유해물질	0	대칭 분포
		지상대상구분_대상	19.256	우측 편향
		초과율	1.752	우측 편향

번호	사업명	특성명	Skewness	불균형 여부
		취급인원	14.778	우측 편향
		물질군명_물리적인자_합	6.005	우측 편향
		물질군명_화학적인자_합	2.304	우측 편향
		물질군명_분진_합	6.986	우측 편향
	작업환경측정 (화학물질취급현황)	취급구분_사용	4.249	우측 편향
		취급구분_제조	35.311	우측 편향
		취급용도_기타	15.502	우측 편향
		취급용도_세척	144.633	우측 편향
		취급용도_실험	3.171	우측 편향
		취급용도_용접	22.343	우측 편향
		취급용도_원료	182.502	우측 편향
		취급물질군명_기타유해물질	13.773	우측 편향
		취급물질군명_노출기준제정물질	4.106	우측 편향
		취급물질군명_제조금지 유해물질	0	대칭 분포
		취급물질군명_허가대상 유해물질	280.096	우측 편향
		취급물질군명_물리적인자_통합	0	대칭 분포
		취급물질군명_분진인자_통합	9.574	우측 편향
		취급물질군명_화학적인자_통합	1.697	우측 편향
15	고용보험ERP근로자 수	고용상시인원수	15.352	우측 편향
		남성근로자수	13.683	우측 편향
		여성근로자수	19.016	우측 편향
		외국인근로자수	30.974	우측 편향
16	특수건강진단 (특검)	총수진자수(명)	7.39	우측 편향
		유해물질군명_노출기준제정물질	5.071	우측 편향
		유해물질군명_야간작업	5.774	우측 편향
		유해물질군명_제조금지 유해물질	0	대칭 분포
		유해물질군명_허가대상 유해물질	8.168	우측 편향
		A판정비율	-0.593	대칭 분포
		C1판정비율	8.758	우측 편향
		C2판정비율	3.365	우측 편향
		D1판정비율	12.441	우측 편향
		D2판정비율	8.554	우측 편향
		CN판정비율	1.587	우측 편향
		DN판정비율	2.148	우측 편향
		유해물질군명_물리적인자_합	3.87	우측 편향
		유해물질군명_화학적인자_합	4.588	우측 편향
		유해물질군명_분진_합	3.185	우측 편향
		사업수행여부	7.858	우측 편향
	특수건강진단 (사업장별 검진내역)	검진종목별 수진자수(일반)	16.534	우측 편향
		검진종목별 수진자수(특수)	13.623	우측 편향

번호	사업명	특성명	Skewness	불균형 여부
		검진종목별 수진자수(배치전)	10.015	우측 편향
		검진종목별 수진자수(수시)	560.2	우측 편향
		검진종목별 수진자수(임시)	792.245	우측 편향
		검진종목별 수진자수(수첩)	792.245	우측 편향
		사업수행여부	7.858	우측 편향
17	산업안전보건실태조사	종사자수	2.772	우측 편향
		교대 근무제 시행	0.402	대칭 분포
		노동조함 유무(예, 아니오)	-1.027	좌측 편향
		현장 산업안전보건위원회 구성 및 운영	0.681	대칭 분포
		1년간 유해, 위험 요인에 대한 위험성 평가 및 필요한 조치 문서작업	-1.758	좌측 편향
		2020년 일반건강진단 결과 사후관리조치대상자 조치 이행 여부 확인	0	대칭 분포
		2020년 특수건강진단 결과 사후관리조치대상자 조치 이행 여부 확인	0	대칭 분포
		2020년 작업 환경 측정 결과를 바탕으로 유해인자의 노출량을 최소화하기 위한 구체적 노력	-18.93	좌측 편향
		사업체 내에 상주하며 연간 계약을 하는 협력 업체	2.709	우측 편향
		사내 상주하며 연간 계약 협력 업체 수	6.702	우측 편향
		사내 상주하며 연간 계약 협력 업체의 총 근로자 수	8.459	우측 편향
		거래하는 원청 회사 여부, 사업체가 원청 회사의 사업체 내에 위치하는지 여부	1.705	우측 편향
		교대근무_종사자_비율	0.837	대칭 분포
		야간근무_종사자_비율	0.847	대칭 분포
		작업 환경 관련 위험 요인 평균	0.334	대칭 분포
		신체적 부담 관련 위험 요인 평균	-0.128	대칭 분포
		생/화학 물질 관련 위험 요인 평균	0.91	대칭 분포
		기계, 전기, 기타 위험 요인 평균	0.11	대칭 분포
		스트레스 심각도 평균	0.173	대칭 분포
		스트레스 관리 노력 정도 평균	-0.5	대칭 분포
		경영진 안전보건의지 평균	-1.494	좌측 편향
		사업장내 안전문화 평균	-0.72	대칭 분포
		근로자 안전보건의지 평균	-0.504	대칭 분포
18	산업재해조사표	재해자동종경력년수_1~3	110.432	우측 편향
		재해자동종경력년수_3~5	78.749	우측 편향

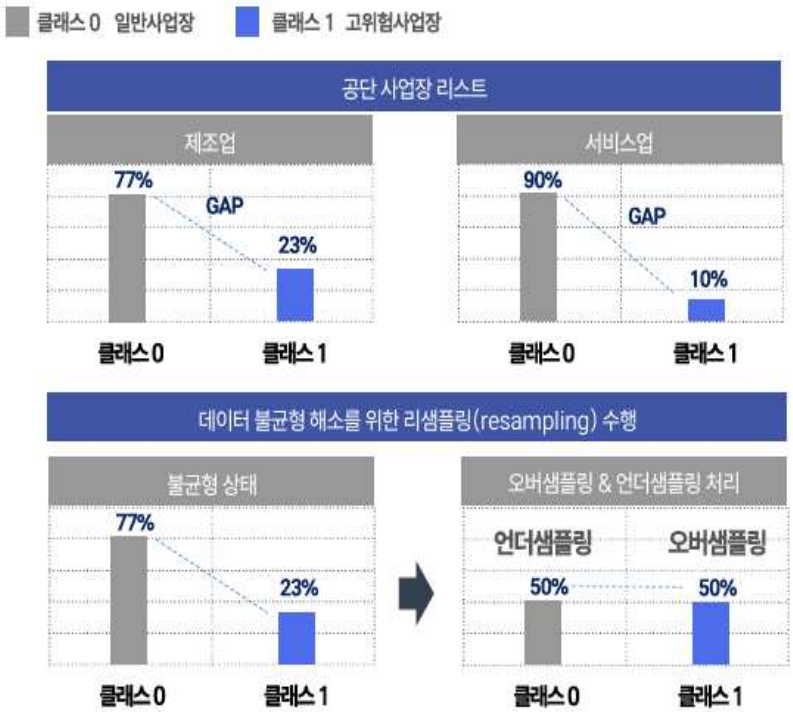
번호	사업명	특성명	Skewness	불균형 여부
		재해자동종경력년수_5~10	72.599	우측 편향
		재해자동종경력년수_10~20	112.122	우측 편향
		재해자동종경력년수_20~	149.198	우측 편향
		상해부위_머리	160.578	우측 편향
		상해부위_몸통	125.054	우측 편향
		상해부위_팔	208.11	우측 편향
		상해부위_전신	372.51	우측 편향
		상해부위_다발성	90.479	우측 편향
		상해부위_기타	52.094	우측 편향
		재해발생요일_주말	140.373	우측 편향
19	작업환경실태조사 (일반현황)	전기계약용량	-1.722	좌측 편향
		야간작업유무	3.704	우측 편향
		정비_보수여부	3.645	우측 편향
		안전관리자	3.046	우측 편향
		안전관리자_유형	-2.115	좌측 편향
		보건관리자	2.908	우측 편향
		보건관리자_유형	-1.994	좌측 편향
		안전보건담당자	3.131	우측 편향
		안전보건담당자수	16.779	우측 편향
		원청_하청여부	-5.464	좌측 편향
		하청사업장수	52.687	우측 편향
		하청근로자수	28.806	우측 편향
		근골격계부담작업대상여부	-2.738	좌측 편향
		유해요인조사실시여부	0.634	대칭 분포
		복지시설_개수	0.526	대칭 분포
	작업환경실태조사 (화학물질취급)	취급/생산	0	대칭 분포
		허용대상물질여부	0	대칭 분포
		허용기준물질여부	0	대칭 분포
		관리대상물질여부	0.709	대칭 분포
		안전검사물질여부	13.909	우측 편향
		안전관리물질여부	1.808	우측 편향
		기타물질여부	-0.171	대칭 분포
		특검대상물질여부	0.24	대칭 분포
		측정대상물질여부	-0.224	대칭 분포
		PSM대상물질여부	5.525	우측 편향
		건강관리수첩대상물질여부	0	대칭 분포
		사고대상물질여부	1.662	우측 편향
		금지대상물질여부	0	대칭 분포
		근로자_월_취급시간	2.721	우측 편향
	작업환경실태조사 (기계기구설비현황)	제조_보유수량총개수	0	대칭 분포
		제조_총종류수량	0	대칭 분포
		비제조_보유수량총개수	256.842	우측 편향
		비제조_총종류수량	112.558	우측 편향
	작업환경실태조사	소음발생공정수_총합	299.464	우측 편향

번호	사업명	특성명	Skewness	불균형 여부
	(작업환경)	밀폐공간수_총합	791.469	우측 편향
		작업환경구분_고열_한랭_다습_및_방사선_취급_작업_총합	297.88	우측 편향
		작업환경구분_밀폐공간(산소결핍_위험장소)_현황_총합	457.401	우측 편향
		작업환경구분_분진_흄_발생작업_총합	227.576	우측 편향
		작업환경구분_사내도급작업_총합	0	대칭 분포
		작업환경구분_소음작업_총합	242.569	우측 편향
		작업환경구분_제조나노물질의_제조_및_취급_작업_총합	0	대칭 분포
		작업환경구분_진동발생작업_총합	457.401	우측 편향
		고열_한랭_다습_및_방사선_취급_작업_종사근로자수합	710.32	우측 편향
		밀폐공간(산소결핍_위험장소)_현황_종사근로자수합	0	대칭 분포
		분진_흄_발생작업_종사근로자수합	441.365	우측 편향
		사내도급작업_종사근로자수합	0	대칭 분포
		소음작업_종사근로자수합	377.494	우측 편향
		제조나노물질의_제조_및_취급_작업_종사근로자수합	0	대칭 분포
		진동발생작업_종사근로자수합	750.6	우측 편향
21	공공기관등급 데이터	등급	66.646	우측 편향
23	사업장수준조사평가+재해율	사업주 · 관리자 마인드	5.017	우측 편향
		근로자 안전보건 행동 수준	4.98	우측 편향
		작업장 및 근로환경 수준	4.98	우측 편향
		3년 평균 재해율	15.601	우측 편향

※ 불균형 여부의 '대칭 분포'의 기준은 Skewness값이 -1~1인 경우임

(2) 데이터 불균형(Data Imbalance)

○ 데이터 불균형은 클래스 간 데이터 수의 차이로 발생하는 문제로, 이는 모델 성능에 큰 영향을 미칠 수 있음. 특히, 고위험 사업장에 비해 저위험 사업장의 데이터가 많아, 모델이 다수 클래스(저위험 사업장)에 대해 더 좋은 성능을 보이는 경향이 있음.



[그림 3-2] 데이터 불균형 분석 결과

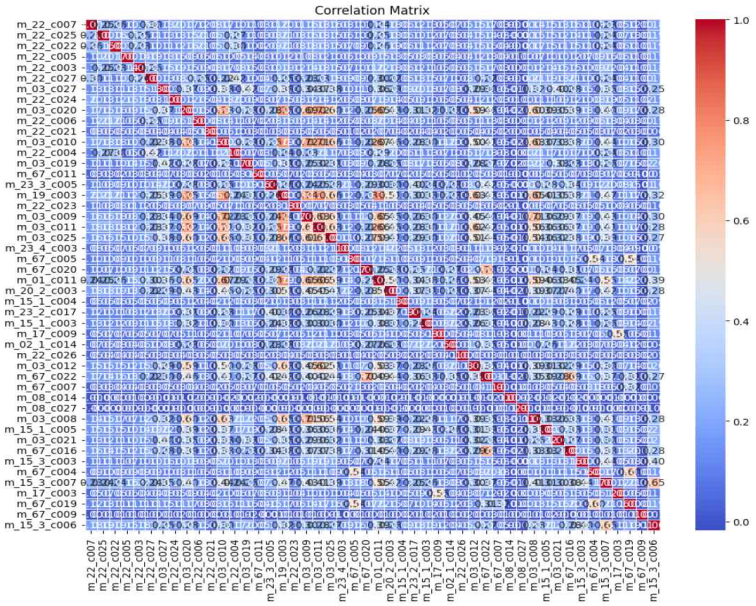
1. 데이터 불균형 문제

- 안전 사업장 데이터의 비율이 위험 사업장 데이터보다 많으며, 이로 인해 모델이 소수 클래스를 정확하게 예측하는 데 어려움이 있음.

2. 해결 방안

- 서비스업에서 특히 데이터 불균형 비율이 커서, 단순한 오버샘플링보다는 오버샘플링과 언더샘플링을 결합한 기법(예:SMOTE-Tomek)을 사용하여 다수 클래스의 불필요한 데이터를 제거하고, 균형 잡힌 데이터 셋을 구성하는 방법을 적용함.

(3) 특성 간 강한 상관관계



[그림 3-3] 다중공선성 분석 결과

○ 다중공선성(Multicollinearity)은 데이터의 여러 특성들이 서로 강한 상관성을 가질 때 발생하며, 이는 모델 학습시 과적합을 유발할 수 있음. 공선성 문제는 모델이 특정 변수에 지나치게 의존하게 되어, 실제 예측에서 불안정한 결과를 초래할 위험이 있음. 이를 해결하기 위해 다중공선성 분석을 통해 불필요한 특성을 제거하는 과정이 필요함.

○ 또한, 학습에 불필요한 특성수를 줄여 모델의 성능 변화에 영향을 미치는 정도를 함께 확인하였음.

1. 근로손실일수와 특성간 상관분석

- 연속값을 가지는 근로손실일수와 각 특성 간 상관계수를 구한 후, 상관계수 0.3 이상 또는 -0.3 이하의 특성만을 모델 학습에 사용하였음. 상관계수는 -1에서 1사이틔 값을 가지며, 이 값이 0에 가까울수록 상관

관계가 약하고, 1에 가까울수록 강한 양의 상관관계를, -1에 가까울수록 강한 음의 상관관계를 나타냄.

2. 다중 공선성 분석

- 상관분석을 통해 0.3 미만의 상관계수를 가진 특성을 제외한 후에도, 여전히 공선성이 높은 일부 특성들이 있어서, 이러한 특성들은 모델의 학습 과정에서 중복된 정보를 제공할 수 있으므로 모델의 성능과 해석력에 부정적인 영향을 미칠 수 있음. 따라서, 공선성이 높은 특성들을 추가로 제거하거나 축소하여, 모델이 보다 일관성 있는 학습을 할 수 있도록 처리함.

○ 〈표 3-16〉는 제조업과 서비스업의 특성별로 근로손실일수와 상관분석을 통해 상관계수 0.3 이상 또는 -0.3 미만의 값을 가진 특성들의 목록을 나타냄.

〈표 3-16〉 업종별 근로손실일수와 특성간 상관분석 결과 상위 50개

제조업			서비스업		
번호	특성명	상관계수	번호	특성명	상관계수
1	재해자동종경력년수_5~10	0.968	1	상해부위_몸통	0.228
2	상해부위_몸통	0.959	2	재해자동종경력년수_5~10	0.224
3	상해부위_다리	0.930	3	재해자동종경력년수_3~5	0.220
4	재해자동종경력년수_20~	0.918	4	상해부위_머리	0.218
5	재해자동종경력년수_10~20	0.909	5	재해발생요일_주말	0.215
6	상해부위_팔	0.901	6	재해자동종경력년수_20~	0.214
7	직종_대분류_예술·디자인·방송·스포츠직_합계	0.882	7	상해부위_팔	0.213
8	상해부위_머리	0.874	8	재해자동종경력년수_1~3	0.209
9	직종_대분류_교육·법률·사회복지·경찰·소방직및군인_합계	0.868	9	재해자동종경력년수_10~20	0.208397
10	재해자동종경력년수_3~5	0.863	10	상해부위_전신	0.170334
11	상해부위_기타	0.857	11	상해부위_다발성	0.158292

제조업			서비스업		
번호	특성명	상관계수	번호	특성명	상관계수
12	연령대_40대_합계	0.848	12	선임자총수	0.145417
13	하청사업장수	0.841	13	안전관리자	0.135299
14	하청근로자수	0.831	14	보건관리자	0.135114
15	재해자동종경력년수_1~3	0.796	15	보건담당자	0.134658
16	직종_대분류_건설·채굴직_합계	0.766	16	선임자종류수	0.133823
17	국소배기장치	0.748	17	전담유무	0.128502
18	성별_남_합계	0.746	18	상해부위_기타	0.125993
19	남성근로자수	0.745	19	소업종명	0.104759
20	기계설비_비제조_보유갯수	0.738	20	사업수행여부	0.102226
21	피보험자_합계	0.727	21	사업수행여부	0.101872
22	고용상시인원수	0.719	22	규모1	0.096148
23	상해부위_다발성	0.712	23	사업수행여부	0.08483
24	연령대_30대_합계	0.694	24	안전보건관리책임자	0.079323
25	연령대_50대_합계	0.643	25	직종_대분류_건설·채굴직_합계	0.077204
26	직종_대분류_연구직및공학기술직_합계	0.642	26	관리자	0.077159
27	소음발생공정수	0.637	27	근속기간_년수_평균	0.062671
28	컨베이어종류_롤러	0.635	28	명예산업안전감독관	0.061971
29	크레인	0.634	29	외국인근로자수	0.061663
30	근로자수	0.633	30	직종_대분류_농림어업직_합계	0.060986
31	검진종목별 수진자수(일반)	0.623	31	성별_남_합계	0.046329
32	검진종목별_총합	0.617	32	검진종목별 수진자수(일반)	0.045631
33	검진종목별 수진자수(특수)	0.610	33	고용상시인원수	0.043999
34	교육분야코드_안전보건관계자	0.586	34	연령대_50대_합계	0.043816
35	근로자_월_취급시간	0.556	35	검진종목별 수진자수(특수)	0.041
36	교육분야코드_관리자	0.540	36	직종_대분류_예술·디자인·방송·스포츠직_합계	0.039
37	사업구분_설치	0.515	37	직종_대분류_연구직및공학기술직_합계	0.037
38	점검차수	0.512	38	연령대_40대_합계	0.037
39	상해부위_전신	0.507	39	융자지원여부	0.037
40	연령대_60대이상_합계	0.492	40	등급	0.037
41	심사결과_적합	0.488	41	피보험자_합계	0.036

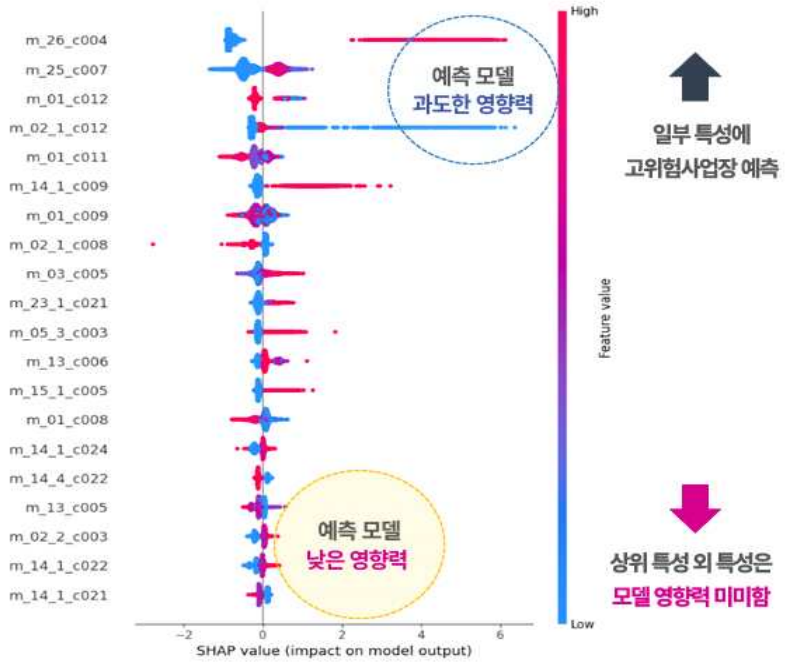
제조업			서비스업		
번호	특성명	상관계수	번호	특성명	상관계수
42	컨베이어종류_버킷	0.480	42	컨베이어(구간내컨베이어종류)_롤러	0.036
43	위험물제조소_종류	0.476	43	선정기준명_기타	0.035
44	이동탱크수	0.448	44	클린지원여부	0.034
45	연령대_10대_20대_합계	0.419	45	산업보건의	0.032
46	교육분야코드_일반근로자	0.393	46	남성근로자수	0.030
47	직종_대분류_농림어업직_합계	0.376	47	교부금액	0.028
48	압력용기	0.359	48	연령대_60대이상_합계	0.027
49	과정구분_온라인	0.353	49	직종_대분류_설치·정비·생산직_합계	0.027
50	컨베이어종류_체인	0.352	50	성별_여_합계	0.027

○ 제조업의 경우, 근로손실일수에 직접적으로 상관성이 있는 특성들이 있으나, 서비스업의 경우는 특성들이 전반적으로 제조업에 비해 상관계수가 낮은 것을 확인하였음.

○ 이러한 특성들은 모델이 각 업종에서 영향을 미치는 영향이 상이함을 시사하며, 제조업의 경우는 강한 상관관계를 보이는 특성을 중심으로 학습을 진행이 필요할 수 있고, 서비스업의 경우는 상관관계가 상대적으로 낮기 때문에 보다 다양한 특성을 고려한 분석이 필요함.

○ 이러한 차이를 반영하여, 모델의 성능이 기대에 미치지 못하는 경우, 맞춤형 모델링 전략을 채택할 필요가 있으며, 제조업에서는 상관계수가 높은 특성을 활용하여 더 간결하고 해석 가능한 모델을 구축할 수있고, 서비스업은 다양한 특성 조합을 통해 모델의 성능을 향상시키는 방법이 고려되어야 함.

(4) 모델의 특성 의존성 편향



[그림 3-4] SHAP을 활용한 특성 기여도 분석 결과

○ 모델이 특정 특성에 과도하게 의존하는 경우, 모델 성능이 왜곡될 수 있으며, 이를 확인하기 위해 XAI(Explainable AI) 기법을 적용하여 모델의 특성 의존성 편향을 분석하였음.

1. XAI 기법을 통한 특성 중요도 분석

- SHAP(Shapley Additive Explanations)는 게임 이론에서 유래한 기법으로, 각 특성이 모델의 예측에 얼마나 기여하는지 정량화 할 수 있음. 이를 통해 모델이 특정 예측을 내린 이유를 설명할 수 있으며, 로컬과 글로벌 해석을 동시에 제공함.

- SHAP Value는 특성별로 상대적 기여도를 나타내며, 이를 통해 모델이 특정 특성에 의존하는 정도를 확인할 수 있음. 일반적으로,

Mean(|SHAP|) 값이 클수록 해당 특성이 모델 예측에 더 중요한 역할을 하고 있다고 판단됨.

2. SHAP Value의 활용

- XGBoost에서 제공되는 Feature Importance는 모델이 학습 과정에서 각 특성을 얼마나 자주 분할 기준으로 사용했는지를 빠르게 확인할 수 있는 도구임. 하지만 SHAP Value는 각 예측에 대해 보다 구체적으로 특성 기여도를 계산하여, 특정 예측 결과가 왜 그런지에 대해 깊은 이해를 제공할 수 있음.

- SHAP 분석을 통해 0에 가까운 값을 가지는 특성은 상대적으로 모델의 예측에 중요하지 않다고 해석할 수 있으며, 값이 크게 분포하는 특성은 중요한 기여를 하고 있다고 판단할 수 있음.

3. 모델의 편향성 해소 방안

- SHAP 분석 결과를 바탕으로, 모델이 특정 특성에 과도하게 의존하고 있음을 발견 시, 모델이 균형 잡힌 학습을 할 수 있도록 특성의 가중치를 낮추거나 제거, 차원축소 등을 수행할 수 있음.

〈표 3-17〉 업종별 특성별 기여도 분석 결과 상위 50개

제조업			서비스업		
번호	특성명	기여도	번호	특성명	기여도
1	재해율3년평균	1.079	1	3년 평균 재해율	1.106
2	안전관리수준평가사업장위험도_현장위험관리수준	0.539	2	근속기간_년수_평균	0.625
3	규모1	0.463	3	대상규모명_under_500	0.461
4	점검결과조치_사업장자체개선 후 종결/점검 종결	0.271	4	대상규모명_over_2000	0.418
5	위탁기관평가_점수	0.236	5	심사결과_적정	0.415
6	근로자수	0.223	6	자료제공_총합	0.402
7	충돌방지장치점수	0.172	7	자료제공_총합	0.389
8	위험기계기구_보유건수	0.163	8	심사결과_조건부적정	0.349

제조업			서비스업		
번호	특성명	기여도	번호	특성명	기여도
9	재해발생수준	0.149	9	사업구분_설치	0.345
10	선정기준명_신규사업장	0.146	10	대상규모명_under_2000	0.342
11	교육분야코드_일반근로자	0.143	11	수료여부_미수료	0.316
12	기술지원_화학사고예방	0.142	12	밀폐실태-위험도평가 총점	0.31
13	점검차수	0.141	13	수료비율	0.272
14	소음발생공정수	0.14	14	사업구분_변경	0.269
15	교부금액	0.137	15	사업주의관심도	0.255
16	검사실시_종류수	0.133	16	이동탱크여부	0.23
17	안전보건관리및개선노력	0.129	17	운전자격점수	0.229
18	소업종명	0.128	18	구성원의참여및이해수준	0.221
19	선정기준명_재해발생사업장	0.128	19	경영진 안전보건의지 평균	0.221
20	근속기간_년수_평균	0.126	20	이동탱크수	0.219
21	보유건수	0.123	21	교육분야코드_일반근로자	0.192
22	사고유발요인_개수	0.12	22	고용상시인원수	0.19
23	위험기계기구종합	0.119	23	관리자이해점수	0.184
24	검사실시_합	0.118	24	위탁기관평가	0.167
25	근로자_월_취급시간	0.114	25	급기팬 보유	0.165
26	자료제공_총합	0.114	26	만나이_평균	0.148
27	위험성평가실행수준	0.114	27	선정기준코드_산재관련	0.147
28	복지시설_개수	0.113	28	교육분야코드_관리자	0.144
29	외국인근로자수	0.11	29	선정기준명_기타	0.14
30	설치기간_5년이상	0.11	30	안전띠점수	0.139
31	작업환경_소음작업	0.108	31	소업종명	0.139
32	점검차수	0.107	32	수료여부_미수료	0.138
33	검사비대상_합	0.106	33	교육대상_재분류_특수형태근로자	0.128
34	중업종명	0.104	34	교육대상_재분류_책임자	0.119
35	검사비대상_종류수	0.103	35	규모1	0.114
36	전담	0.1	36	겸직	0.11
37	기술지원_사고성재해예방	0.099	37	충돌방지장치점수	0.109
38	등급_초급	0.096	38	상해부위_팔	0.094
39	탱크여부	0.092	39	기타물질여부	0.094
40	등급_중급	0.09	40	여성근로자수	0.092
41	설치기간_5년미만	0.088	41	교대 근무제 시행	0.092
42	안전띠점수	0.088	42	등급_중급	0.091
43	기계설비_비제조_보유갯수	0.085	43	유해물질군명_야간작업	0.09

제조업			서비스업		
번호	특성명	기여도	번호	특성명	기여도
44	행정구역	0.085	44	생/화학 물질 관련 위험 요인 평균	0.08
45	점검차수	0.084	45	성별_여_합계	0.078
46	안전보건수준평가종합	0.081	46	위험성평가실행수준	0.076
47	탱크총합	0.08	47	근로자 안전보건 행동 수준	0.072
48	겸직	0.074	48	재해자동종경력년수_1~3	0.072
49	관리자이해점수	0.073	49	관리대상물질여부	0.07
50	노동지청	0.069	50	피보험자_합계	0.067

○ 서로 다른 사업을 통해 수집된 데이터 중 서로 유사한 특성이 존재하며, 모델에 유사한 영향력을 미침을 확인하였음.

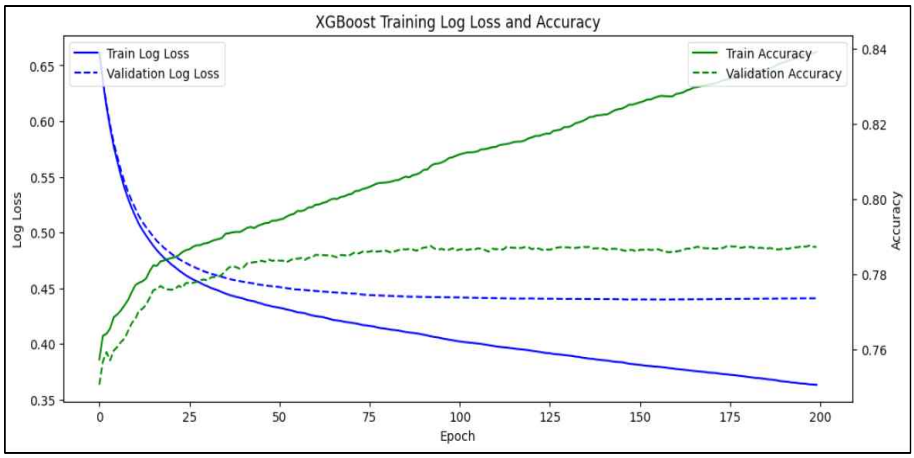
○ 400여개의 특성 중, 모델 학습에 크게 영향을 미치지 못하는 특성들이 다수 존재하며, 이러한 특성들은 모델에 의해 자주 사용하지는 않더라도, 모델의 성능 향상을 위해 사전에 분석하여 제거하는 것이 필요함.

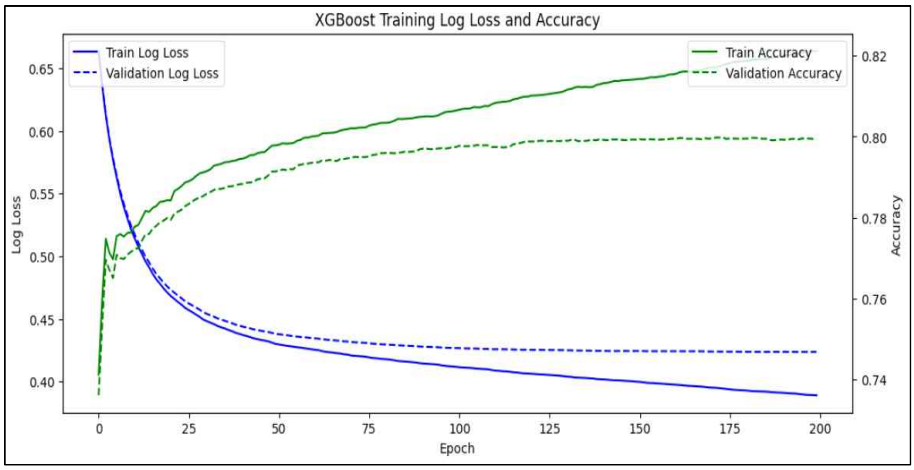
○ 과도하게 영향력이 높은 특성은 모델 학습 시 제거하거나 두 개 이상의 특성이 매우 유사할 경우 하나의 특성만 남기고 나머지를 제거하는 등의 처리를 한 후 학습에 반영하였음.

3) 다양한 데이터 전처리 및 모델 적용

○ XGBoost 외에도 각각의 고유한 강점을 가진 CatBoost, LightGBM(LGBM), 그리고 딥러닝 모델을 추가로 적용하여 성능을 비교하였음. 데이터 전처리 기법과 하이퍼파라미터 최적화를 통해 각 모델의 성능을 극대화 하였음.

(1) XGBoost 모델 하이퍼파라미터 최적화 후 학습 결과

구분	내용
작업명	제조업_XGBoost_RandomSearch최적화
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.78 2. 정밀도(Precision): - Label 0 (0.77): 모델이 안전 사업장으로 예측한 것 중 77%가 실제 안전 사업장임을 의미함 - Label 1 (0.82): 모델이 위험 사업장으로 예측한 것 중 82%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.83): 실제 안전 사업장 중 83%가 정확하게 예측함 - Label 1 (0.74): 실제 위험 사업장 중 74%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.80 - Label 1 : 0.78

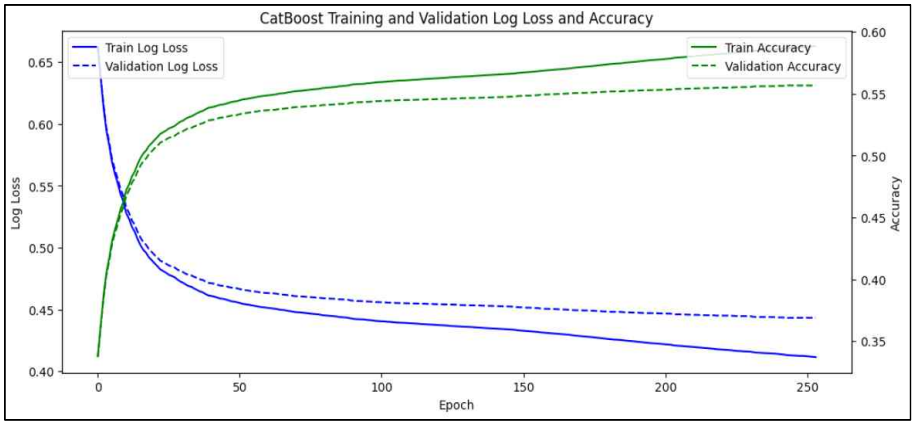
구분	내용
작업명	서비스업_XGBoost_RandomSearch최적화
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.80 2. 정밀도(Precision): - Label 0 (0.77): 모델이 안전 사업장으로 예측한 것 중 77%가 실제 안전 사업장임을 의미함 - Label 1 (0.84): 모델이 위험 사업장으로 예측한 것 중 84%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.86): 실제 안전 사업장 중 86%가 정확하게 예측함 - Label 1 (0.74): 실제 위험 사업장 중 74%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.81 - Label 1 : 0.79

○ 최적화 전과 비교하여 제조업과 서비스업 모두 정확도는 다소 낮아졌지만, Label 0(안전사업장)에 과도하게 집중되던 현상이 사라지고, Label0과 Label1(위험사업장)에 대한 예측 성능 균형이 이루어짐.

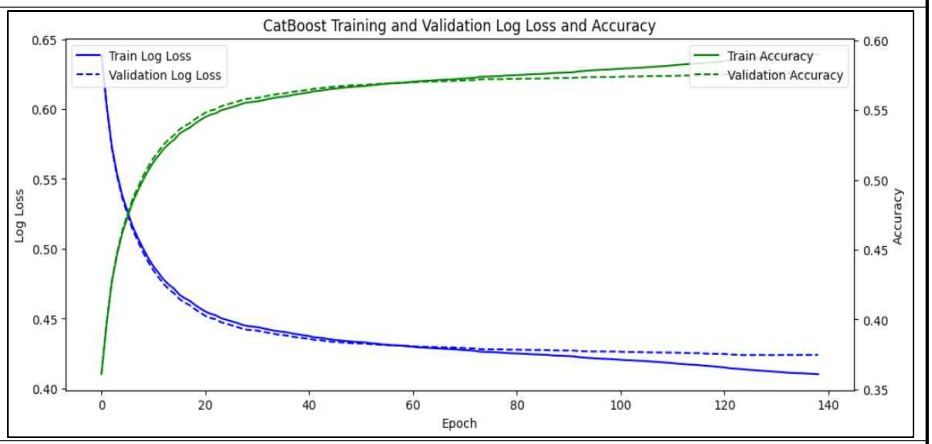
(2) CatBoost 모델 하이퍼파라미터 최적화 후 학습 결과

○ CatBoost(Categorical Boosting)는 Yandex에서 개발된 Gradient Boosting 기반의 머신러닝 알고리즘으로, 특히 범주형 데이터 처리에 강점을 지니고 있음. CatBoost는 자동으로 범주형 데이터를 인코딩하고, 일반적인 부스팅 모델에서 발생하는 과적합 문제를 방지할 수 있는 새로운 방식을 도입하였으며 다음과 같은 특징을 가지고 있음.

1. **범주형 데이터 처리**: CatBoost는 범주형 데이터를 효율적으로 처리할 수 있도록 순서 코딩(Order-based encoding) 기법을 사용하여, One-Hot Encoding과 같은 전통적인 방식에 비해 더 빠르고 메모리 효율적임.
2. **과적합 방지**: CatBoost는 학습 과정에서 랜덤화 기법을 적용하여 과적합을 방지하며, 데이터의 순서와 관계없이 일관된 성능을 유지할 수 있음.
3. **빠른 학습**: GPU 가속을 지원하여 대규모 데이터셋에서도 빠른 학습 가능함.

구분	내용
작업명	제조업_CatBoost_RandomSearch최적화
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.78 2. 정밀도(Precision): - Label 0 (0.77): 모델이 안전 사업장으로 예측한 것 중 77%가 실제 안전 사업장임을 의미함 - Label 1 (0.82): 모델이 위험 사업장으로 예측한 것 중 82%가 실제 위험 사업장임을

	의미함 3. 재현율(Recall) - Label 0 (0.83): 실제 안전 사업장 중 83%가 정확하게 예측함 - Label 1 (0.75): 실제 위험 사업장 중 75%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.80 - Label 1 : 0.78

구분	내용
작업명	서비스업_CatBoost_RandomSearch최적화
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.79 2. 정밀도(Precision): - Label 0 (0.76): 모델이 안전 사업장으로 예측한 것 중 76%가 실제 안전 사업장임을 의미함 - Label 1 (0.84): 모델이 위험 사업장으로 예측한 것 중 84%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.86): 실제 안전 사업장 중 86%가 정확하게 예측함 - Label 1 (0.74): 실제 위험 사업장 중 74%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.81 - Label 1 : 0.78

○ 학습 그래프의 곡선은 XGBoost보다 안정적으로 학습하는것처럼 보이나, 성능은 XGBoost와 큰 차이를 보이지 않았음.

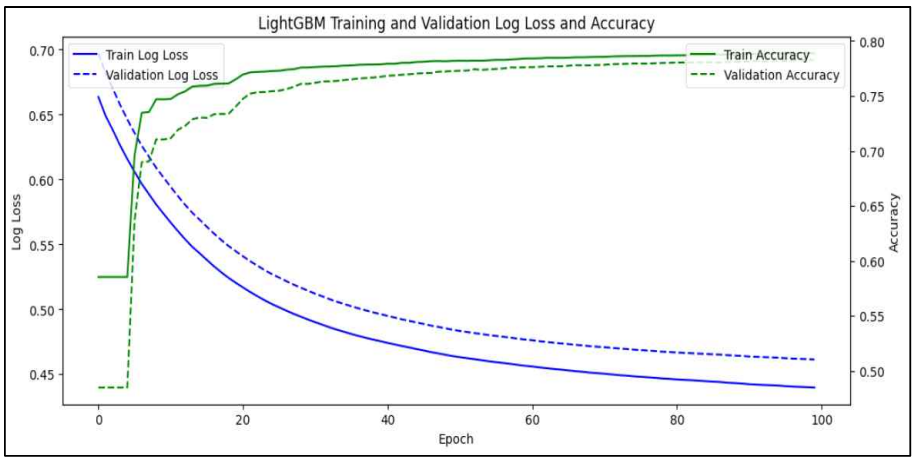
(3) LightGBM 모델 하이퍼파라미터 최적화 후 학습 결과

○ LightGBM(LGBM)은 Microsoft에서 개발한 Gradient Boosting Decision Tree(GBDT) 기반의 알고리즘으로, 특히 대규모 데이터와 고차원 데이터를 처리하는데 효율적임. Histogram-based 방식을 통해 메모리 사용량을 줄이고, 매우 빠른 학습 속도를 지원함.

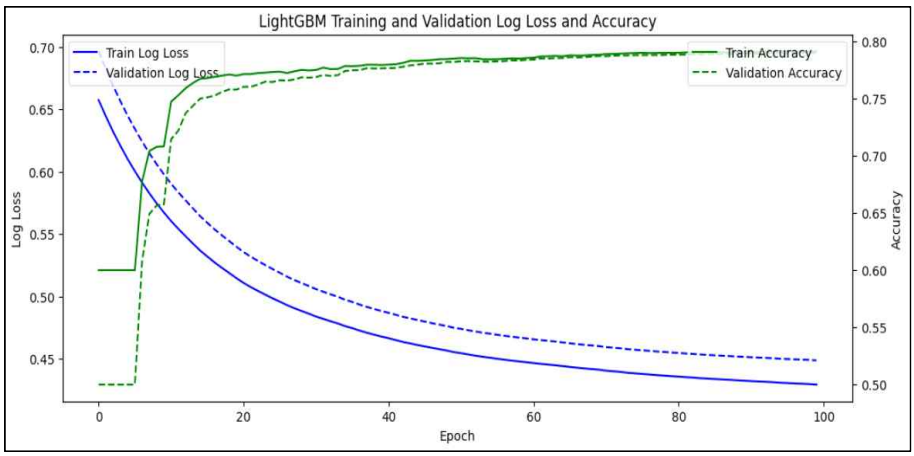
1. **Leaf-wise Tree Growth**: LGBM은 Depth-wise 방식이 아닌 Leaf-wise로 트리를 확장하여 오차 감소에 집중함. 이는 모델의 성능을 빠르게 향상시킬 수 있지만, 과적합 위험이 있음.

2. **빠른 학습 속도**: 데이터를 구간화하여 처리하는 방식으로 메모리 효율을 극대화하고, 대규모 데이터에서도 빠른 학습을 수행할 수 있음.

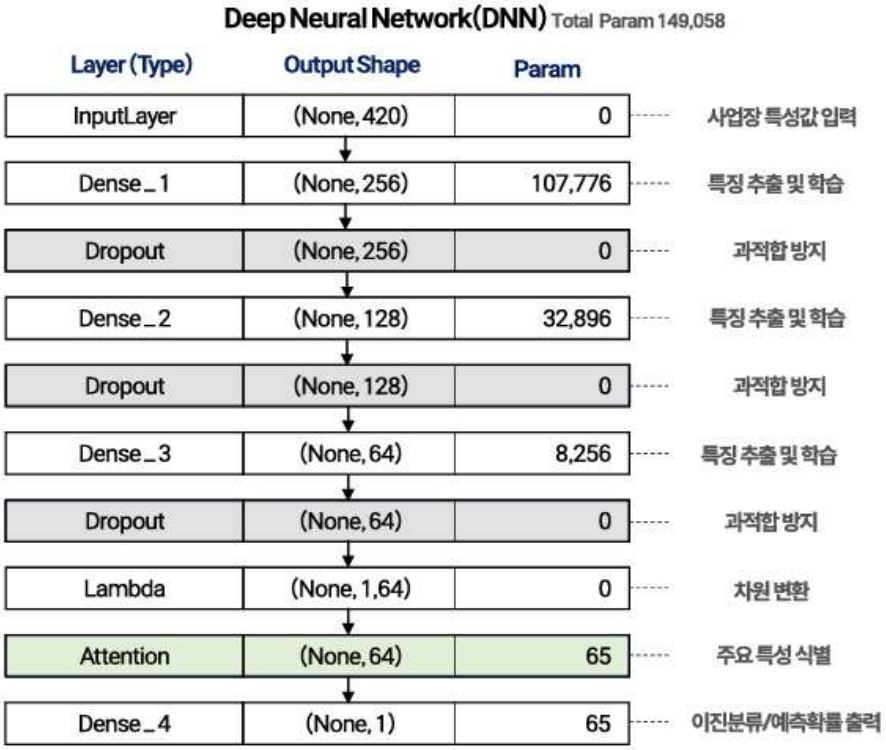
3. **GPU 지원**: GPU를 활용한 학습을 지원하여 더 빠른 성능을 낼 수 있으며, 대규모 데이터셋에 적합함.

구분	내용
작업명	제조업_LGBM_RandomSearch최적화
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.77 2. 정밀도(Precision): - Label 0 (0.80): 모델이 안전 사업장으로 예측한 것 중 00%가 실제 안전 사업장임을

	의미함 - Label 1 (0.76): 모델이 위험 사업장으로 예측한 것 중 00%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.76): 실제 안전 사업장 중 76%가 정확하게 예측함 - Label 1 (0.80): 실제 위험 사업장 중 80%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.78 - Label 1 : 0.78

구분	내용
작업명	서비스업_LGBM_RandomSearch최적화
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.79 2. 정밀도(Precision): - Label 0 (0.80): 모델이 안전 사업장으로 예측한 것 중 00%가 실제 안전 사업장임을 의미함 - Label 1 (0.78): 모델이 위험 사업장으로 예측한 것 중 00%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.78): 실제 안전 사업장 중 00%가 정확하게 예측함 - Label 1 (0.80): 실제 위험 사업장 중 00%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.79 - Label 1 : 0.79

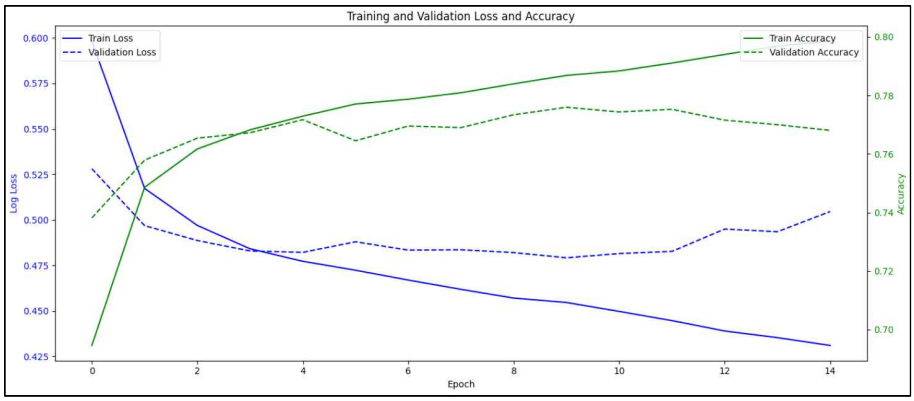
(4) Deep Neural Network 모델 학습 결과



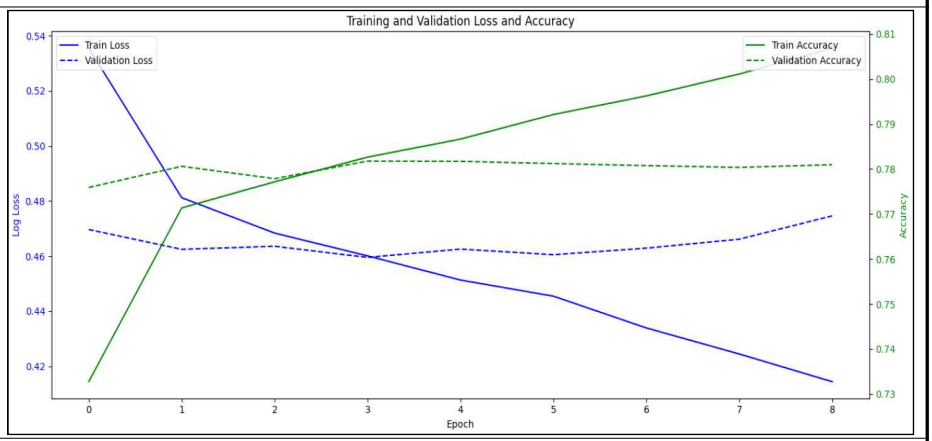
[그림 3-5] 딥러닝 모델 구성

○ 딥러닝 모델은 인공 신경망(Artificial Neural Networks, ANN) 기반의 모델로, 특히 비선형성이 강한 데이터와 복잡한 상호작용이 있는 데이터에서 우수한 성능을 보임.

○ 본 연구에서는 149,058개의 파라미터를 구성하였으며, 과적합 방지를 위해 Dropout층과 Attention 층을 추가하였음.

구분	내용
작업명	제조업_DNN_RandomSearch최적화
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.77 2. 정밀도(Precision): - Label 0 (0.75): 모델이 안전 사업장으로 예측한 것 중 75%가 실제 안전 사업장임을 의미함 - Label 1 (0.80): 모델이 위험 사업장으로 예측한 것 중 80%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.82): 실제 안전 사업장 중 82%가 정확하게 예측함 - Label 1 (0.73): 실제 위험 사업장 중 73%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.79 - Label 1 : 0.76

○ 학습 그래프가 수렴하지 않고, 타 모델 대비 상대적으로 성능 역시 떨어지는 것을 확인하였음.

구분	내용
작업명	서비스업_DNN_RandomSearch최적화
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.77 2. 정밀도(Precision): - Label 0 (0.74): 모델이 안전 사업장으로 예측한 것 중 74%가 실제 안전 사업장임을 의미함 - Label 1 (0.83): 모델이 위험 사업장으로 예측한 것 중 83%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.85): 실제 안전 사업장 중 85%가 정확하게 예측함 - Label 1 (0.70): 실제 위험 사업장 중 70%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.79 - Label 1 : 0.75

4) 기존 고위험사업장 선정 모델 분석 결과 및 추가 개선 방안

○ 하이퍼파라미터는 학습률, 배치 크기, 정규화 계수 등의 요소로 구성되어 있으며, 이들은 서로 복잡하게 상호작용을 함. 사람이 이를 일일이 조정하는 것은 상호작용을 정확히 파악하고, 최적의 조합을 찾는데 한계가 있음. 특히, 데이터 조건이 조금만 바뀌거나 전처리 방법이 달라져도 하이퍼파라미터 튜닝이 다시 필요해지는 문제가 발생함.

○ 이러한 문제를 해결하기 위해 최적화 알고리즘을 사용하면, 모델의 성능을 전반적으로 유사한 수준으로 유지하면서도, 효율적이고 일관된 결과를 얻을 수 있었음. 최적화 알고리즘은 수작업으로 하기 어려운 복잡한 튜닝 작업을 자동화함으로써 더 나은 결과를 빠르게 도출할 수 있음.

○ 추가적인 성능 개선을 위해서는 데이터의 품질 개선 또는 라벨링 방식의 수정, 그리고 차원 축소 기법 등을 적용하여 데이터의 품질 변화가 모델 성능에 미치는 영향을 분석할 필요가 있음. 이 과정에서 데이터의 다양성이나 특성 변환이 성능에 어떤 변화를 가져오는지 확인이 필요함.

3. 고위험사업장 선별 모델 설계 및 개발

1) 모델 설계 및 실험 개요

○ 이번 연구에서는 라벨링된 데이터 조건을 추가하여 범위를 조절하고, 위험 수준 현장평가 데이터를 추가로 활용하여 고위험 사업장 선별 모델의 성능을 개선하고자 하였음.

1. 산재발생 데이터 적용

- 산재발생 이력이 없는 사업장의 경우, 향후 발생할 사고가 경미한 수준인지, 큰 사고로 이어질지 불확실하기 때문에, 이를 Label 0(안전사업장)으로 처리할 때 노이즈가 발생할 수 있다는 가정을 두고 학습을 진행함. 이를 통해 데이터가 모델에 미치는 영향을 확인하였음.

2. 위험 수준 현장평가 데이터 적용

- 사업장 현장에서 직접 평가한 3개의 문항 합계를 기준으로 임계치(10~13점)을 조정하여 Label(안전/위험)에 변화를 주었음. 현장 평가 데이터가 모델 성능에 미치는 영향을 확인하였으며, 사람이 직접 현장에서 평가한 위험도를 모델이 정확하게 반영할 수 있는지를 확인하였음.

3. 산재발생 데이터와 위험 수준 현장평가 데이터의 결합

- 산재발생 데이터와 위험 수준 현장평가 데이터를 결합하여 학습을 진행한 결과, 두 데이터가 상호보완하여 모델의 성능이 향상되는 효과를 확인하였음.

4. 중요 특성 활용

- 모델 학습에서 특성 간 강한 상관관계와 불필요한 특성들이 모델 성능에 미치는 영향을 확인하기 위해, 근로손실일수와 관련된 영향력이 낮

은 특성들을 제외하고 상위 50개의 특성만을 적용하여 추가 학습을 진행함.

2) 모델 설계 및 실험의 상세 내용

○ 학습데이터에 조건을 추가 적용하여 개별로 학습을 진행하여 고위험 사업장 선별 모델의 성능을 개선하고 평가하였음. 각 데이터의 적용 방식과 모델 실험의 결과를 정리하였음.

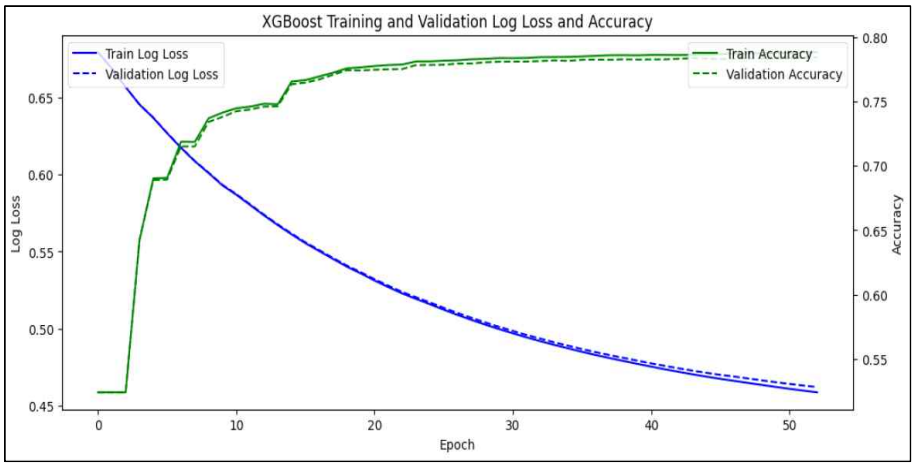
〈표 3-18〉 데이터 추가 및 실험 내용

NO	구분	내용
1	산재발생 데이터 적용	• Label 1(위험사업장)의 경우 위험성을 확실하게 확인이 가능하지만, 산재발생 이력이 없는 사업장은 향후 사고 발생 시, 경미한 사고가 주로 발생하게 될지, 큰 피해가 발생할지 알 수 없으므로, 이를 Label 0(안전사업장)으로 사용 시 노이즈가 될 수 있음을 가정하였음
2	위험 수준 현장평가 데이터 적용	• 사업장 현장에서 평가한 5점 기준 3개 문항 합계를 기반으로, 임계치를 10~13 사이로 조정하여 위험/안전 Label에 변화를 주었음
3	산재발생 데이터 + 위험 수준 현장평가 데이터 적용	• 1번 산재발생 데이터와 2번 위험 수준 현장평가 데이터를 동시에 적용하여 학습하여 상호보완적으로 작용하여 모델 성능이 향상됨을 확인하였음

○ 추가로, 모델 학습에서 특성간 강한 상관관계와 불필요한 특성들이 모델에 미치는 영향을 확인하기 위해, 근로손실일수와 영향력이 낮은 특성들을 제외하고, 상위 50개 특성만을 적용하여 추가 학습을 진행하였음.

(1) 산재발생 데이터 적용

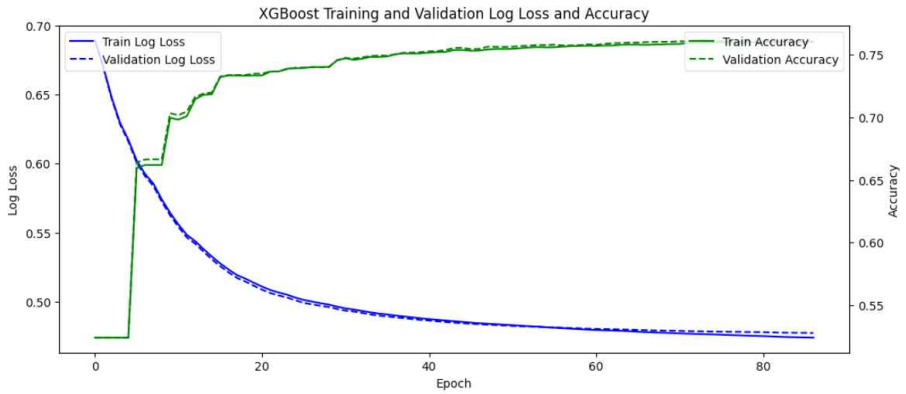
○ 제조업 사업장 데이터 내 산재발생 데이터 적용 후 학습 결과.

구분	내용
작업명	제조업_XGBoost_RandomSearch최적화_산재발생 데이터 적용
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.78 2. 정밀도(Precision): - Label 0 (0.75): 모델이 안전 사업장으로 예측한 것 중 75%가 실제 안전 사업장임을 의미함 - Label 1 (0.82): 모델이 위험 사업장으로 예측한 것 중 82%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.82): 실제 안전 사업장 중 82%가 정확하게 예측함 - Label 1 (0.76): 실제 위험 사업장 중 76%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.78 - Label 1 : 0.79

○ 제조업 사업장은 사고사례가 있는 사업장의 개수는 67,581건이고, 사업장의 위험사업장 여부를 나타내는 원본 레이블링 값을 그대로 사용함.

○ 모델 중 가장 성능이 좋았으며, 이는 실제 사고사례가 있는 데이터를 기반으로 분석한 결과로 모델 성능에 기여한 것으로 보여짐.

○ 제조업 사업장 데이터 내 산재발생 데이터 적용 및 상위 50개 특성으로 학습한 결과

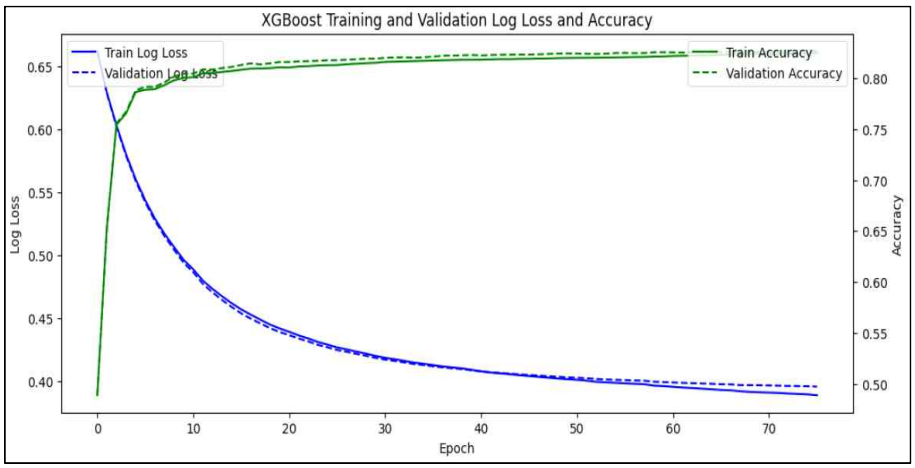
구분	내용
작업명	제조업_XGBoost_RandomSearch최적화_산재발생 데이터 적용_상위50특성
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.75 2. 정밀도(Precision): - Label 0 (0.75): 모델이 안전 사업장으로 예측한 것 중 75%가 실제 안전 사업장임을 의미함 - Label 1 (0.76): 모델이 위험 사업장으로 예측한 것 중 76%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.73): 실제 안전 사업장 중 73%가 정확하게 예측함 - Label 1 (0.78): 실제 위험 사업장 중 78%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.74 - Label 1 : 0.77

○ 제조업 사업장 데이터 내 산재발생 데이터 적용 및 상위 50개 특성으로 학습한 결과.

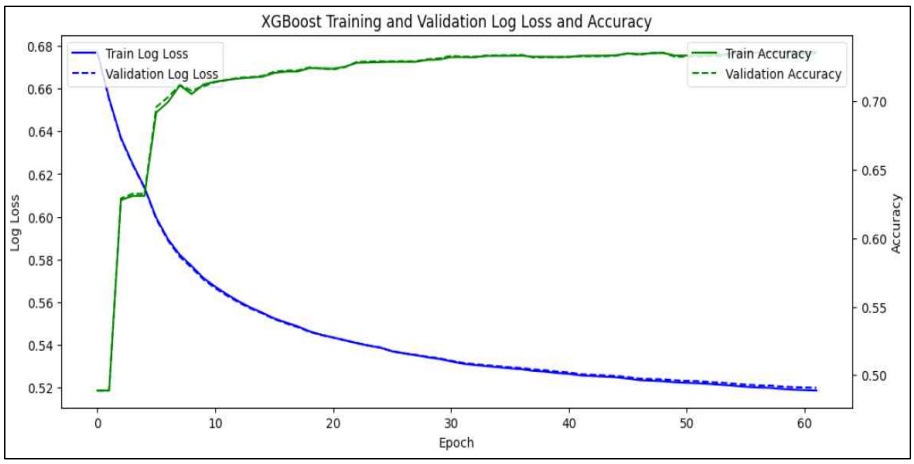
○ 안전/위험 사업장 판별 기준이 되는 특성인 'target'과 상관분석을 통해 상관계수가 높은 상위 50개의 특성으로 학습한 결과임.

○ 전체 특성 417개 중 약 12% 정도만 학습에 사용돼 모델 성능이 전체 특성을 사용한 모델보다는 낮은 성능을 보임.

○ 서비스업 사업장 데이터 내 산재발생 데이터 적용 결과

구분	내용
작업명	서비스업_XGBoost_RandomSearch최적화_산재발생 데이터 적용
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.81 2. 정밀도(Precision): - Label 0 (0.81): 모델이 안전 사업장으로 예측한 것 중 81%가 실제 안전 사업장임을 의미함 - Label 1 (0.83): 모델이 위험 사업장으로 예측한 것 중 83%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.85): 실제 안전 사업장 중 85%가 정확하게 예측함 - Label 1 (0.79): 실제 위험 사업장 중 79%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.83 - Label 1 : 0.81

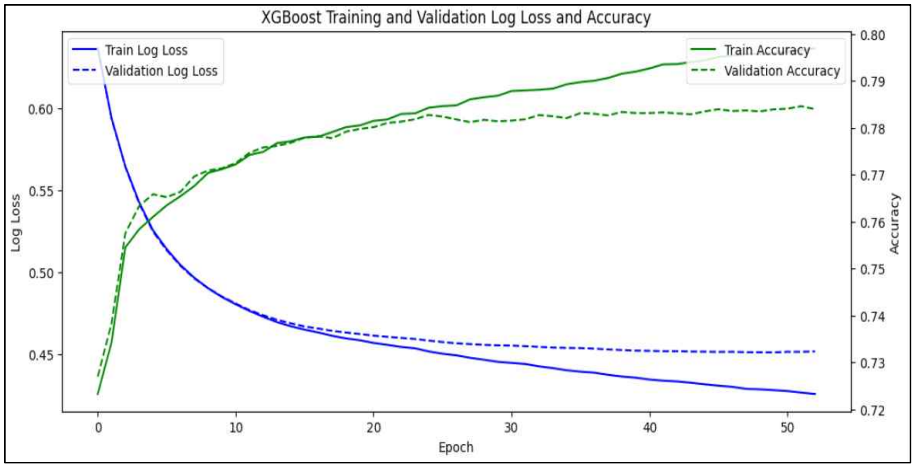
○ 서비스업 사업장 데이터 내 산재발생 데이터 적용 및 상위 50개 특성으로 학습한 결과

구분	내용
작업명	서비스업_XGBoost_RandomSearch최적화_산재발생 데이터 적용_상위50특성
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.73 2. 정밀도(Precision): - Label 0 (0.75): 모델이 안전 사업장으로 예측한 것 중 75%가 실제 안전 사업장임을 의미함 - Label 1 (0.72): 모델이 위험 사업장으로 예측한 것 중 72%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.71): 실제 안전 사업장 중 71%가 정확하게 예측함 - Label 1 (0.76): 실제 위험 사업장 중 76%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.73 - Label 1 : 0.74

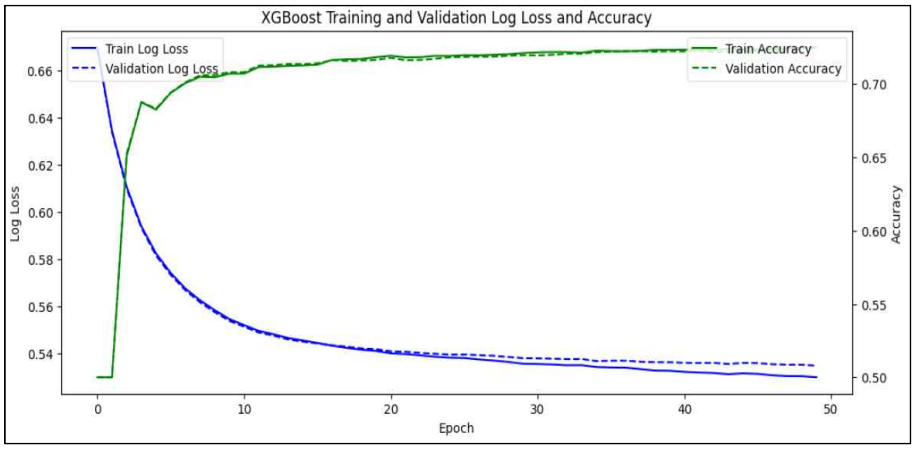
○ 전체 특성 473개 중 약 10% 정도만 학습에 사용돼 모델 성능이 전체 특성을 사용한 모델보다는 낮은 성능을 보임.

(2) 위험 수준 현장평가 데이터 적용

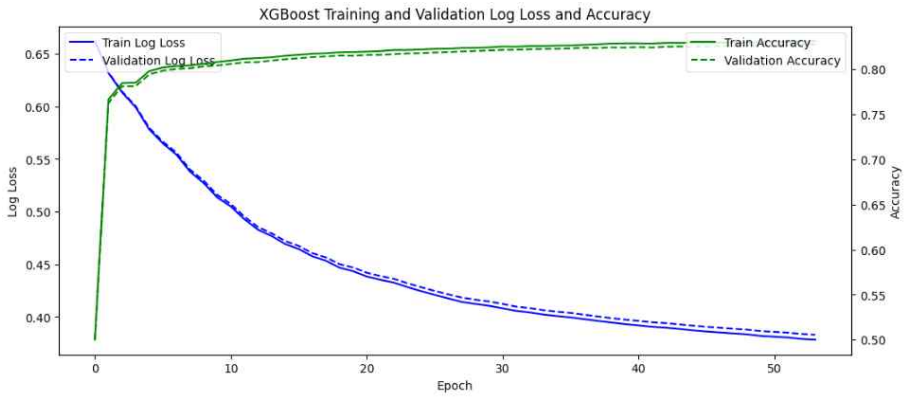
○ 제조업 사업장 데이터 내 위험 수준 현장평가 적용 결과

구분	내용
작업명	제조업_XGBoost_RandomSearch최적화_위험 수준 현장평가 적용
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.77 2. 정밀도(Precision): - Label 0 (0.77): 모델이 안전 사업장으로 예측한 것 중 77%가 실제 안전 사업장임을 의미함 - Label 1 (0.79): 모델이 위험 사업장으로 예측한 것 중 79%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.79): 실제 안전 사업장 중 79%가 정확하게 예측함 - Label 1 (0.76): 실제 위험 사업장 중 76%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.78 - Label 1 : 0.77

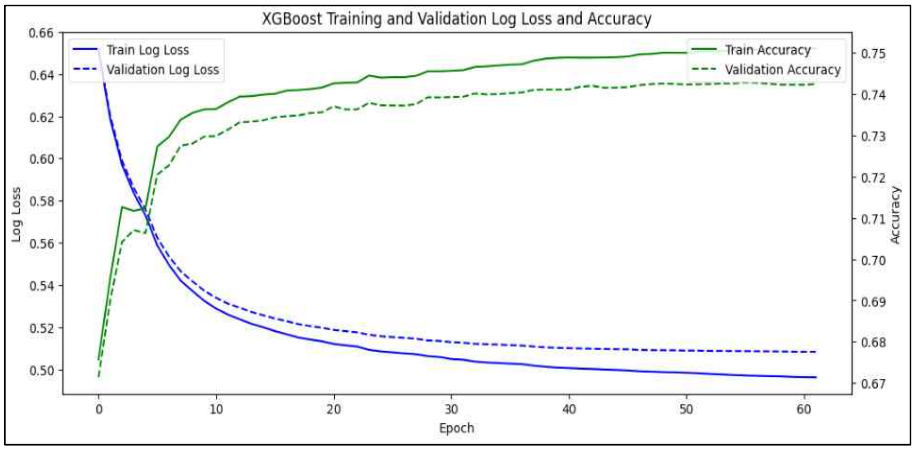
○ 제조업 사업장 데이터 내 위험 수준 현장평가 적용 및 상위 50개 특성으로 학습한 결과

구분	내용
작업명	제조업_XGBoost_RandomSearch최적화_위험 수준 현장평가 적용_상위50특성
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.71 2. 정밀도(Precision): - Label 0 (0.74): 모델이 안전 사업장으로 예측한 것 중 74%가 실제 안전 사업장임을 의미함 - Label 1 (0.70): 모델이 위험 사업장으로 예측한 것 중 70%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.67): 실제 안전 사업장 중 67%가 정확하게 예측함 - Label 1 (0.77): 실제 위험 사업장 중 77%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.71 - Label 1 : 0.73

○ 서비스업 사업장 데이터 내 위험 수준 현장평가 적용 결과

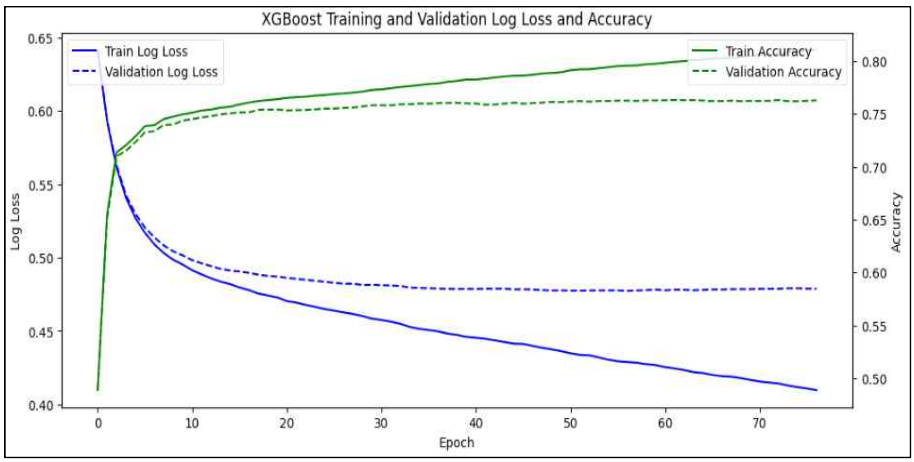
구분	내용
작업명	서비스업_XGBoost_RandomSearch최적화_위험 수준 현장평가 적용
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.83 2. 정밀도(Precision): - Label 0 (0.81): 모델이 안전 사업장으로 예측한 것 중 81%가 실제 안전 사업장임을 의미함 - Label 1 (0.85): 모델이 위험 사업장으로 예측한 것 중 85%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.87): 실제 안전 사업장 중 87%가 정확하게 예측함 - Label 1 (0.74): 실제 위험 사업장 중 74%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.81 - Label 1 : 0.78

○ 서비스업 사업장 데이터 내 위험 수준 현장평가 적용 및 상위 50개 특성으로 학습한 결과

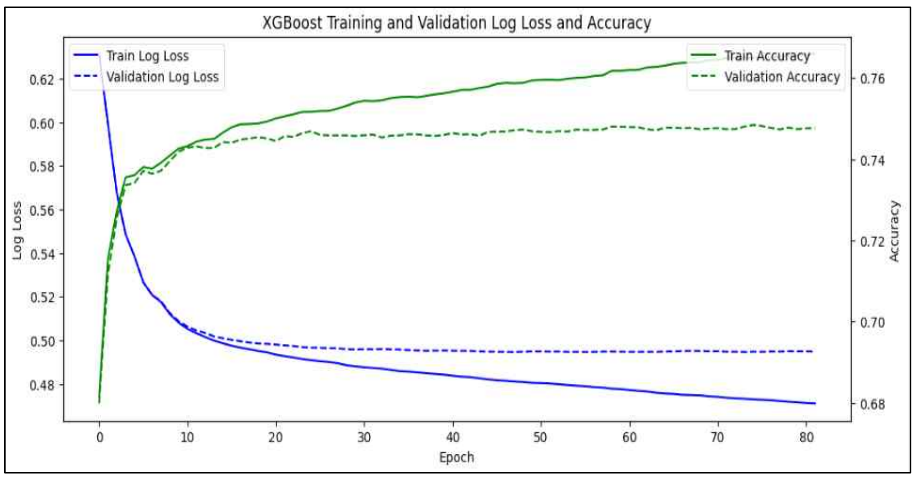
구분	내용
작업명	서비스업_XGBoost_RandomSearch최적화_위험 수준 현장평가 적용_상위50특성
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.74 2. 정밀도(Precision): - Label 0 (0.72): 모델이 안전 사업장으로 예측한 것 중 72%가 실제 안전 사업장임을 의미함 - Label 1 (0.77): 모델이 위험 사업장으로 예측한 것 중 77%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.80): 실제 안전 사업장 중 80%가 정확하게 예측함 - Label 1 (0.69): 실제 위험 사업장 중 69%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.76 - Label 1 : 0.73

(3) 산재발생 데이터 + 위험 수준 현장평가 데이터 적용

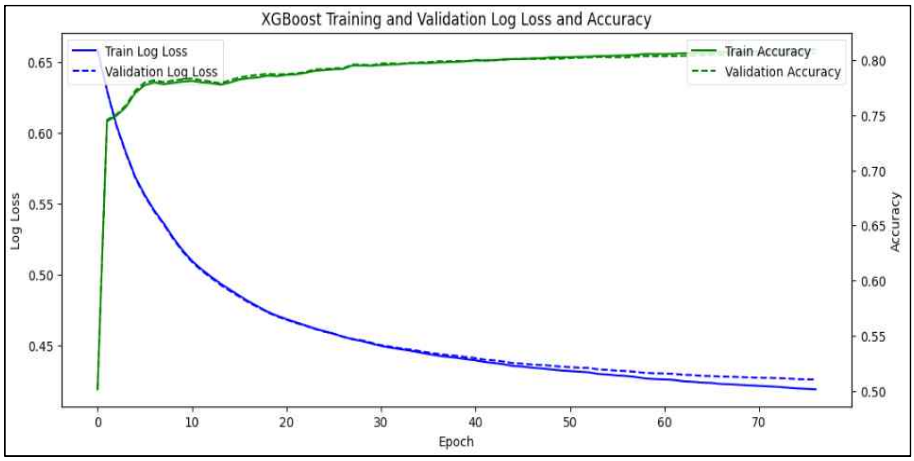
○ 제조업 사업장 데이터 내 산재발생 데이터 + 위험 수준 현장평가 적용 결과

구분	내용
작업명	제조업_XGBoost_RandomSearch최적화_산재발생 데이터 + 위험 수준 현장평가 적용
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.76 2. 정밀도(Precision): - Label 0 (0.79): 모델이 안전 사업장으로 예측한 것 중 79%가 실제 안전 사업장임을 의미함 - Label 1 (0.73): 모델이 위험 사업장으로 예측한 것 중 73%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.72): 실제 안전 사업장 중 72%가 정확하게 예측함 - Label 1 (0.80): 실제 위험 사업장 중 80%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.76 - Label 1 : 0.77

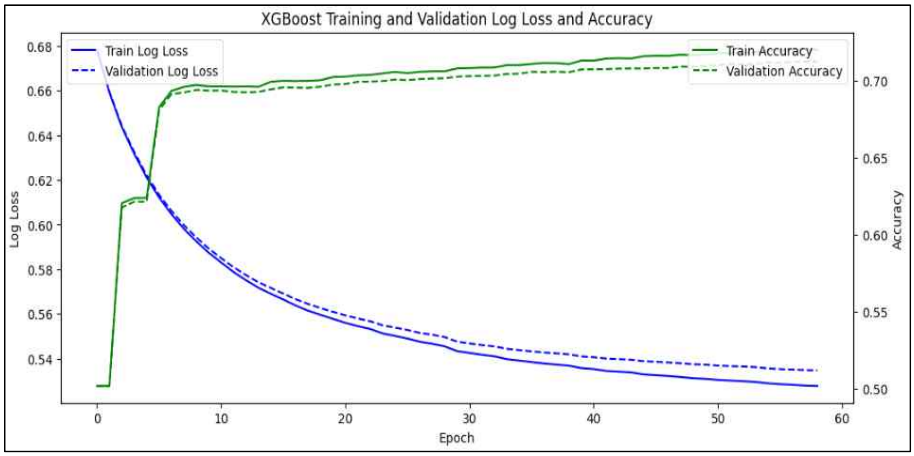
○ 제조업 사업장 데이터 내 산재발생 데이터 + 위험 수준 현장평가 적용 및 상위 50개 특성으로 학습한 결과

구분	내용
작업명	제조업_XGBoost_RandomSearch최적화_산재발생 데이터 + 위험 수준 현장평가 적용_상위50특성
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.74 2. 정밀도(Precision): - Label 0 (0.72): 모델이 안전 사업장으로 예측한 것 중 72%가 실제 안전 사업장임을 의미함 - Label 1 (0.78): 모델이 위험 사업장으로 예측한 것 중 78%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.82): 실제 안전 사업장 중 82%가 정확하게 예측함 - Label 1 (0.66): 실제 위험 사업장 중 66%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.77 - Label 1 : 0.72

○ 서비스업 사업장 데이터 내 산재발생 데이터 + 위험 수준 현장평가 적용 결과

구분	내용
작업명	서비스업_XGBoost_RandomSearch최적화_산재발생 데이터 + 위험 수준 현장평가 적용
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.80 2. 정밀도(Precision): - Label 0 (0.78): 모델이 안전 사업장으로 예측한 것 중 78%가 실제 안전 사업장임을 의미함 - Label 1 (0.84): 모델이 위험 사업장으로 예측한 것 중 84%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.86): 실제 안전 사업장 중 86%가 정확하게 예측함 - Label 1 (0.76): 실제 위험 사업장 중 76%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.81 - Label 1 : 0.80

○ 서비스업 사업장 데이터 내 산재발생 데이터 + 위험 수준 현장평가 적용 및 상위 50개 특성으로 학습한 결과

구분	내용
작업명	서비스업_XGBoost_RandomSearch최적화_산재발생 데이터 + 위험 수준 현장평가 적용_상위50특성
학습 그래프	
모델 평가	1. 정확도(Accuracy): 0.71 2. 정밀도(Precision): - Label 0 (0.74): 모델이 안전 사업장으로 예측한 것 중 74%가 실제 안전 사업장임을 의미함 - Label 1 (0.70): 모델이 위험 사업장으로 예측한 것 중 70%가 실제 위험 사업장임을 의미함 3. 재현율(Recall) - Label 0 (0.68): 실제 안전 사업장 중 68%가 정확하게 예측함 - Label 1 (0.76): 실제 위험 사업장 중 76%가 정확하게 예측함 4. F1-Score: - Label 0 : 0.71 - Label 1 : 0.73

4. 산업안전 도메인 맞춤형 언어모델 개선 및 모델 성능 비교 분석

1) 산업안전 도메인 맞춤형 언어모델 튜닝

(1) 언어모델 개발 개요

○ 본 연구에서는 고위험 사업장 분류모델의 해석력을 향상시키기 위해 사전학습된 언어모델을 활용하여 XGBoost 모델의 예측 결과를 보다 직관적이고 설명 가능하게 만들고자 하였음.

○ XGBoost는 성능이 뛰어나고, 해석 가능한 AI(XAI) 기법인 Feature Importance와 SHAP를 사용해 예측 결과를 해석하는 데 일반적으로 사용되고 있음. 그러나 이 방법은 수치적 해석에 의존하여, 비전문가가 이해하기 어려울 수 있으며, 특정 도메인에서 보다 구체적인 설명을 제공하는 데 한계가 있음.

○ 이를 보완하기 위해, 언어모델을 활용하여 예측 결과를 직관적이고 자연어로 설명하는 방식을 적용시켰음. XGBoost 모델이 사업장의 위험성을 예측하면, 해당 사업장에서 어떤 요인이 위험 요소로 작용하는지를 자연어로 설명하는 모델을 구축하는 것을 목표로 함. 이를 통해 사업장 위험 요소를 파악하고, 개선 방안을 제시하며, 비전문가도 쉽게 이해할 수 있는 설명 가능한 정보를 제공함.

(2) 모델 선정 및 학습 방법 비교

○ XGBoost와 같은 전통적인 머신러닝 모델에서 사용되는 해석 기법인 Feature Importance와 SHAP는 정량적 분석을 통해 예측 결과를 해석하는 방법임. 이러한 기법들은 모델이 어떤 특성을 중요하게 사용했는지 설명하는 데 유용하지만, 비전문가에게는 직관적이지 않다는 단점이

존재함.

○ 따라서, 사전학습된 언어모델을 통해 예측 결과를 자연어로 설명하는 방법이 대안으로 제안하였으며, 특히 Gemma와 같은 모델을 튜닝하여, 모델의 예측 결과를 사용자가 쉽게 이해할 수 있는 방식으로 설명하는 접근법을 도입하게 되었음. 이를 위해 다양한 학습 방법을 검토한 결과, 서버 자원과 효율성을 고려하여 LoRA(Low-RANK Adaptation) 기법을 적용하게 되었음.

○ 〈표 3-19〉는 언어모델 학습 방법을 비교한 것임.

〈표 3-19〉 언어모델 학습 방법의 비교 표

구분	개념	장점	단점
전체 파인튜닝 (Full Fine-Tuning)	사전 학습된 언어모델의 모든 파라미터를 대상 도메인 데이터로 재학습	모델 전체를 최적화하므로 도메인 적합성이 높아짐	대용량 모델의 경우 학습에 막대한 계산 자원과 시간이 소요되며, 과적합 위험이 있음
미세 조정 (Fine-Tuning)	모델의 일부 계층만 재학습하여 도메인 특성을 반영함	계산 비용이 전체 파인튜닝보다 적으며, 도메인 적합성도 향상됨	업데이트하지 않는 계층의 한계로 인해 성능 향상에 제한이 있을 수 있음
어댑터 방법 (Adapter Methods)	기존 모델에 소규모 어탭터 모듈을 추가하여 도메인 특성을 학습함	메모리 효율성이 높고, 다양한 도메인에 쉽게 적용 가능	추가적인 모듈로 인해 추론 속도가 느려짐
LoRA (Low-Rank Adaptation)	모델의 파라미터 업데이트를 저랭크 행렬 분해로 표현하여 학습 파라미터 수를 줄임	메모리 사용량과 계산량을 크게 줄여 대용량 모델의 학습을 효율적으로 수행함	저랭크 근사로 인해 표현력 감소

(3) 모델 선정 이유

○ Gemma는 Google DeepMind에서 개발된 오픈 모델로, 텍스트 생성

과 명령어 기반 상호작용에 최적화되어 있음. 특히, 경량화된 구조 덕분에 제한된 자원에서도 학습이 가능함.

○ Gemma-2 9B 모델은 27B 모델에서 지식을 증류한 방식으로 학습되어, 연산 및 메모리 자원을 절약하면서도 대규모 모델에 접근한 성능을 제공함. 이는 사전 학습된 모델을 기반으로, 새로운 도메인에 맞춘 미세 조정과 효율적인 학습을 가능하게 함.

○ 또한, Rotary Positioning Embeddings(RoPE)와 Local Attention 및 Global Attention을 교차적용하여, 텍스트의 세부 정보와 전체 맥락을 모두 고려하는 구조를 가지고 있음.

(4) LoRA를 적용한 Gemma 모델 튜닝 방법

○ LoRA(Low-Rank Adaptation)는 Transformer 구조의 Self-Attention 모듈에 적용되는 기법으로, Query(Q)와 Value(V) 행렬에 저차원 행렬을 추가하여 파라미터 수를 줄이고, 효율적인 학습을 가능하게 함. LoRA를 적용한 학습 방법은 다음과 같음.

1. **사전학습된 모델 로드**: ko-gemma-2-9b-it 모델을 로드하고, 초기 파라미터를 고정함.
2. **LoRA 적용**:기존 파라미터를 고정한 채, 저차원 행렬만을 업데이트하여 학습을 진행함.
3. **모델 튜닝 및 학습**: 사업장 정보와 XGBoost의 예측 결과에 대해, 위험 요소와 그 이유를 설명하는 자연어 생성 모델을 구현하고, 사용자가 이해하기 쉬운 포맷으로 출력함.

○ 언어모델 학습에 사용된 시스템은 NVIDIA RTX 계열 GPU 4대를 장착하여, 멀티 GPU 학습을 수행할 수 있도록 하여 신속한 모델 학습이 가능하도록 구성하였음.

〈표 3-20〉 언어모델 학습서버 구성 정보

구분	내용	비고
CPU	4.0GHz-5, 3GHz, 32-core, 64-Thread, 128MB Cache	
RAM	512GB	
SSD	4TB	
HDD	12TB	
GPU	NVIDIA RTX (VRAM 24G, 82 TFLOPS) x 4ea	멀티 GPU 학습
NIC	10G Port	
SW	Ubuntu, CUDA Toolkit 12.1, cuDNN, Nvidia-Docker, Python 3.10	

○ 〈표 3-21〉은 본 연구에서 LoRA 기법을 활용하여 학습 시 사용된 하이퍼파라미터 및 설정값임.

〈표 3-21〉 Gemma모델 LoRA학습 설정 내용

구분	설정값	설명
시퀀스 길이	600	입력 시퀀스의 최대 길이(토큰 단위)
배치 크기	32	실질적인 배치 크기는 미니 배치 크기와 그라디언트 축적 단계의 곱으로 계산됨
에폭 수	300	전체 학습 데이터셋을 몇 번 반복할지 설정
학습률	0.0002	학습 중 사용되는 초기 학습률
워밍업 스텝	300	모델 학습 초반의 워밍업 스탭 수로, 일정 학습률로 시작한 후 점차적으로 학습률을 높임
평가 주기	4	에폭 당 평가 횟수
LoRA r 값	16	LoRA에서 저차원 행렬의 랭크. 메모리 사용량과 모델 표현력에 영향을 줌
LoRA alpha 값	32	LoRA의 스케일링 파라미터로, 학습 강도에 영향을 미침
LoRA Dropout	0.05	과적합을 방지하기 위해 사용

구분	설정값	설명
LoRA 타겟 모듈	lora_target_linear: true	모든 선형 계층에 LoRA를 적용
최적화 기법	paged_adamw_8bit	메모리 효율성을 위해 8비트 최적화 기법 사용
학습 스케줄러	cosine	학습률을 조절하는 코사인 스케줄러 사용
gradient_checkpointing	true	메모리 사용량 절감을 위해 그라디언트 체크포인트 사용
flash_attention	true	Attention 성능을 향상시키기 위한 Flash Attention 활성화

2) 언어모델 학습 데이터 가공 및 생성

(1) 언어모델 학습용 특성 선별 및 데이터 구성

○ 본 연구에서는 XGBoost 모델의 예측 결과를 기반으로 데이터를 구성하고, AI 학습에 활용하였음. XGBoost 모델은 고위험 사업장일 확률을 예측하는데 사용되며, 예측된 위험 확률과 주요 입력 파라미터들을 바탕으로 텍스트 데이터를 재구성 함.

- 데이터 구성 방식: 각 사업장에서 위험 요소로 작용할 수 있는 설비 종류, 작업 환경, 근무 환경 등의 주요 변수를 선별하여 텍스트 데이터로 가공하였음. 약 400개의 특성을 모두 사용한 것이 아니라, 비슷한 특성명은 제외하고 모델 학습에 중요한 주요 특성들을 선별하여 구성하였고, 이 과정에서 선택된 변수를 상위 그룹으로 묶어 보다 단순화된 구조로 데이터가 자연스럽게 연결될 수 있도록 구성하였음.

- 텍스트 구성 목적: 텍스트 데이터는 언어모델이 데이터를 보다 효율적으로 학습할 수 있도록 주요 학습할 수 있도록, 주요 변수를 핵심 정보 중심으로 최적화하여 구성하였음. 이러한 구성은 언어모델이 데이터를 이해하고, 위험 요소를 분석하는 데 적합하도록 설계되었으며, 모델이

학습하는 데이터는 주요 위험 요소들에 대한 설명력을 높이고, 예측된 위험 확률과 그 원인에 대한 구체적인 분석을 제공할 수 있게 됨.

〈표 3-22〉 언어모델에 사용하기 위한 데이터 선별 및 그룹화

구분	특성 그룹	데이터 정제 예시
근무환경	위험설비 종류	분쇄, 파쇄기, 사출성형기, 산업용 로봇, 승강기
	작업환경 밀폐공간(위험환경)	강재등시설, 기타밀폐공간
	작업환경 밀폐공간(안전환경)	급기팬, 가스농도측정기
	근골격계부담작업 여부	있음, 없음
	복지시설 여부	많음, 적음, 없음
	야간작업 유무	있음, 없음
	유해요인 조사여부	있음, 없음
시설	대량위험물 제조소 여부	있음, 없음
	석유화학단지 내 사업장 여부	있음, 없음
	예방규정 제출대상 여부	있음, 없음
	근로자 작업환경	고열/한랭/다습 및 방사선 취급 작업, 산소결핍 위험장소 작업, 분진/흄발생작업, 사내하도급작업, 소음작업, 제조나노물질의 제조 및 취급작업, 진동발생작업
교육	안전검사 점수	최하, 하, 중, 상, 최상
	안전교육 수료비율	높음, 보통, 낮음
	관리자 업무 이해점수	최하, 하, 중, 상, 최상
	법정방호장치점수	최하, 하, 중, 상, 최상
	심사결과 부적합 비율	최하, 하, 중, 상, 최상
	안전보건 관리인력	관리자, 보건관리자, 보건담당자

구분	특성 그룹	데이터 정제 예시
사고	위험 기인물	끼임, 떨어짐, 부딪힘
	유해위험 물질수	최하, 하, 중, 상, 최상
	재해발생률(3년평균)	상, 중, 하, 없음
	상해발생 빈도	상, 중, 하, 없음
	취급위험물질 여부	PSM 대상물질, 건강관리수첩 대상물질, 관리대상물질, 금지대상물질, 안전관리물질, 안전검사물질, 특검대상물질, 측정대상물질, 허용기준설정물질, 허가대상유해물질
근로자	근로자 정보	근로자는 N명이며, 평균 M년의 근속기간을 가짐
	성별	근로자는 N%의 남성과 M%의 여성으로 이루어져 있음
	연령대	근로자들 중 '50대'와 '40대'가 가장 많음
	근로자 직종	건설*채굴직이 가장 많은 비율을 차지하고 있으며, 농림어업직, 보건의료직이 다음으로 많은 비율을 차지하고 있음
XGBoost를 통한 위험사업장 판단 결과	위험사업장 확률	0.95

(2) 프롬프트 처리 및 학습데이터 생성

○ XGBoost 예측 결과를 바탕으로 텍스트 데이터를 생성하고, 이를 프롬프트로 변환하여 언어모델에 입력하였음. 〈표 Ⅲ-23〉는 사용한 프롬프트 예시 일부임.

〈표 3-23〉 언어모델 입력 프롬프트 예시 일부

구분	입력 프롬프트
제조업 Prompt-01	'위험사업장으로 판단된 근거', '안전사업장으로 판단된 근거', '위험요소', '개선방안'을 설명해줘. --- 해당 사업장이 위험사업장일 확률은 8.14%로 안전사업장에 가까움. ### 사업장 정보 - 사업분야 : 제조업 - 근로자 정보 : 근로자는 6명 이며, 평균 11.67년의 근속기간을 가짐 - 근로자 남녀 성비 정보 : 근로자는 66.67%의 남성과 33.33%의 여성으로 이루어져 있음 - 연령대 정보 : 근로자들 중 '40대'와 '30대'가 가장 많음 - 최근 재해 발생률 : 재해 없음 - 상해사건 발생경력 : 없음 ### 설비 및 시설 관련 정보 - 작업에 사용하는 위험설비 종류 : 압력용기 - 대량위험물 제조소 여부 : 없음 - 석유화학단지내 사업장 여부 : 없음 - 예방규정 제출대상 여부 : 없음 - 근로자 위험 작업환경 : 없음 - 작업환경(위험환경) 밀폐공간 종류 : 없음 - 작업환경(안전환경) 밀폐공간 종류 : 급기팬, 밀폐실태-가스농도측정기를 보유하고 있음 - 위험물질에 의한 위험도 : 매우 낮음 - 취급위험물질 종류 : 없음 ### 근무환경 정보 - 근골격계부담작업 여부 : 없음 - 복지시설 여부 : 없음 - 야간작업 유무 : 없음 - 유해요인 조사 여부 : 없음 - 근무환경에서 발생가능한 위험 기인물 종류 : 화재 ### 교육 관련 - 안전검사 점수(최하, 하, 중, 상, 최상) : 중 - 안전교육 수료비율 : 낮음 - 관리자 업무 이해 점수(최하, 하, 중, 상, 최상) : 하

구분	입력 프롬프트
	- 법정방호장치 점수(최하, 하, 중, 상, 최상) : 중 - 심사결과 부적합 비율 : 낮음 - 안전보건 관리인력 : 없음
제조업 Prompt-02	'위험사업장으로 판단된 근거', '안전사업장으로 판단된 근거', '위험요소', '개선방안'을 설명해줘. --- 해당 사업장이 위험사업장일 확률은 0.7%로 안전사업장에 가까움. ### 사업장 정보 - 사업분야 : 제조업 - 근로자 정보 : 근로자는 1명 이며, 근속년수에 대한 정보가 없음 - 근로자 남녀 성비 정보 : 근로자는 100.0%의 남성과 0.0%의 여성으로 이루어져 있음 - 연령대 정보 : 근로자들 중 '50대'가 가장 많음 - 최근 재해 발생률 : 재해 없음 - 상해사건 발생경력 : 없음 ### 설비 및 시설 관련 정보 - 작업에 사용하는 위험설비 종류 : 없음 - 대량위험물 제조소 여부 : 없음 - 석유화학단지내 사업장 여부 : 없음 - 예방규정 제출대상 여부 : 없음 - 근로자 위험 작업환경 : 없음 - 작업환경(위험환경) 밀폐공간 종류 : 없음 - 작업환경(안전환경) 밀폐공간 종류 : 없음 - 위험물질에 의한 위험도 : 매우 낮음 - 취급위험물질 종류 : 없음 ### 근무환경 정보 - 근골격계부담작업 여부 : 없음 - 복지시설 여부 : 없음 - 야간작업 유무 : 없음 - 유해요인 조사 여부 : 없음 - 근무환경에서 발생가능한 위험 기인물 종류 : 없음 ### 교육 관련 - 안전검사 점수(최하, 하, 중, 상, 최상) : 상 - 안전교육 수료비율 : 낮음

구분	입력 프롬프트
	- 관리자 업무 이해 점수(최하, 하, 중, 상, 최상) : 중 - 법정방호장치 점수(최하, 하, 중, 상, 최상) : 상 - 심사결과 부적합 비율 : 낮음 - 안전보건 관리인력 : 없음
제조업 Prompt-03	'위험사업장으로 판단된 근거', '안전사업장으로 판단된 근거', '위험요소', '개선방안'을 설명해줘. --- 해당 사업장이 위험사업장일 확률은 93.41%로 위험사업장에 가까움. ### 사업장 정보 - 사업분야 : 제조업 - 근로자 정보 : 근로자는 7명 이며, 평균 5.62년의 근속기간을 가짐 - 근로자 남녀 성비 정보 : 근로자는 75.0%의 남성과 25.0%의 여성으로 이루어져 있음 - 연령대 정보 : 근로자들 중 '50대'와 '40대'가 가장 많음 - 최근 재해 발생률 : 높음 - 상해사건 발생경력 : 없음 ### 설비 및 시설 관련 정보 - 작업에 사용하는 위험설비 종류 : 지게차 - 대량위험물 제조소 여부 : 없음 - 석유화학단지내 사업장 여부 : 없음 - 예방규정 제출대상 여부 : 없음 - 근로자 위험 작업환경 : 소음작업 - 작업환경(위험환경) 밀폐공간 종류 : 없음 - 작업환경(안전환경) 밀폐공간 종류 : 없음 - 위험물질에 의한 위험도 : 매우 낮음 - 취급위험물질 종류 : 안전관리물질 ### 근무환경 정보 - 근골격계부담작업 여부 : 없음 - 복지시설 여부 : 적음 - 야간작업 유무 : 없음 - 유해요인 조사 여부 : 없음 - 근무환경에서 발생가능한 위험 기인물 종류 : 없음 ### 교육 관련 - 안전검사 점수(최하, 하, 중, 상, 최상) : 하

구분	입력 프롬프트
	- 안전교육 수료비율 : 높음 - 관리자 업무 이해 점수(최하, 하, 중, 상, 최상) : 최하 - 법정방호장치 점수(최하, 하, 중, 상, 최상) : 중 - 심사결과 부적합 비율 : 낮음 - 안전보건 관리인력 : 없음
제조업 Prompt-04	'위험사업장으로 판단된 근거', '안전사업장으로 판단된 근거', '위험요소', '개선방안'을 설명해줘. --- 해당 사업장이 위험사업장일 확률은 5.87%로 안전사업장에 가까움. ### 사업장 정보 - 사업분야 : 제조업 - 근로자 정보 : 근로자는 1명 이며, 근속년수에 대한 정보가 없음 - 근로자 남녀 성비 정보 : 정보없음 - 연령대 정보 : 정보없음 - 최근 재해 발생률 : 재해 없음 - 상해사건 발생경력 : 없음 ### 설비 및 시설 관련 정보 - 작업에 사용하는 위험설비 종류 : 크레인(천장,갠트리) - 대량위험물 제조소 여부 : 없음 - 석유화학단지내 사업장 여부 : 없음 - 예방규정 제출대상 여부 : 없음 - 근로자 위험 작업환경 : 소음작업 - 작업환경(위험환경) 밀폐공간 종류 : 없음 - 작업환경(안전환경) 밀폐공간 종류 : 급기팬, 밀폐실태-가스농도측정기를 보유하고 있음 - 위험물질에 의한 위험도 : 매우 낮음 - 취급위험물질 종류 : 없음 ### 근무환경 정보 - 근골격계부담작업 여부 : 없음 - 복지시설 여부 : 없음 - 야간작업 유무 : 없음 - 유해요인 조사 여부 : 없음 - 근무환경에서 발생가능한 위험 기인물 종류 : 끼임 ### 교육 관련

구분	입력 프롬프트
	- 안전검사 점수(최하, 하, 중, 상, 최상) : 중 - 안전교육 수료비율 : 낮음 - 관리자 업무 이해 점수(최하, 하, 중, 상, 최상) : 하 - 법정방호장치 점수(최하, 하, 중, 상, 최상) : 중 - 심사결과 부적합 비율 : 낮음 - 안전보건 관리인력 : 없음
제조업 Prompt-05	'위험사업장으로 판단된 근거', '안전사업장으로 판단된 근거', '위험요소', '개선방안'을 설명해줘. --- 해당 사업장이 위험사업장일 확률은 40.05%로 안전사업장에 가까움. ### 사업장 정보 - 사업분야 : 제조업 - 근로자 정보 : 근로자는 1명 이며, 평균 19.0년의 근속기간을 가짐 - 근로자 남녀 성비 정보 : 근로자는 100.0%의 남성과 0.0%의 여성으로 이루어져 있음 - 연령대 정보 : 근로자들 중 '60대이상'가 가장 많음 - 최근 재해 발생률 : 재해 없음 - 상해사건 발생경력 : 없음 ### 설비 및 시설 관련 정보 - 작업에 사용하는 위험설비 종류 : 지게차 - 대량위험물 제조소 여부 : 없음 - 석유화학단지내 사업장 여부 : 없음 - 예방규정 제출대상 여부 : 있음 - 근로자 위험 작업환경 : 소음작업 - 작업환경(위험환경) 밀폐공간 종류 : 없음 - 작업환경(안전환경) 밀폐공간 종류 : 없음 - 위험물질에 의한 위험도 : 매우 낮음 - 취급위험물질 종류 : 없음 ### 근무환경 정보 - 근골격계부담작업 여부 : 없음 - 복지시설 여부 : 없음 - 야간작업 유무 : 없음 - 유해요인 조사 여부 : 없음 - 근무환경에서 발생가능한 위험 기인물 종류 : 없음

구분	입력 프롬프트
	### 교육 관련 - 안전검사 점수(최하, 하, 중, 상, 최상) : 중 - 안전교육 수료비율 : 낮음 - 관리자 업무 이해 점수(최하, 하, 중, 상, 최상) : 중 - 법정방호장치 점수(최하, 하, 중, 상, 최상) : 중 - 심사결과 부적합 비율 : 낮음 - 안전보건 관리인력 : 없음

○ 위 프롬프트를 언어모델에 입력 시, 출력된 형태는 〈표 Ⅲ-24〉와 같으며, 언어모델 학습 시 그대로 학습 데이터로 활용됨.

〈표 3-24〉 언어모델 출력 결과 예시 일부

구분	언어모델 출력 결과
제조업 Prompt-01	## 사업장 위험도 분석 결과 및 근거 해당 사업장의 위험사업장 확률은 8.14%로 안전사업장에 가까운 수치입니다. 하지만 몇 가지 위험 요소와 개선 가능성을 확인할 수 있습니다. **1. 위험사업장으로 판단된 근거:** * **낮은 안전교육 수료비율:** 안전교육은 근로자의 안전 의식을 높이고 위험 상황에 대처하는 능력을 향상시키는 중요한 요소입니다. 낮은 수료비율은 안전 사고 발생 가능성을 높일 수 있습니다. * **관리자 업무 이해도 부족:** 관리자의 안전보건 관련 지식과 이해도가 부족하면 안전 관리 시스템 구축 및 운영에 어려움을 겪을 수 있습니다. 이는 안전사고 예방 및 대응에 부정적인 영향을 미칠 수 있습니다. * **안전보건 관리인력 부재:** 안전보건 관리인력은 안전 관리 시스템을 구축하고 운영하며, 근로자에게 안전 교육을 제공하는 중요한 역할을 수행합니다. 안전보건 관리인력이 부재할 경우 안전 관리 체계가 약화될 수 있습니다. **2. 안전사업장으로 판단된 근거:** * **최근 재해 발생률 및 상해사건 발생경력 부재:** 최근 재해 발생이 없고 상해사건 경력도 없는 것은 안전 관리 시스템이 어느 정도 효과적으로 작동하고 있음을 시사합니다.

구분	언어모델 출력 결과
	* **위험설비 사용 및 위험물질 취급 부재:** 압력용기 사용은 주의가 필요하지만, 대량위험물 제조소나 석유화학단지 내 사업장이 아니고, 위험물질을 취급하지 않는 점은 위험 요소를 크게 줄여줍니다. * **안전환경 시설 확보:** 급기팬과 가스농도측정기를 보유하고 있는 것은 작업 환경의 안전성을 높이는 데 기여합니다. **3. 위험요소:** * **낮은 안전교육 수료비율:** 안전 교육의 중요성을 강조하고, 근로자들의 참여를 유도하기 위한 노력이 필요합니다. * **관리자 업무 이해도 부족:** 안전보건 관련 교육 프로그램을 통해 관리자의 안전 관리 역량을 강화해야 합니다. * **안전보건 관리인력 부재:** 안전보건 관리인력을 채용하거나, 외부 전문가를 활용하여 안전 관리 시스템을 구축하고 운영하는 것이 좋습니다. **4. 개선방안:** * **안전교육 강화:** 안전교육 프로그램을 정기적으로 실시하고, 참여율을 높이기 위한 다양한 방법을 모색해야 합니다. * **관리자 안전보건 교육:** 안전보건 관련 법규 및 지침, 안전 관리 시스템 운영 등에 대한 교육을 통해 관리자의 역량을 강화해야 합니다. * **안전보건 관리인력 확보:** 안전보건 관리인력을 채용하거나, 외부 전문가를 활용하여 안전 관리 시스템을 구축하고 운영하는 것이 좋습니다. * **안전 점검 및 관리 시스템 구축:** 정기적인 안전 점검을 실시하고, 문제점을 개선하기 위한 시스템을 구축해야 합니다. * **근로자 참여 유도:** 안전 관리에 대한 근로자의 참여를 유도하고, 안전 문제를 신속하게 보고할 수 있는 시스템을 마련해야 합니다.
제조업 Prompt-02	## 사업장 위험도 분석 결과 및 근거 해당 사업장의 위험사업장 확률은 0.7%로 매우 낮으며, 안전사업장에 가까운 수준입니다. **1. 안전사업장으로 판단된 근거:** * **최근 재해 발생률 및 상해사건 발생경력:** 최근 재해 발생 및 상해사건 발생 경력이 없음. * **위험설비 및 위험물질:** 작업에 사용하는 위험설비가 없고, 대량위험물 제조소, 석유화학단지 내 사업장이 아니며, 취급위험물질도 없음. * **근무환경:** 근골격계 부담 작업, 야간 작업, 유해요인이 없는 안전한

구분	언어모델 출력 결과
	근무 환경. * **안전관리 점수:** 안전검사, 법정방호장치 점수가 상으로 높고, 심사결과 부적합 비율도 낮음. **2. 위험요소:** * **안전교육 수료비율:** 안전교육 수료비율이 낮다는 점은 근로자의 안전 의식 및 지식 부족 가능성을 시사합니다. * **안전보건 관리인력 부재:** 안전보건 관리인력이 없는 상황은 안전 관리 시스템 구축 및 운영에 어려움을 초래할 수 있습니다. * **복지시설 부재:** 복지시설이 없는 것은 근로자의 안전 및 건강 관리에 부정적인 영향을 미칠 수 있습니다. **3. 개선방안:** * **안전교육 강화:** 안전교육 프로그램을 정기적으로 실시하고, 수료율을 높이기 위한 노력이 필요합니다. * **안전보건 관리인력 확보:** 안전보건 관리인력을 확보하여 안전 관리 시스템을 구축하고 운영하는 것이 중요합니다. * **복지시설 확충:** 근로자의 건강과 안전을 위해 필수적인 복지시설을 확충해야 합니다. * **정기적인 위험 요소 점검:** 위험 요소를 정기적으로 점검하고 개선하여 안전사업장으로서의 수준을 유지해야 합니다. * **관리자 안전보건 교육 강화:** 관리자의 안전보건 지식과 책임감을 강화하기 위한 교육을 실시하여 안전 문화 조성에 기여해야 합니다.
제조업 Prompt-03	## 사업장 위험도 분석 결과 및 근거 해당 제조업 사업장의 위험사업장 확률은 93.41%로 매우 높습니다. **1. 위험사업장으로 판단된 근거:** * **높은 재해 발생률:** 최근 재해 발생률이 높다는 점은 가장 큰 위험 요소입니다. 이는 안전 관리 시스템의 부족이나 작업 환경의 문제점을 시사합니다. * **낮은 안전검사 점수:** 안전검사 점수가 '하'로 낮은 것은 안전 관리 시스템 및 작업 환경에 대한 철저한 점검 부족을 의미합니다. * **최저 관리자 업무 이해 점수:** 관리자의 안전보건 관련 업무 이해도가 낮은 것은 안전 문화 조성 및 안전 지침 준수에 큰 영향을 미칩니다. * **안전보건 관리인력 부재:** 안전보건 관리인력이 없는 상황은 안전 관리 시스템 구축 및 운영에 큰 어려움을 초래합니다.

구분	언어모델 출력 결과
	* **낮은 복지시설 여부:** 근로자 복지 시설이 부족한 것은 근로자의 안전과 건강을 위협할 수 있습니다. **2. 위험요소:** * **지게차 사용:** 지게차는 운전자의 실수나 정비 부실 시 심각한 사고를 유발할 수 있는 위험 설비입니다. * **소음 작업:** 장기간 소음 작업은 청력 손상 및 심혈관 질환 등의 건강 문제를 유발할 수 있습니다. * **낮은 안전교육 점수:** 안전 교육 수료 비율이 높더라도 교육의 질이 낮거나 실제 작업 환경에 적용되지 않으면 효과가 제한적입니다. **3. 개선방안:** * **안전 관리 시스템 강화:** 안전검사 점수 향상을 위해 정기적인 안전 점검 및 관리 계획을 수립하고 시행해야 합니다. * **안전 교육 프로그램 개선:** 실제 작업 환경에 맞춘 안전 교육 프로그램을 개발하고, 교육 참여도를 높이기 위한 노력이 필요합니다. * **관리자 안전보건 교육 강화:** 관리자의 안전보건 관련 지식과 역량을 강화하기 위한 교육 프로그램을 제공해야 합니다. * **안전보건 관리인력 확보:** 안전보건 관리인력을 채용하여 안전 관리 시스템 구축 및 운영을 책임지도록 해야 합니다. * **복지시설 확충:** 근로자 복지 시설을 확충하여 근로자의 안전과 건강을 증진시켜야 합니다. * **지게차 안전 관리 강화:** 지게차 운전자에 대한 정기적인 교육 및 시험을 실시하고, 지게차의 정기적인 점검 및 유지보수를 철저히 해야 합니다. * **소음 작업 관리 강화:** 소음 작업 환경을 개선하고, 소음 차단 장비를 제공하여 근로자의 청력 손상을 예방해야 합니다. * **재해 예방 프로그램 개발 및 시행:** 재해 발생 가능성을 분석하고, 예방을 위한 프로그램을 개발 및 시행하여 재해 발생률을 감소시켜야 합니다.
제조업 Prompt-04	## 사업장 위험도 분석 결과 및 근거 해당 사업장의 위험사업장 확률은 5.87%로 안전사업장에 가까운 수치입니다. 하지만, 몇 가지 위험 요소와 개선 가능한 부분들이 존재합니다.

구분	언어모델 출력 결과
	1. 위험사업장으로 판단된 근거: * **안전교육 수료비율이 낮음:** 안전 교육은 근로자의 안전 의식을 높이고 위험 상황에 대처하는 능력을 향상시키는 중요한 요소입니다. 낮은 수료율은 근로자들이 안전 지침을 충분히 이해하고 준수하지 못할 가능성을 높입니다. * **관리자 업무 이해 점수가 낮음:** 안전보건 관리자는 사업장의 안전 관리를 책임지는 중요한 역할을 합니다. 관리자의 안전 관련 지식 부족은 안전 관리 시스템의 약점으로 이어질 수 있습니다. * **안전보건 관리인력 부재:** 안전보건 관리인력은 안전 점검, 교육, 위험 요소 관리 등을 담당하여 안전한 작업 환경을 조성합니다. 관리인력이 부재할 경우, 안전 관리 시스템이 미흡해지고 사고 발생 가능성이 높아질 수 있습니다. **2. 안전사업장으로 판단된 근거:** * **최근 재해 발생률 및 상해사건 발생 경력이 없음:** 이러한 기록은 사업장이 안전 관리에 노력하고 있다는 것을 보여줍니다. * **위험물질 취급 및 대량위험물 제조 여부가 없음:** 위험물질은 사고 발생 가능성을 높이는 요소입니다. 이러한 위험 요소가 없는 것은 안전성을 높이는 요인입니다. * **밀폐공간 및 소음 작업 외에 다른 위험 작업 환경이 없는 것:** 밀폐공간과 소음 작업은 주의가 필요한 작업 환경이지만, 다른 위험 요소가 없는 것은 안전성을 높이는 요소입니다. * **가스농도측정기를 보유하고 있음:** 밀폐공간에서 작업 시 가스농도를 측정하고 관리하는 것은 안전을 위한 중요한 조치입니다. **3. 위험요소:** * **크레인(천장, 갠트리) 사용:** 크레인은 무거운 물체를 이동시키는 장비로, 제대로 된 안전 교육과 점검 없이 사용될 경우 끼임, 낙하 등의 사고로 이어질 수 있습니다. * **소음 작업:** 장시간 소음 작업은 청력 손상을 유발할 수 있습니다. 소음 차단 장비를 사용하고, 소음 노출 시간을 제한하는 등의 조치가 필요합니다. * **안전 교육 수료율 및 관리자 업무 이해도 부족:** 안전 교육의 중요성을 강조하고, 관리자의 안전 관련 교육을 통해 안전 관리 시스템을 강화해야 합니다. **4. 개선방안:**

구분	언어모델 출력 결과
	* **안전 교육 강화:** 모든 근로자에게 정기적인 안전 교육을 실시하고, 교육 수료율을 높이기 위한 노력이 필요합니다. 특히 크레인 사용 안전 교육을 강화해야 합니다. * **관리자 안전 교육:** 안전보건 관리자를 지정하고, 안전 관련 법규, 안전 관리 시스템, 위험 요소 관리 등에 대한 교육을 실시하여 안전 관리 역량을 강화해야 합니다. * **안전보건 관리인력 확보:** 안전보건 관리인력을 확보하여 안전 점검, 교육, 위험 요소 관리 등을 체계적으로 수행해야 합니다. * **소음 관리:** 소음 작업 시 소음 차단 장비를 사용하고, 근로자의 소음 노출 시간을 제한하는 등의 조치를 취해야 합니다. * **정기적인 안전 점검:** 크레인 및 기타 작업 장비의 정기적인 점검을 통해 안전성을 확보해야 합니다. * **안전보건 시스템 구축:** 안전보건 관련 법규를 준수하고, 안전 관리 계획을 수립하여 안전한 작업 환경을 조성해야 합니다.
제조업 Prompt-05	## 사업장 위험도 분석 결과 및 근거 해당 사업장의 위험사업장 확률은 40.05%로 안전사업장에 가까운 수치입니다. 하지만, 몇 가지 위험 요소와 개선 가능한 부분들이 존재합니다. **1. 위험사업장으로 판단된 근거:** * **낮은 안전교육 수료비율:** 안전교육은 근로자의 안전 의식을 높이고 위험 상황에 대처하는 능력을 향상시키는 데 필수적입니다. 낮은 수료비율은 안전 사고 발생 가능성을 높일 수 있습니다. * **안전보건 관리인력 부재:** 안전보건 관리인력은 사업장 내 안전 관리 시스템을 구축하고 운영하며, 안전 교육 및 점검을 담당합니다. 관리인력 부재는 안전 관리 체계의 약점으로 이어질 수 있습니다. * **중간 수준의 안전검사, 안전교육 관리자 업무 이해도 점수:** 안전검사 점수, 안전교육 관리자 업무 이해도 점수가 중간 수준으로 나타났습니다. 이는 안전 관리 시스템 및 안전 교육의 질적 수준이 충분하지 않음을 시사합니다. **2. 안전사업장으로 판단된 근거:** * **최근 재해 발생률 및 상해사건 발생경력 부재:** 최근 재해 발생률과 상해사건 발생 경력이 없는 것은 안전 관리 노력의 효과를 보여주는 지표입니다.

구분	언어모델 출력 결과
	* **위험설비 종류가 지게차로 제한적:** 지게차는 안전 교육 및 운영 규정 준수가 중요하지만, 석유화학물질이나 고위험 설비를 사용하지 않는다는 점에서 위험 요소가 상대적으로 적습니다. * **위험물질 취급 및 밀폐공간 부재:** 위험물질 취급 및 밀폐공간 부재는 작업 환경의 위험성을 크게 낮추는 요소입니다. **3. 위험요소:** * **낮은 안전교육 수료비율:** 안전 교육 부족으로 인해 근로자들이 위험 상황에 대처하는 데 어려움을 겪을 수 있습니다. * **안전보건 관리인력 부재:** 안전 관리 시스템 구축 및 운영 부족으로 인해 안전 사고 예방 및 대응에 어려움을 겪을 수 있습니다. * **중간 수준의 안전 관리 지표:** 안전검사, 안전교육 관리자 업무 이해도 점수가 중간 수준으로 안전 관리 시스템의 질적 부족을 시사합니다. **4. 개선방안:** * **안전교육 강화:** 안전 교육 프로그램을 정기적으로 실시하고, 참여율을 높이기 위한 노력이 필요합니다. 특히, 지게차 안전 운전 교육을 강화하는 것이 중요합니다. * **안전보건 관리인력 확보:** 안전보건 관리인력을 채용하여 안전 관리 시스템을 구축하고 운영하는 것이 필수적입니다. * **안전 관리 시스템 개선:** 안전검사, 안전교육 관리자 업무 이해도 등 안전 관리 지표를 향상시키기 위한 노력이 필요합니다. * **근로자 참여 유도:** 안전 관리에 대한 근로자들의 참여를 유도하고, 안전 문제를 신속하게 보고할 수 있는 시스템을 구축하는 것이 중요합니다.

3) LoRA 학습 적용

○ 본 연구에서는 XGBoost 모델의 고위험 사업장 확률 점수와 그에 대한 다양한 사업장 상황 정보를 결합하여 LoRA기법으로 언어모델을 학습시킴. 모델이 각 사업장의 위험 요소와 상황별 특성에 대해 보다 구체적이고 맞춤형 설명을 할 수 있도록 다음과 같이 적용함.

1. XGBoost 확률 점수와 상황 정보의 결합

- XGBoost 모델은 각 사업장의 위험도를 확률 점수로 예측함. 이 점수는 사업장이 고위험인지 저위험인지를 평가하는 데 중요한 기준이 됨.

- 이와 함께, 사업장의 상황 정보(예: 설비 종류, 작업 환경,근무 환경 등)을 결합하여 언어모델 학습 데이터로 활용함.

- 고위험 확률이 높은 사업장일수록 위험 요소에 대한 구체적인 설명이 필요하므로, 상황 정보와 함께 이러한 확률 점수를 모델에 학습시키는 방법임.

2. LoRA 학습 적용

- LoRA 기법을 통해, 사전 학습된 언어모델의 일부 파라미터만 미세조정하여 고위험사업장에 특화된 모델로 학습하였음. 이로 인해 언어모델이 일반적인 상황에서 응답하는 것보다, 고위험 상황에 맞춤형 대응을 할 수 있도록 최적화함.

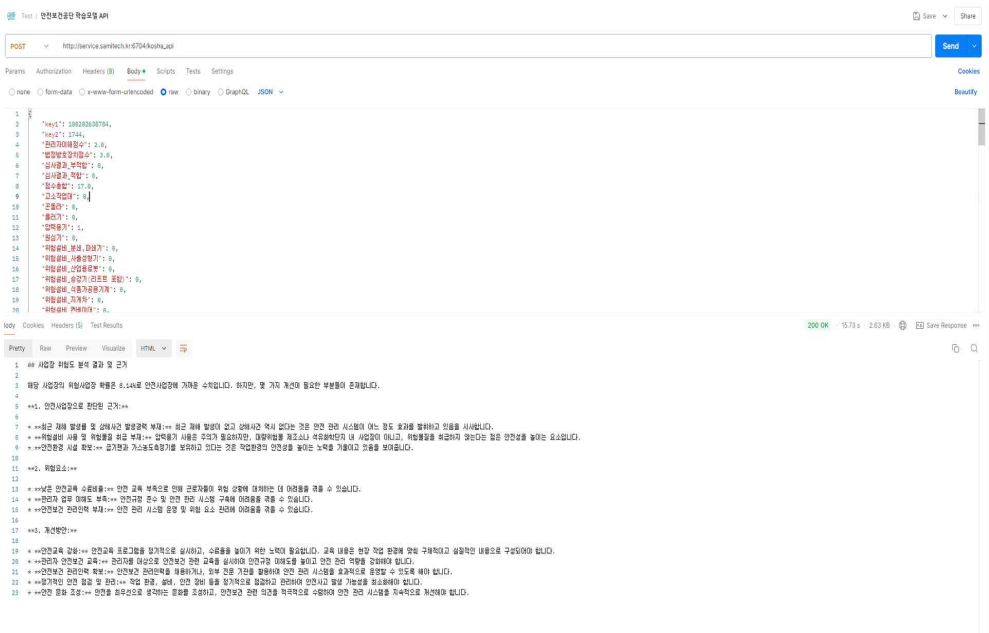
- 특히 저장 공간과 연산 자원을 절약하면서도 모델의 표현력을 높일 수 있는 LoRA의 특성으로 약 40만 건 데이터로 효율적으로 학습하였음.

3. 추론 요청 시 차이점

- 기존 모델에서 프롬프트를 기반으로 추론을 요청할 경우, 모델은 일반적인 답변을 제공하게 됨. 이는 다양한 상황 정보를 학습하지 않은 상태이기 때문에 정확한 해석이 어려울 수 있음. 그러나 충분한 데이터로

LoRA로 미세 조정된 모델은 학습된 데이터에 맞춰 답변을 제공함.

- 예를 들어, 새롭게 주어진 상황 정보가 특정 위험 요소와 관련이 있는 경우, 모델은 학습된 데이터를 바탕으로 해당 상황에 대한 구체적인 위험 분석을 제공할 수 있으며, 위험 확률 점수와 결합된 정보는 모델이 왜 특정 사업장이 고위험인지 또는 안전한지에 대한 구체적인 이유와 설명을 자연어로 제시할 수 있게됨.



[그림 3-6] 언어모델 LoRA 튜닝 결과 API 구성 및 테스트 결과

4) 기존 방법과의 비교

○ 이번 연구에서는 안전보건공단의 기존 고위험사업장 예측 모델을 개선하기 위한 실험을 진행하였음. 기존 모델은 XGBoost를 활용하여 위험 사업장 예측 점수를 산출하고, Feature Importance, SHAP을 통해 예측에 기여한 특성의 영향력과 변수명 및 수준 정보를 제공하는 방식으로 이루어짐.

○ 그러나, 본 연구에서는 보다 설명력과 접근성을 높이기 위해 최적화된 모델을 활용하여 위험사업장 예측 점수를 제공하는 방식은 유지하되, Feature Importance나 SHAP과 같은 수치적 해석 대신, 언어모델을 통해 비전문가들도 이해할 수 있는 방식으로 위험 요인과 개선 방안을 설명하는 방법을 제안함. 이를 통해 고위험 사업장으로 판단된 이유와 개선사항을 텍스트로 제공하여, 비전문가도 모델의 예측 결과를 쉽게 이해하고 활용할 수 있도록 함.

〈표 3-25〉 언어모델 및 수치적 해석의 특장점 비교 표

구분	특장점
언어모델을 통한 해석	• **비전문가 이해 가능**: 언어모델은 결과를 자연어로 설명하여 비전문가도 쉽게 이해할 수 있도록 제공함. 예를 들어, “이 사업장은 안전 규정 준수율이 낮아 위험합니다”와 같은 직관적인 설명 제공 • **구체적인 개선 방안 제시**: 위험 요인을 설명하면서, 이를 개선하기 위한 구체적인 조치나 주의사항 제시
Feature Importance 해석	• **정량적인 근거:** 각 특성이 모델 예측에 기여한 정도를 수치적으로 표현하여 객관적인 근거 제공. • **일관성 및 재현성:** 계산에 기반한 해석 방식이기 때문에, 결과가 일관되며 재현 가능함.
SHAP 해석	• **개별 예측 해석 가능:** 사업장 단위로 개별 예측 결과에 대해서도 기여도를 분석할 수 있어, 개별 사업장의 위험 요인을 더 세밀하게 분석 가능함. • **기여도 산출:** 게임 이론을 기반으로 각 특성이 예측에 기여한 정도를 정확하게 계산하므로, 모델의 예측 결과를 보다 투명하게 제공

○ 해석 방법의 한계점 또한 중요하며, 각각의 해석 방법이 가지고 있는 한계는 해석의 완성도를 떨어뜨릴 수 있으므로, 이를 보완할 수 있는 방안도 전략도 함께 고려되어야 함.

〈표 3-26〉 언어모델 및 수치적 해석의 한계점 비교 표

구분	한계점
언어모델을 통한 해석	• **일관성 부족:** 언어모델은 같은 입력 데이터에 대해서도 출력이 일관되지 않을 수 있음. • **불확실성:** 언어모델은 학습 데이터에 강하게 의존하기 때문에, 때로는 부정확하거나 불확실한 정보를 제공할 가능성 있음
Feature Importance 해석	• **직관성 부족**: 수치적 정보는 비전문가에게 직관적으로 이해하기 어려움 • **맥락 부족**: 중요한 특성은 설명하지만, 왜 그 특성이 중요한지에 대한 맥락 제공 없음
SHAP 해석	• **계산 복잡도**: 계산량이 많고 대규모 데이터에서 시간이 오래 걸릴 수 있음 • **이해 어려움**: 비전문가에게는 SAHP의 시각적 해석이 복잡할 수 있음

○ 향후에는 모델 훈련 시 다양한 상황정보를 더 포함시키고, 예방대책 등의 정보를 포함한 데이터셋을 학습시켜, 특정 입력에 대해 더 일관되고 다양한 정보를 제공할 수 있도록 해야 함. 또한 학습 데이터 검증 및 사후 검증을 통해 잘못된 정보를 걸러낼 수 있는 필터링 메커니즘의 도입이 필요함.

Ⅳ. 결론 및 제언

Ⅳ. 결론 및 제언

1. 결론

1) 산업안전 데이터 수집 및 전처리

○ 이번 연구에서는 산업안전 데이터를 수집하고, 전처리 과정을 통해 모델 학습에 적합한 데이터셋을 구성하는데 중점을 두었음. 이를 위해 모델에서 처리되는 이상치 처리, 코드값 변환, 결측치 처리 등 기본적인 전처리 기법 외에도, 사업장 데이터의 특성에 맞춘 추가 전처리 방법들을 적용하여 데이터를정제하였음. 이러한 정제된 데이터셋은 모델의 기본적인 성능확보와 학습을 수행하는 필수적인 단계로 작용하였음.

2) 기존 고위험사업장 선정 모델 및 데이터 분석

○ 기존의 고위험사업장 선정 모델은 XGBoost 알고리즘을 사용하여 Confusion Matrix를 기반으로 정밀도, 재현율, F1 Score 등의 지표를 통해 모델의 성능을 평가하였음. 초기 모델 학습 및 데이터를 분석하여 안전사업장과 고위험사업장의 불균형, 특성값의 불균형, 특성간 상관관계 등이 확인되었으며, Feature Importance와 SHAP 분석을 통해 모델이 일부 특성에 과도하게 의존하는 경향도 확인하였음.

○ XGBoost 외에도 LGBM(LightGBM), CatBoost, 그리고 딥러닝 모델을 추가로 적용하여 다른 모델에서도 데이터 학습 성능을 확인하였음. 최종적으로 LGBM 모델이 정밀도, 재현율이 균형이 있으며 전반적으로 높은 성능을 보임.

3) 고위험사업장 선별 모델 설계 및 개발

○ 데이터의 적용 방식에 따라 모델 성능의 변동을 확인하기 위해 산재발생 데이터와 위험 수준 평가 데이터를 각각 또는 결합하여 모델을 학습하였음. 그 결과, 전체 데이터셋으로 학습한 것보다 성능이 향상되었음을 확인하였고, 또한 분석을 통해 상위 50개의 주요 특성만을 적용하여 학습했을 때, 성능이 더 개선되는 결과를 얻었음. 학습에 필요한 주요 특성의 선별과 데이터셋의 품질이 향후에도 중요한 역할을 할 수 있음을 확인함.

4) 산업안전 도메인 맞춤형 언어모델 개선 및 모델 성능 비교 분석

○ 기존의 수치적 해석 방식 대신, 본 연구에서는 언어모델을 도입하여 비전문가도 쉽게 이해할 수 있는 설명 방식을 개발하였음. 고위험사업장 선정 모델에서 산출한 예측 결과를 기반으로, LoRA(Low-Rank Adaptation)기법을 적용하여 학습된 언어모델을 통해 고위험사업장 판단 이유, 위험요인, 개선사항 등을 자연어로 설명함으로써 접근성과 설명력을 높임.

○ 언어모델을 활용한 해석 방식은 비전문가도 이해할 수 있는 텍스트 기반 설명을 제공하여 현장 적용 가능성을 높였으며, 이는 Feature Importance나 SHAP과 같은 수치적 해석과 함께 보완할 수 있는 도구로 활용될 수 있음.

5) 데이터 기반 감독·점검체계 구축 지원

○ 본 선정 모델은 기존의 사업장 감독·점검 시 생성한 데이터를 활용하여 데이터 기반 고위험 사업장 선정 모델을 통하여 실제 산업재해가 발생할 가능성이 얼마나 되는지, 발생할 수 있는지, 또는 사업장이 어느 정

도의 산재위험에 노출되었는지를 판단하는데 활용할 수 있도록 하였음.

○ 본 연구의 목적인 데이터 기반의 감독·점검을 위한 기본 프레임워크로 [1]사업장 선정·배포, [2]지방관서 지시.감독 실시, [3]결과 분석·평가, [4]다음연도 계획 수립하는 데이터 기반 감독·점검체계 구축을 완성할 수 있도록 지원함.

2. 제언

1) 데이터의 양적 및 질적 확보

○ 모델 성능을 더욱 향상시키기 위해서는 양질의 데이터 확보가 필수적임. 특히 산재 발생 이력이 부족하거나 산재에 관련성이 떨어지는 특성, 결측된 데이터는 예측 성능을 저하시킬 수 있으므로, 많은 데이터를 확보할 수 있는 방안이 필요함. 특히, 사업장 평가 기준을 세밀하게 설정하고 실제 현장에 맞는 평가를 하였는지 지속적인 검토 방안이 필요함. 또한 시간에 따라 변화를 파악할 수 있는 시계열 데이터의 확보 또한 중요하며, 모델이 장기적인 추세를 반영하고, 더 정밀한 예측을 할 수 있음.

2) 라벨링 기준 개선

○ 라벨링의 기준 설정은 모델의 예측 성능에 큰 영향을 미칠 수 있음. 안전/고위험 사업장으로 분류되는 기준이 명확하지 않거나 모호할 경우, 모델 성능의 저하로 이어질 수 있음. 조건에 부합하지 않거나 부여할 수 없는 데이터에 대한 처리 기준을 정립하여 불확실성을 최소화하고, 데이터의 일관성을 유지해야 함.

3) 언어모델을 통한 현장 적용 확대

○ 이번 연구에서 사용된 언어모델은 비전문가도 쉽게 이해할 수 있는 방식으로 고위험 사업장 판단 근거를 제공함으로써 현장 적용 가능성을 높였음. 향후 더 다양한 언어모델과 인프라를 도입하여, 각 사업장의 궂레적 상황에 맞춘 맞춤형 해석을 제공할 수 있도록 발전시킬 필요가 있음. 또한 의사결정 지원 도구로서 언어모델을 활용하여 감독관, 기업 관리자, 근로자가 모두 이해하고 활용할 수 있는 의사소통 도구로 발전시킬 수 있음.

4) 비정형 데이터 활용 체계 구축

○ 비정형 데이터는 사업장의 텍스트 기록, 보고서, 현장 평가 내용 등을 포함하며, 이를 효과적으로 처리할 수 있는 시스템 및 기존 자료를 전산화할 수 있는 방법이 필요함. 비정형 데이터의 구조화 및 분석을 통해 더 많은 데이터 자산을 활용할 수 있으며, 모델 성능을 한층 더 끌어올릴 수 있음. 특히 AI기반 자연어 처리(NLP) 기술을 활용하여 비정형 데이터를 정형 데이터로 변환함으로써, 보다 정교한 분석과 예측이 가능해질 것임.

5) 고위험사업장 선별 모델의 활용

○ 데이터 기반 감독·점검체계 구축을 완성하기 위하여 산재발생 현황 모니터링 웹페이지 개발 시 분석 및 시각화 데이터로 활용

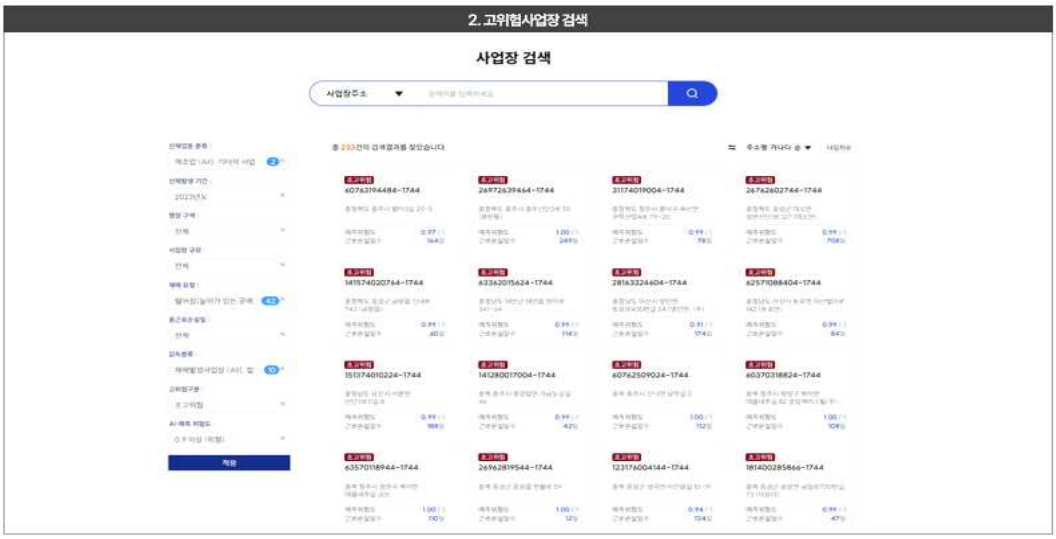
○ 다음은 산재데이터, 고위험사업장 선별 모델, 데이터 분석결과 등을 산재발생 현황 모니터링 웹페이지를 통하여 '고위험사업장 디지털맵', '고위험사업장 정보 검색', '사업장별 산재정보 시각화', '고위험사업장 상세정보'. '사업장 위험도 분석 결과 및 근거 리포트 생성' 등 활용예임.

(1) 고위사업장 디지털맵

- 사업장 위치 정보와 함께 요약 중심 사업장별 산재정보를 미리 볼 수 있도록 제공

(2) 고위사업장 정보 검색

- 산재업종분류, 산재발생 기간, 행정구역, 사업장규모, 재행 유형, 총근로손실일, 감독분류, 고위험구분 등 사업장별 산재정보 검색 제공



(3) 사업장별 산재정보 시각화

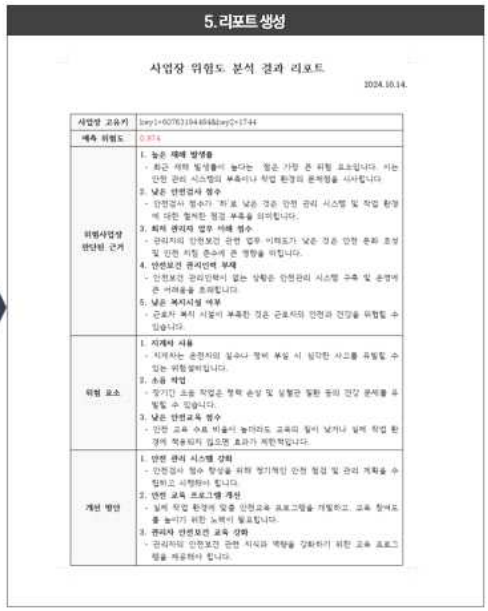
- 사업장 기초정보, 고위험 사업장 여부, 고위험사업장 예측점수, 사업장 감성평가 및 3년간 평균 재해율 등을 시각화 정보로 제공

(4) 고위사업장 상세정보

- 정보 속성별로 안전 및 보건관리, 작업환경, 재정지원 및 투자, 인증 및 평가, 교육 및 훈련, 민간위탁 기술지도로 구분하여 정보 제공

(5) 사업장 위험도 분석 결과 및 근거 리포트 생성

- 사업장별 예측 위험도, 위험사업장 판단 근거, 위험 요소, 개선 방안 등을 리포트 형태로 제공



참고문헌

고용노동부, 인공지능 알고리즘을 활용한 재해개요 분류모델 시범 개발, 2023.

안전보건공단, 사업장 정량정보를 활용한 산재 고위험사업장 선별 효과성평가 및 개선방안 연구, 2022.

안전보건공단, 한국의 산업별 산업재해 발생 추이와 경기적 영향요인 연구, 2021.

안전보건공단, 직종별 특성과 사망사고 발생 위험분석 연구(Ⅰ), 2021.

Tianqi Chen, Carlos Guestrin, XGBoost: A Scalable Tree Boosting System. 2016.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 2017.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, Andrey Gulin. CatBoost: Unbiased Boosting with Categorical Features. 2017.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. 2021.

연구진

연 구 기 관 : 주식회사 사미텍

연구책임자 : 김재두 (연구소장, ㈜사미텍)
연 구 원 : 김기형 (선임연구원, ㈜사미텍)
연 구 원 : 우은경 (연구원, ㈜사미텍)
연 구 원 : 김수경 (감사, ㈜사미텍)
연 구 원 : 이형용 (선임연구원, ㈜사미텍)
연구보조원 : 양지호 (연구원, ㈜사미텍)
연구보조원 : 박주연 (연구원, ㈜사미텍)
연구보조원 : 오민근 (연구원, ㈜사미텍)
연구보조원 : 김혜민 (대표이사, ㈜사미텍)

연구상대역 : 박재석 (선임연구위원, 안전연구실)

연구기간

2024. 04. 29. ~ 2024. 10. 31.

본 연구는 산업안전보건연구원의 2024년도 위탁연구 용역사업에 의한 것임

본 연구보고서의 내용은 연구책임자의 개인적 견해이며,
우리 연구원의 공식견해와 다를 수도 있음을 알려드립니다.

산업안전보건연구원장

고위험 사업장 선정 모델 개선을 통한 감독·점검 효과성 제고방안 연구

(2024-산업안전보건연구원-504)

발 행 일 : 2024년 10월 31일
발 행 인 : 산업안전보건연구원 원장 박승현
연구책임자 : ㈜사미텍 연구소장 김재두
발 행 처 : 안전보건공단 산업안전보건연구원
주 소 : (44429) 울산광역시 중구 종가로 400
전 화 : 052-703-0841
팩 스 : 052-703-0334
Homepage : http://oshri.kosha.or.kr
I S B N : 979-11-94453-14-7
공공안심글꼴 : 무료글꼴, 한국출판인회의, Kopub바탕체/돋움체