1 단계 문제 제기

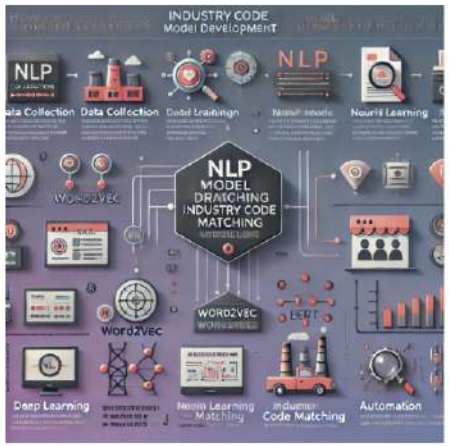
수요기업의 업종 매칭 필요를 분석하다

수요기업은 위험성 평가 모형과 맞춤형 보험료 산정 솔루션 개발을 위해 빅데이터 분석을 도입하고 있다. 현재 다양한 데이터 출처 간 융합 과정에서 "업종코드"를 수작업으로 통일하고 있어 시간과 비용 문제가 발생하고 있다. 이를 해결하기 위해 업종 매칭 알고리즘이 필요하며, 이를 통해 데이터를 자동화하고 분석의 정확성과 효율성을 높일 수 있다. 이와 함께, 데이터 융합과 분석 결과를 기반으로 위험성 평가 모형과 맞춤형 보험료 산정 솔루션을 개발하여 고객 맞춤형 서비스를 제공하고 기업 경쟁력을 강화할 계획이다. 빅데이터 분석은 이러한 과제를 해결하고 지속 가능한 성장을 지원하는 핵심 도구로 자리 잡고 있다.

업종코드 매칭을 위해 다양한 출처에서 업종 유추 데이터를 수집해야 한다. 이를 통해 데이터 융합의 신뢰성과 정확성을 강화할 수 있다. 업종코드 매칭을 위한 주요 데이터는 다음과 같다.

- [빅데이터플랫폼] 경기지역경제포털 사업자데이터: 기업체명, 업종명, 제품명, 주생산품, 지적재산권 명칭 및 내용 등 기업의 업종 특성을 나타내는 정보.
- [공공데이터] 통계청 제10차, 11차 한국표준산업분류 데이터: 세세분류 업종명, 국가 단위의 표준화된 업종 데이터.
- [공공데이터] 국세청 산업 분류 데이터: 세세분류 업종명, 과세 및 신고 기반의 정확한 업종 정보.
- [공공데이터] 국세청 전국 가동사업자현황 데이터: 지역별, 성별, 존속연수별 가동 사업자 현황, 생활밀접 업종 사업자 정보.
- [공공데이터] 전국등록공장현황 데이터: 공장에서 생산하는 주요 생산품과 원자재 정보.

이 데이터들은 업종 매칭 알고리즘 개발에 필요한 핵심 자료로, 업종 분류의 정확성을 높이고 데이터 처리 과정을 자동화하는 데 기여할 것이다. 이를 통해 업종코드 매칭의 효율성을 높이고, 위험성 평가와 맞춤형 솔루션 개발에 있어 데이터 기반 의사결정을 지원할 수 있다.



2 단계 데이터활용

정확한 업종코드 매칭을 위해 NLP(자연어처리) 모델을 도입하다

개별 기업들은 서로 다른 업종코드를 활용하는 문제가 있었다. 따라서 이를 해결하기 위해, 업종명 혹은 이와 관련된 업종 유추 데이터, 즉 텍스트 데이터를 이용한 업종코드 매칭 기술이 필요하였다. 기업의 업종과 관련된 텍스트 데이터를 얻기 위해, KoDATA의 기업 데이터, 한국산업단지공단의 전국등록공장현황 데이터를 이용하였다. 사용된 데이터는 기업의 제품 혹은 생산품, 원자재, 지적재산권, 용도 등과 관련된 데이터였으며, 궁극적으로는 업종명과 관련된 어떤 키워드가 오더라도 정확한 업종코드 매칭이 가능하도록 최대한 많은 학습 데이터를 수집하였다. 이와 더불어 Chat GPT의 추론 특화 모델인 o1-preview를 이용하여, 국세청 업종코드 약 1,200개에 대한 관련 키워드를 생성, 이를 추가적인 학습 데이터로 사용하였다. 이렇게 구축한 학습데이터에서도 최대한 관련이 있는 키워드만 추출하기 위해 Word2Vec 기술을 이용하였고, 최종 학습 데이터를 기반으로 NLP(자연어처리) 모델인, 한국어 기반 KoBERT에 Fine Tuning(특정 도메인에 적용하기 위한 추가 학습) 하였다. 학습 결과, 업종명과 관련된 키워드를 입력하면 적절한 국세청 업종코드를 매칭하는 AI 모델을 개발하게 되었다.

활용 데이터 현황	
수집 채널	데이터명
지역경제 빅데이터 플랫폼	사업자 데이터
외부 데이터	전국 등록 공장현황 데이터 (한국산업단지공단)
	한국표준산업분류(KSIC) 데이터(통계청)
	업종코드 데이터, 100대 생활업종 신규사업자 현황(국세청)
수요기업 데이터	해당 없음

구분	사용 전문 목록	사용 칼럼명	비고
사업자 데이터 (경기지역 경제포털)	2. 기업개요 (DB0538_1)	기업체명	기업을 특징하는 컬럼
		업종명 (표준산업분류 11차)	
		제품명	
		소재지우편번호주소	전국등록공장현황 데이터와 조인 시기 업체명과 함께 사용
	14. 사업장 (DB0538_4)	주생산품	
	24. 지적재산권 (DB0538_8)	명칭	
		내용	
	68. 주요제품매출구성 (DB0538_10)	용도	
공장데이터 (한국산업단지공단 전국등록공장현황) 기준:2023-10-04	전국등록공장현황	회사명	기업데이터 데이터와 조인시 사용
		법인주소	
		생산품	추출된 키워드는 사업자데이터의 키워드와 합쳐져 하나의 칼럼으로 사용
		원자재	
GPT 활용 데이터	o1-preview 기반 생성키워드	업종명 기반 키워드	기존 사업자데이터, 공장데이터와 결합 후 필터링을 통해 최종데이터 구축

업종코드 매칭 AI 모델을 통해, 매칭 자동화 가능성을 열다

01 분석결과

텍스트 기반 업종코드 매칭 AI 모델 개발

업종명과 관련된 키워드로 학습 데이터를 구축, 한국어 기반 KoBERT로 학습한 뒤, '국세청 100대 생활업종'으로 테스트를 실시하였다. 이 결과, 정확도 90%를 상회하는 AI 모델을 개발할 수 있었다. 이는 분류해야 하는 클래스(업종코드)가 1,200개가 넘는 어려운 작업임을 감안하면, 상당히 높은 정확도이다. 특히, 개별 기업들이 서로 다른 업종코드를 활용하더라도, 텍스트 기반으로 업종코드를 매칭하기 때문에, 향후 수기로 매칭할 필요 자체를 줄였다는 점에서 의미 있다. 또한, 이번 분석을 통해 분류 클래스가 매우 많은 경우에도, 고품질 학습 데이터를 확보할 수 있다면, 충분히 좋은 분류 결과를 낼 수 있다는 점에서 의의를 갖는다.

밀접 업종명 (X)	KSIC11_BZC_NM (Y, 실제업종명)	예측 업종명	일치 여부	정확도
휴대폰 가게	이동전화기 제조업	이동전화기 제조업	일치	0.92857
화장품 가게	화장품, 비누 및 방향제 소매업	화장품, 비누 및 방향제 소매업	일치	92.8571
호프 주점	일반 유흥주점업	일반 유흥주점업	일치	
헬스 클럽	체력단련시설 운영업	체력단련시설 운영업	일치	
한식 음식점	한식 음식점업	한식 일반 음식점업	불일치	
한방병원, 한의원	한의원	한의원	일치	
피부, 비뇨기과의원	일반 의원	일반 의원	일치	
피부 관리업	피부 미용업	피부 미용업	일치	
편의점	체인화 편의점	체인화 편의점	일치	

02 활용방안

업종 매칭 작업 자동화와 보험 가입 솔루션 기능 고도화

개발된 코드 매칭 AI 모델을 이용하면, 수기로 진행되었던 업종 매칭 작업이 자동화되어 업무 효율성이 증대된다. 이는 기업 업종에 따른 맞춤형 보험료 산정에 활용될 것이다. 또한 이러한 자동화 기능을 현재 수요기업의 보험가입 솔루션에 웹 혹은 어플에 적용시키면, 잠재 보험 가입자의 편의성이 증대되고 수요기업 역시 업무를 빠르게 처리할 수 있게 된다. 특히, 이러한 효과는 본인 사업장의 업종을 잘 모르는 소상공인 및 중소기업 사업주에게 유효하게 작용될 것으로 보인다.

'23년도 사업종류별 산재보험료율

단위:천분율(%)

사업종류	요율	사업종류	요율
1. 광업		**4. 건설업**	36
석탄광업 및 채석업	185	**5. 운수.창고.통신업**	
석회석·금속·비금속·기타광업	57	철도·항공·창고·운수 관련 서비스업	8