# Assignment 2

(산대특)_빅데이터 분석 기반 AI 알고리즘 개발 과정_육성 과정 **김종현, 김한열, 박지환, 이진규 훈련생**

## 1. 연구 개요

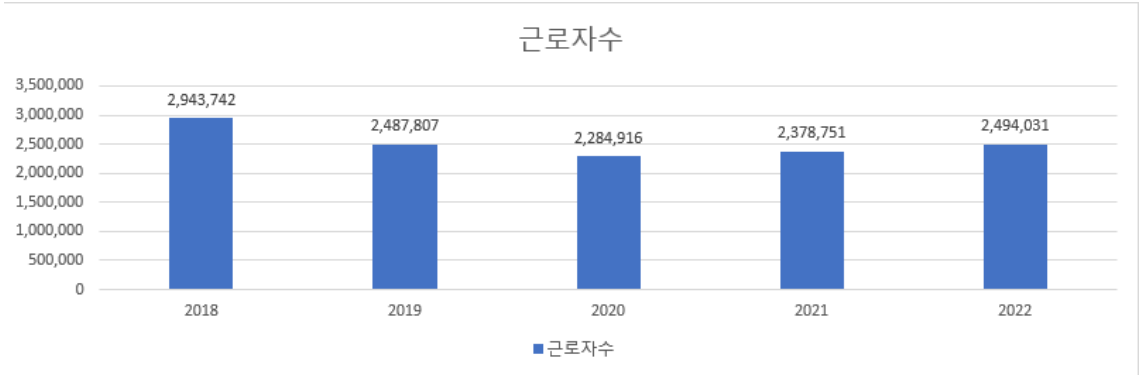
### 1) 연구의 배경 및 목적

고용노동부의 「산업재해 현황분석」에 따르면 산업재해 발생 현황은 1988년부터 2003년까지 증가 추세를 보이다가 2003년부터 2017년까지 꾸준히 감소 추세를 보였으나 2017년 이후 다시 증가 추세를 보인다. 근로자의 안전을 위하여 다양한 정책을 실시하고 있으나 아직 현장에서 벌어지는 사고 예방의 효과가 미비한 것으로 보인다.

2018~2022년까지 5년 동안 전체 산업군(산업군 중분류 기준)의 평균 사업 장수는 2,654,107개소이며, 근로자 수는 19,073,438명, 그 가운데 요양재해자 수는 102,305명으로 요양재해율은 0.54%이다. 전체 산업군 중에서 제조업과 건설업의 사업장 수는 전체 대비 각각 394,201명(14.17%), 389,398명(14.00%)를 차지하고 있으며, 요양재해자 수는 제조업 29,751명(25.96%), 건설업 28,577명(24.94%)으로 5년 평균 요양재해자 수(114,597명)의 약 절반가량을 차지하고 있다. 요양재해율의 경우, 5년 평균 0.59%에 비해 제조업 1.24%, 건설업 1.92%로 약 2~3배 이상의 수치를 보인다.

근로자의 안전을 보장하기 위하여 2022년 중대재해처벌법이 시행되었지만 안전보건 관리공단의 「통계로_보는_2022년_산업재해」에 따르면 2022년 전체 업종에서 발생한 사고사망자 874명 중 약 절반가량인 402명이 건설업에서 발생하였고, 두 번째로 높은 184명이 제조업에서 발생하였다. 중대재해처벌법의 기준인 50인 이상 사업장에서 사고사망자는 총 874명 중 167명(19.11%)이 발생한 반면 해당하지 않는 50인 미만 사업장에서 총 874명 중 707명(80.89%)이 발생하였다.

이러한 산업재해를 대표적으로 보여주는 것이 2022년 발생한 광주 화정 아이파크 붕괴 사고(2022.1.11)이다. 날씨를 고려하지 않은 무리한 공사 일정으로 발생한 사고로 사회에 큰 파장을 불러일으켰다. 여러 산업 중 빈번한 산업재해의 발생과 큰 인명피해를 동반하는 건설업에 산재 지수를 제공 및 부여하여 근로자들이 안전한 근무지를 선택할 수 있도록 하고, 기업은 구인을 위해 자발적인 안전관리를 진행하도록 유도하며 산업재해로 발생하는 사회적 비용을 감소시키고자 본 연구를 진행하였다.




출처

## 2) 연구 범위

| |
|---|
| 산업 : 건설업 |
| 기간 : 6 |
| 지역 : |

여전 data 수집중?

# 2. 연구의 목표 및 내용

## 1) 연구 목표

건설업 혹은 현장이라는 용어 넣을것.

| |
|---|
| 특정 상황 및 현장에 대한 위험을 수치화하여 산업재해 예방에 기여한다. |

## 2) 연구 내용

(1) 산업재해 발생에 영향을 미치는 요인을 선정한다.

- 자료의 기술 통계량을 파악하기 위해 빈도분석을 수행한다.

(2) 연구에 사용된 요인의 신뢰도를 확인하기 위해 요인분석을 시행한다.

- 신뢰도 검증은 샤피로 윌크검정을 수행한다.

(3) 선정된 요인과 신뢰도 검증을 토대로 연구에 대한 변수를 재구성한다.

(4) 요인별로 산업재해 발생확률을 추정한다.

- 여러 모델을 구축한 후 비교·분석하여 가장 좋은 성능의 모델을 선택한다.

추후 process 도식화.

# 3. 연구 설계

## 1) 연구 문제

선행연구를 통하여 건축을 포함한 산업 전반에서 산업재해 발생 및 그 위험도를 예측하는 다

수의 연구가 진행되었다는 사실이 확인되었다. 이에 따라 산재 위험에 대한 지수를 사고 규모(경상, 중상, 사망 등)에 따라 몇 종류의 범주로 분류하여 각각에 대하여 수치화할 수 있는지를 본 연구에서 확인할 예정이다.

## 2) 측정 및 자료 수집

건축업에서 발생한 산업재해에 관한 상세한 데이터는 국토교통부 산하의 건설공사 안전관리 종합정보망에서 각 산재 사건 단위로 상세한 데이터를 제공하고 있다. 데이터 다운로드는 엑셀로 정리된 데이터로 6개월 단위로 다운로드가 가능하다. 해당 데이터는 주요 내용은 아래와 같다.

| 컬럼명 | 데이터유형 | 컬럼명 | 데이터유형 | 컬럼명 | 데이터유형 |
|---|---|---|---|---|---|
| 사고명 | 명목형 | 발생일시 | Datetime | 공공/민간 | 명목형 |
| 기상-날씨 | 명목형 | 기상-기온 | 연속형 | 기상-습도 | 연속형 |
| 시설물종류 | 명목형 | 개인보호조치 | 명목형 | 공종 | 명목형 |
| 사고장소 | 명목형 | 사망자수 | 연속형 | 부상자수 | 연속형 |
| 사고원인 | 명목형 | 공사비 | 명목형 | 작업자수 | 명목형 |

사고가 발생하지 않은 대조군 데이터는 고용노동부에서 발표하는 전체 사업 진행 / 노동자 데이터에서 추출할 예정이며, 산업재해가 발생한 데이터와 대조하여 사고 발생률 등의 수치를 산출할 수 있다. 건설 환경에 대한 데이터는 기온과 습도가 제공되나 풍향 등의 추가적인 데이터가 필요할 경우 기상청에서 제공하는 데이터를 바탕으로 날짜와 위치 데이터를 이용하여 환경 데이터를 추가할 수 있다.

## 3) 분석 방법

· 데이터 수집 및 파악 : 상기한 데이터 수집 방법을 바탕으로 데이터를 수집하고, 수집한 데이터가 연구의 방향성과 맞는 데이터인지 확인하고 적합성을 검증한다. 탐색적 데이터 분석을 수행하여 데이터에서 확인할 수 있는 특징을 찾고, 이후 연구 방향 결정을 결정한다. 연구 방향 결정에는 종속변수를 무엇으로 선정할 것인지가 포함된다.

· 데이터 전처리 및 추가 수집 : 추가적인 탐색적 데이터 분석이나 모델 훈련의 가시성 및 성능 향상을 위한 데이터 전처리 작업을 진행한다. 데이터 전처리 작업은 한 번에 국한하여 시행하지 않고 데이터 분석 및 모델 훈련 과정에서 필요하다고 판단되면 여러 번 수행할 수 있다. 처리를 진행한 데이터로도 연구 진행에 지장이 있다면 데이터를 추가 수집하여 데이터의 양을 늘리거나 새로운 인사이트를 도출할 수 있는 추가적인 피쳐를 탐색한다.

· 모델 훈련 및 평가 : 탐색적 데이터 분석의 결과와 전처리를 진행한 데이터를 바탕으로 학습을 진행할 머신러닝 모델과 그 기법을 선정한다. 데이터 확인 과정에서 결정한 종속변수를 높은 정확도로 예측할 수 있는 머신러닝 모델을 완성하기 위하여 여러 가지 방법의 데이터 전처리, 모델 선정, 하이퍼파라미터 튜닝, 앙상블을 포함한 다양한 모델링 기법을 사용한다. 수집한 데이터는 시계열 데이터를 포함하고 있으므로 모델 성능 검증은 과거의 데이터로 미래의 산업재해 발생을 예측하는 방식으로 진행한다.

· 최종 모델 선정 및 보고서 작성 : 상기한 과정을 통하여 최종 모델 훈련을 완료하고 이를 바탕으로 연구 결과를 보고서로 작성한다. 모델 훈련을 통하여 산출한 산재 위험 지수에 대한 정확도 및 전문성을 검증하기 위해서는 유사한 연구를 진행한 전문적인 논문 등을 참조하여 연구에 대한 근거를 제시해야 한다.

## 4) 연구 추진 일정

계획이지 일정은 아닌.

※ 데이터 전처리

수집한 데이터는 분석을 위한 적합한 형태로 데이터 전처리를 진행한다. 아래는 작업이 진행될 것으로 예상되는 몇 가지의 전처리 프로세스와 이를 적용할 특성을 나열한 것이다.

- 명목형 데이터에 대한 재군집화 : 고유한 값들이 많은 데이터에 대하여 일정한 기준을 확인하여 해당 기준대로 데이터를 재분류하여 새로운 명목형 데이터를 만들어 데이터 분석에 대한 가시성과 모델 성능 향상을 기대할 수 있다. (적용 가능한 데이터 예시 : 사고 명과 피해 내용을 바탕으로 사고를 경상, 중상 등으로 재분류)

- 연속형 데이터에 대한 명목형 데이터 변환 : 직관적이지 않은 연속형 데이터를 특정 기준에 따라 명목형 데이터로 가공하여 파생 변수를 생성할 수 있다. 해당 기준은 선행연구나 비지도 학습의 군집화를 통하여 선정할 수 있다. (적용 가능한 데이터 예시 : 습도 데이터를 다습, 건조 등의 명목형 데이터로 재분류)

- 스케일링, 라벨링을 포함한 머신러닝을 위한 기초적인 데이터 가공은 필요에 따라 진행한다. 이상치 / 결측치 처리와 같이 데이터의 특성 및 서열이 바뀔 수 있는 경우에는 선행연구 등 해당 데이터 가공에 대한 타당성을 확인한 후 진행하며, 이를 확보하지 못할 시 데이터 정제보다는 추가 데이터 수집을 진행하는 방향으로 진행한다.

※ 모델 훈련 및 평가

모델 훈련을 위해서는 최종적으로 도식화하기 위한 종속변수를 어떤 것으로 할지를 확정한 후 진행한다. 종속변수에 따라 사용할 모델을 선정하고 모델 훈련을 진행한다.

- 가용한 모델을 최대한 많이 선정하여 모델 학습을 진행하고 성능이 우수한 모델을 선정한다. 모델별로 우수한 성능을 내는지 확인하기 위하여 검증 데이터는 별도로 분리하여 모델 성능을 검증한다. 데이터는 날짜 데이터를 포함하고 있으므로 Timeseries split 기법으로 과거의 데이터로 미래의 산재를 예측하는 방법을 사용할 수 있으며, 그 외에도 산재 유형별로 데이터를 분리하여 유형별로 모델 성능을 측정하여 모델 튜닝에 활용하는 방법 등을 사용할 수 있다.

- 추가적인 모델 성능 향상을 위하여 앙상블 기법으로 여러 가지 모델을 조합하여 새로운 모델을 생성한다. 앙상블로 생성된 모델은 단일 모델 중 성능이 가장 우수한 모델과 비교하며, 필요에 따라 앙상블 기법은 사용하지 않을 수 있다. 가능할 경우 딥러닝 모델을 사용하는 방안도 고려할 수 있다.

- 모델 성능 평가를 위해서는 연구 방향에 알맞은 올바른 평가지표 선정이 선행되어야 한다. 산업재해 발생을 예측하는 모델의 경우 실제로 산업재해가 발생할 가능성이 큰 고위험군을 위험하다고 판단하여야 하는 것이 무엇보다도 중요하기 때문에 양성 데이터를 실제 양성으로 예

측하는 비율인 재현율에 관한 지표는 반드시 확인해야 한다.

※ 연구과정에서 발생할 것으로 예상되는 문제와 해결 방법

- 데이터 부족 : 데이터가 부족한 적합한 모델 성능을 확보하지 못하였을 경우 추가적인 데이터 수집을 진행한다. 같은 출처에서 구할 수 있는 데이터나 타 정부 부처나 기업체에서 제공하는 데이터를 기존 데이터에 추가하여 활용한다. 그럼에도 데이터가 부족할 경우 해외에서 발생한 건축 산업재해 데이터도 사용할 수 있다.

- 데이터 병합 이슈 : 데이터의 출처가 파편화되어 있으면 데이터 병합 과정에서 join의 기준이 되는 id가 없을 경우가 발생할 수 있다. 이 경우 병합하려는 데이터끼리의 공통된 요소를 최대한 발견한 후 해당 값을 기준으로 한 통계값 (평균값, 중앙값 등)을 이용하여 데이터를 병합한다. raw 데이터의 단일 항목을 기준으로 한 병합이 아니므로 데이터 중복이 필연적으로 발생하며 성능 또한 반드시 향상된다고 보장할 수 없으므로 여러 가지의 데이터 병합 시도하거나 다른 출처의 데이터를 탐색하는 방향으로 연구를 진행할 수 있다.

# 4. 선행연구의 내용 및 결과

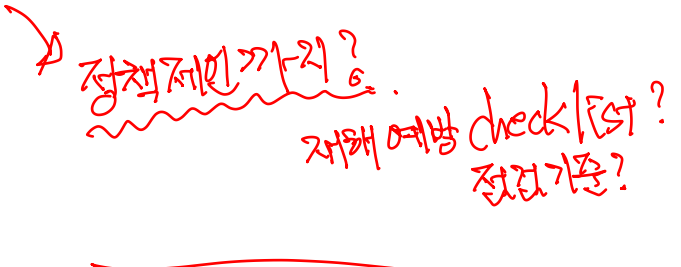
- 산업재해 발생의 상대위험도 분석 및 순환분포 모형 추정(김학열, 허태영)
  - 주요내용: 재해 발생의 패턴을 다양한 유형(산업별, 재해유형별, 발생 시간별, 입사 근속기간별 등)으로 구분하여 비교·분석하고, 시간대별 발생 건수를 이용하여 산업재해 발생에 관한 통계적 순환모형을 개발
  - 결론: 광업과 제조업이 고위험도 산업으로 분류, 제조업은 상대적으로 낮은 사망 상대위험도, 입사 근속이 아주 짧거나 긴 근로자가 고위험군에 속함, 새벽 및 아침 시간대에서 상대적 높은 위험도
  - 한계점: 전체 산업 업종 중 부상자와 사망자만을 대상으로 하여 구체적으로 어떤 형태의 사고로 부상과 사망이 일어났는지를 알 수 없다. 사고가 일어났을 때의 사업장의 상황과 근로자의 상황에 대한 자료가 부족하다.

- 건설공사의 정량적 위험도 산정 방법론(김현수, 이현수, 박문서)
  - 주요내용: 기존의 재해지표에 대한 한계점을 분석한 뒤, 위험도 산정에 필요한 변수들을 설정하고, 이를 조합하여 정량적 위험도 산정 방법론을 제시
  - 결론: 공정별 상대 비교를 통해 위험 강도와 빈도 각각의 크기를 구한 뒤 기하평균으로 위험도 산정
  - 한계점: 근로자 수, 실제 근로시간. 근로 손실일 수, 보험 지급액 등의 자료를 활용하여 위험도 산정과정에 반영하여 더 정확한 공정별 위험도 산정 가능

- 건설 현장 정형·비정형데이터 활용한 건설 재해 예측 모델 개발(조민건, 이동환, 박주영, 박승희)
  - 주요내용: 건설 현장의 정형데이터(기상, 건설 현장 상황, 장소 등)와 비정형데이터(사고 경위)를 활용하여 재해를 사전에 예측할 수 있는 모델 개발
  - 결론: 정형데이터만 활용했을 때, 비정형데이터만 활용했을 때, 정형, 비정형 둘 다 사용했을 때를 비교했을 때 둘 다 사용한 건설 재해 예측 모델의 정확도가 95.41%로 다른 모델보다 20% 향상됨
  - 한계점: 데이터 불균형으로 인해 예측 모델에 반영되지 못한 요인들이 있을 수 있고 이러한 요인들을 합한 위험도를 예상하고 이를 안전관리 대책과 연계시킬 수 있는 연구가 필요
- 기업 특성이 산업재해 발생에 미치는 영향 : 중소기업과 대기업 비교(김명중, 박선영)
  - 주요내용: 기업의 규모에 따라 산업재해 발생에 영향을 주는 요인이 다르다.
  - 결론: 기업의 규모가 클수록 재정건전성 및 외국인의 비중이 산업재해 발생에 영향을 준다.
  - 한계점: 재무제표를 기반으로 한 분석으로 산업재해 발생 대상자에 대한 개인정보가 빠짐으로 인해 개개인의 경우로는 대입하기 어려움
- 인공지능을 활용한 산업재해 예방 현황과 전망(서동민)
  - 주요내용: 산업재해 예방 분야에서의 AI 활용 인식변화와 기술 동향을 확인하여 국내 산업재해 분야에서의 AI 기술 활용에 대한 고찰을 정리함.
  - 결론: AI 솔루션을 활용하여 작업 환경을 감시하고 위험 예지, 의사결정 등을 수행함으로써 재해 발생 가능성을 최소화하는 긍정적인 전망이 제시됨.
  - 한계점: 기술 개발의 필요성에 대해서만 언급하고 실천을 위한 방법은 제시하지 않음

## 5. 연구 결과의 기대 성과

본 연구를 활용하여 건설 현장에서 발생하는 산업재해 위험도를 수치화하여 누구나 간편하게 산업 안전에 참고할 수 있는 모델을 사용할 수 있을 것으로 기대된다. 특히 산재의 심각성(경상, 중상 등)별로 위험도를 별도 수치화하면 산업 종사자, 기업, 정부 입장별로 우선하여 고려해야 할 분야를 선택적으로 활용할 수 있게 된다. 예를 들어 산업 종사자의 경우 산재 발생 시 산재보험을 적용하기 위한 기업 측의 책임 여부를 본 연구의 모델을 활용하여 뒷받침할 자료로 사용할 수 있다.




## 6. 참고문헌

김학열, 허태영, “산업재해 발생의 상대위험도 분석 및 순환분포 모형 추정,”(서울도시연구 제11권 제1호, 2010. 3), 127~138.

김현수, 이현수, 박문서, “건설공사의 정량적 위험도 산정 방법론,”(한국건설관리학회 학술발표대회 논문집, 2008), 463~466

조민건, 이동환, 박주영, 박승희, “건설현장 정형·비정형데이터를 활용한 기계학습 기반의 건설재해 예측 모델 개발,”(대한토목학회 논문집 제42권 제1호(통권 제220호), 2022), 127~134

김명중, 박선영, “기업 특성이 산업재해 발생에 미치는 영향 : 중소기업과 대기업 비교,”산업연구 Journal of Industrial Studies(J.I.S)(2023) Vol.4 No.2

서동민, “인공지능을 활용한 산업재해 예방 현황과 전망,” 한국콘텐츠학회(2023) Vol.21 No.1

송태호, ”건축시공현장관리를 위한 가설공사 위험도 지수 모델 제안,“ (석사학위, 금오공과대학교 산업대학원 토목,환경 및 건축공학과, 2019)