**󰊱 명칭**

|  |
| --- |
| **명칭: K-FACT (Korean Factory Analysis and Choice Tool)** |

**󰊲 제안배경**

|  |
| --- |
| 산업단지와 같은 공장 집적지는 비슷한 업종의 공장들이 밀집해 있어 다양한 시너지 효과를 창출할 수 있습니다. 이러한 효과는 기업들이 상호 협력하고 경쟁하며, 기술을 빠르게 공유하고, 인재를 유치하는 데 큰 도움을 줍니다. 아래는 한국에서 성공적으로 발생한 산업 집적지의 세 가지 사례입니다.   **홍릉 바이오 클러스터:** 서울 홍릉에 위치한 바이오 클러스터는 고려대학교 병원, 경희대학교 병원 등과 연계되어 있습니다. 이 지역의 바이오 메디컬 산업은 의약품 제조, 의료 기기, 보건업 등으로 구성되어 있으며, 관련 기업들이 집중되어 있습니다. 조사에 따르면, **이 지역에 위치한 기업들은 70.2%가 시너지 효과를 누리고 있으며, 이는 비집적지의 19.2%보다 훨씬 높은 수치입니다​​.**    (https://www.metroseoul.co.kr/article/2019070300105)   **창원국가산업단지:** 창원은 한국 최대의 기계 산업 집적지로, 다양한 기계 제조 기업들이 밀집해 있습니다. 이 지역의 기업들은 부품 조달, 기술 개발, 인력 교육 등의 측면에서 큰 이점을 누리고 있으며, 서로 협력하여 생산성을 높이고 있습니다. 창원국가산업단지의 기계 산업은 국가 경제의 중요한 부분을 차지하며, 지속적인 성장을 이루고 있습니다​​.    (<https://www.cwnic50th.or.kr/main/main.php>)   **여의도 증권가:** 여의도는 한국의 금융 중심지로, 주요 증권사와 금융 기관들이 밀집해 있습니다. 이 지역의 금융 기관들은 서로의 정보를 빠르게 교환하고, 다양한 금융 상품을 공동 개발하며, 금융 시장의 변동에 신속하게 대응할 수 있는 능력을 갖추게 됩니다. 이러한 집적은 금융 서비스의 품질을 높이고, 금융 산업 전체의 성장을 촉진합니다. 여의도 증권가의 성공 사례는 금융 업종에서의 집적지가 얼마나 중요한 역할을 하는지 보여줍니다​​. 관련 현황 및 문제점 산업 집적지의 성공적인 사례에도 불구하고, 공장 부지 선택 과정에서 여전히 많은 기업들이 어려움을 겪고 있습니다. 다음은 대표적인 문제점과 그에 따른 빅데이터 분석의 필요성입니다   1. **부지 선택의 복잡성** 공장 부지 선택은 부지의 물리적 조건, 인프라, 노동력, 비용 등 다양한 요소를 고려해야 하는 복잡한 과정입니다. 이 과정에서 많은 기업들이 데이터 부족이나 분석 능력의 한계로 인해 최적의 부지를 선택하지 못하고 있습니다. 2. **시너지 효과의 극대화 부족** 비슷한 업종의 기업들이 모여 있는 집적지에서는 시너지 효과를 극대화할 수 있지만, 이를 위한 정확한 데이터와 분석이 부족한 경우 효과를 충분히 누리지 못할 수 있습니다. 기업들은 자신들에게 최적화된 집적지를 찾기 위해 신뢰할 수 있는 데이터와 분석이 필요합니다. 3. **빅데이터의 중요성 및 활용 빈도** 빅데이터 분석은 공장 부지 선택 과정에서 중요한 역할을 할 수 있습니다. 데이터 기반의 접근은 부지 선택의 복잡성을 줄이고, 시너지 효과를 극대화할 수 있는 최적의 지역을 추천하는 데 도움을 줍니다. 빅데이터 분석을 통해 기업들은 보다 정확한 의사 결정을 내릴 수 있으며, 이는 기업의 성공과 직결됩니다.  제안 모델 공장 부지 추천 모델은 이러한 문제점을 해결하기 위해 개발되었습니다. 이 모델은 다음과 같은 데이터를 활용하여 공장 부지 선택을 지원합니다:   1. **업종별 특성 분석**    * 공장구분, 설립구분, 입주형태, 보유구분, 등록구분 등의 데이터를 통해 각 업종의 특성을 철저히 분석하여, 업종별로 최적의 입지 조건을 정의합니다. 2. **종업원 관련 데이터**    * 남자종업원, 여자종업원, 외국인남자종업원, 외국인여자종업원, 종업원합계 등의 데이터를 분석하여 노동력의 구성과 가용성을 평가합니다. 3. **공장 규모 및 시설 데이터**    * 공장규모, 용도지역, 지목, 용지면적, 제조시설면적, 부대시설면적, 건축면적 등의 데이터를 통해 부지의 물리적 조건과 적합성을 평가합니다. 4. **지식산업센터 및 대표업종 데이터**    * 지식산업센터명, 대표업종 등의 데이터를 통해 특정 업종이 밀집된 지역을 분석하고, 해당 업종에 최적화된 지역을 추천합니다.  결론 공장 부지 선택은 기업의 성공과 실패를 좌우할 수 있는 중요한 요소입니다. 한국의 창원 국가 산업단지, 홍릉 바이오 클러스터, 여의도 증권가와 같은 성공적인 사례들은 집적지의 중요성을 잘 보여줍니다. 이러한 집적지는 기업들이 상호 협력과 경쟁을 통해 기술 혁신과 생산성 향상을 이루어낼 수 있는 최적의 환경을 제공합니다.  공장 부지 추천 모델은 이러한 성공 사례를 바탕으로 개발되었습니다. 이 모델은 다양한 데이터를 분석하여 업종별 특성과 필요 요건을 평가하고, 최적의 부지를 추천함으로써 기업들이 최대의 시너지 효과를 누릴 수 있도록 돕습니다. 이를 통해 기업은 경쟁력을 강화하고, 생산성을 높이며, 장기적인 성공을 이룰 수 있을 것입니다. |

**󰊳 분석 내용 및 분석 결과**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 모델 데이터셋 개요 사용 데이터: 한국산업단지공단\_공장등록 현황 통계정보(24-06-03)  (https://www.data.go.kr/data/3041646/fileData.do)  이 데이터셋은 공공데이터포탈의 한국산업단지공단\_공장등록 현황 통계정보의 데이터를 사용하였으며, 공장의 위치, 대표업종, 직원 수, 생산품, 시설 면적 등의 데이터를 포함합니다. 본 프로젝트에서는 공장 부지 추천 모델을 개발하기 위해 데이터셋의 일부 변수를 사용하였습니다. **사용된 변수 설명** 모델 학습에 사용된 주요 변수들은 다음과 같습니다:   1. 공장구분: 공장의 구분 (개별, 계획) 2. 설립구분: 공장의 설립 구분 (일반, 국가산업단지 등) 3. 입주형태: 공장의 입주 형태 (일반, 경매 등) 4. 보유구분: 공장 부지의 보유 구분 (임대, 자가) 5. 등록구분: 공장의 등록 구분 (신규, 등록, 부분, 완료) 6. 남자종업원 7. 여자종업원 8. 외국인 남자종업원 9. 외국인 여자종업원 10. 종업원 합계 11. 공장규모: 소기업, 중기업, 대기업 12. 용도지역: 공장이 위치한 용도 지역 13. 지목 14. 용지면적: 공장 부지의 면적 15. 제조시설면적: 제조 시설의 면적 16. 부대시설면적: 부대 시설의 면적 17. 건축면적 18. 지식산업센터명: 공장이 위치한 지식산업센터의 이름 19. 대표업종: 공장의 대표 업종 20. 공장주소(구/동/읍/면) (Target 데이터)  **데이터 분석 절차**  1. **데이터 전처리**    * **전체 데이터 소개**: 시도명, 시군구명, 관리기관, 회사명, 공장구분, 단지명, 설립구분, 입주형태, 보유구분, 등록구분, 전화번호, 남자종업원, 여자종업원, 외국인남자종업원, 외국인여자종업원, 종업원합계, 생산품, 원자재, 공장규모, 용도지역, 지목, 용지면적, 제조시설면적, 부대시설면적, 건축면적, 지식산업센터명, 대표업종, 업종명, 업종코드, 차수, 법인주소, 필지수, 공장주소, 공장주소\_지번, 공장관리번호    * **데이터 삭제**: 분석에 필요하지 않은 다음 데이터를 삭제하였습니다: 시도명, 시군구명, 관리기관, 회사명, 단지명, 전화번호, 법인주소, 필지수, 공장주소, 공장관리번호, 생산품, 원자재, 업종명, 업종코드, 차수. 2. **기본 모델 선정:** 위에서 설정한 삭제할 칼럼을 제외한 후 최적의 모델을 찾기 위해 XGBoost, LightGBM, CatBoost, KNN 모델들을 활용하여 정확도가 가장 높은 모델 2개를 선정합니다.   **모델 설명 및 장단점**   1. XGBoost (Extreme Gradient Boosting): XGBoost는 Gradient Boosting 알고리즘을 기반으로 하는 강력한 머신 러닝 알고리즘 Gradient Boosting은 여러 개의 약한 학습자를 결합하여 강력한 예측 모델을 만드는 앙상블 기법  장점  * **높은 예측 성능**: 대부분의 데이터셋에서 뛰어난 성능을 보임. * **효율성**: 병렬 처리 및 분산 컴퓨팅을 지원하여 학습 속도가 빠름. * **유연성**: 다양한 커스터마이징 옵션과 매개변수 조정 가능.  단점  * **복잡성**: 많은 하이퍼파라미터가 있어 최적의 모델을 찾기 어려울 수 있음. * **메모리 사용량**: 대규모 데이터셋에서는 메모리 사용량이 많아질 수 있음.  1. LightGBM (Light Gradient Boosting Machine): LightGBM은 Microsoft에서 개발한 Gradient Boosting 알고리즘. LightGBM은 대용량 데이터셋을 효율적으로 처리할 수 있도록 설계되었으며, 학습 속도와 메모리 사용 효율성이 뛰어남.  장점  * **빠른 학습 속도**: Histogram-based 알고리즘을 사용하여 학습 속도가 매우 빠름. * **메모리 효율성**: 메모리 사용량이 적음. * **고성능**: 높은 예측 성능을 보임.  단점  * **복잡성**: 많은 하이퍼파라미터가 있어 최적화가 어려울 수 있음. * **균형 잡힌 데이터셋 필요**: 데이터가 균형 잡혀 있지 않으면 성능이 저하될 수 있음.  1. CatBoost: CatBoost는 Yandex에서 개발한 Gradient Boosting 알고리즘. 특히 범주형 데이터 처리를 최적화하여 높은 성능을 보여주며, 자동으로 범주형 데이터를 인코딩하여 처리할 수 있는 특징.  장점  * **범주형 데이터 처리**: 범주형 데이터를 자동으로 처리하여 성능을 극대화. * **고성능**: 다양한 데이터셋에서 높은 예측 성능을 보임. * **사용 용이성**: 별도의 데이터 전처리가 거의 필요 없음.  단점  * **학습 속도**: 다른 Gradient Boosting 알고리즘에 비해 다소 느릴 수 있음. * **복잡성**: 많은 하이퍼파라미터가 있어 최적화가 어려울 수 있음.  1. K-Nearest Neighbors (KNN): KNN은 새로운 데이터 포인트를 기존 데이터와 비교하여 가장 가까운 K개의 이웃을 기준으로 분류하거나 회귀하는 비모수적 방법. 거리 측정 방식을 통해 데이터의 유사성을 기반으로 예측.  장점  * **단순함**: 이해하고 구현하기 쉬움. * **유연성**: 분류와 회귀 모두 사용 가능.  단점  * **느린 예측 속도**: 새로운 데이터 포인트를 예측할 때 모든 데이터를 비교해야 하므로 느릴 수 있음. * **메모리 사용량**: 모든 데이터를 저장해야 하므로 메모리 사용량이 많음. * **불균형 데이터**: 데이터가 불균형할 경우 성능이 저하될 수 있음.  1. **예측 데이터 설정**   최적의 공장부지 '구'(특정 지역의 경우 ‘동’, ‘읍’. ‘면’)를 추천하려고 합니다. 공장주소\_지번의 데이터를 시, 구, 동으로 나눠서 '구' 예측을 진행합니다.  **예시) ‘서울특별시 종로구 창선동 87-1 번지’를 서울특별시 / 종로구 / 창선동 으로 분리한 후 ‘창성동’을 예측 값으로 선정** **모델 성능 결과 (전처리 없이 진행)** **데이터 분할**  데이터셋을 학습, 검증, 테스트 세트로 분할하였습니다. 이를 통해 모델의 성능을 보다 정확하게 평가할 수 있습니다.   * 전체 데이터의 15%는 테스트 데이터로 사용하였습니다. * 나머지 85%의 데이터 중 20%는 검증 데이터로 사용하였고, 나머지 80%는 학습 데이터로 사용하였습니다. * 결과적으로, 전체 데이터의 68%는 학습 데이터, 17%는 검증 데이터, 15%는 테스트 데이터로 구성되었습니다.  |  |  |  | | --- | --- | --- | | 모델 | 정확도 (Accuracy) | F1-Score | | **XGBoost** | **0.6759** | **0.3561** | | **LightGBM** | **0.6795** | **0.3488** | | CatBoost (Original) | 0.6100 | 0.3300 | | CatBoost (Label Encoding) | 0.6500 | 0.3300 | | KNN | 0.4700 | 0.1500 |   \*Accuracy(정확도): Accuracy는 전체 데이터 중에서 모델이 올바르게 예측한 데이터의 비율을 의미합니다. 이는 가장 직관적인 성능 지표로, 특히 데이터가 균형 잡혀 있을 때 유용합니다.  \*F1-Score: F1-Score는 Precision과 Recall의 조화 평균을 사용하여 계산됩니다. 이는 데이터가 불균형 할 때 모델의 성능을 더 잘 평가할 수 있는 지표입니다. 이 평가 결과는 데이터 전처리를 적용하기 전의 모델 성능을 보여주며, 이후 단계에서 적용될 다양한 전처리 기법이 모델 성능에 미치는 영향을 평가하기 위한 기준점으로 활용됩니다.분석 알고리즘 세부 내용  1. **Scaler 적용:** 스케일링은 데이터의 크기를 일정한 범위로 조정하여 모델의 학습 성능을 높이고 학습 시간을 단축시키는 데 도움을 주는 방식입니다. 다양한 스케일링 방법을 적용하여 모델의 성능을 평가하였습니다.  Min-Max Scaler: Min-Max Scaler는 데이터의 최소값을 0으로, 최대값을 1로 변환합니다. 이 스케일링 방법은 데이터의 범위를 [0, 1] 또는 사용자가 지정한 다른 범위로 조정합니다. **공식:**  python - Can someone explain to me how MinMaxScaler() works? - Stack  Overflow  **장점**   * 직관적이고 이해하기 쉬움. * 데이터가 고르게 분포된 경우 효과적.   **단점**   * 이상치(outlier)에 민감함. 이상치가 있는 경우 스케일링 결과가 왜곡될 수 있음.  Standard Scaler: Standard Scaler는 데이터의 평균을 0, 표준 편차를 1로 변환합니다. 이는 데이터가 정규 분포를 따를 때 효과적입니다. **공식:**  Which data scaling technique should I use ? | by Hasan Hüseyin Coşgun |  Medium  **장점**   * 이상치에 덜 민감함. * 대부분의 머신 러닝 알고리즘에서 기본적으로 선호됨.   **단점**   * 데이터가 정규 분포를 따르지 않는 경우 효과가 떨어질 수 있음.  Robust Scaler: Robust Scaler는 중앙값(median)과 IQR(interquartile range, 1사분위수와 3사분위수의 차이)를 사용하여 스케일링합니다. 이는 이상치에 덜 민감하도록 설계되었습니다. **공식:**  Feature Scaling - Part 3 - GeeksforGeeks  **장점**   * 이상치에 매우 강함. * 데이터가 이상치를 포함하는 경우 효과적.   **단점**   * 데이터가 고르게 분포되지 않은 경우 효과가 떨어질 수 있음.  Scaler 적용 결과 3개의 Scaler를 적용했을 때의 정확도와 F1-Score 값은 다음과 같습니다:   |  |  |  | | --- | --- | --- | | Scaler | XGBoost Accuracy | LGBM Accuracy | | Min-Max Scaler | 0.51 | 0.58 | | Standard Scaler | 0.28 | 0.44 | | Robust Scaler | 0.53 | 0.30 |   3가지의 다양한 Scaler 방식을 진행하였으나, 오히려 정확도가 내려가는 결과를 보였습니다. 따라서, Scaler는 적용하지 않기로 결정하였습니다.   1. **Encoding 방법:** 범주형 데이터를 처리하기 위해 두 가지 주요 인코딩 기법을 적용하였습니다: Label Encoding과 Target Encoding.   각 방법의 장단점과 적용 결과는 다음과 같습니다. Label Encoding: Label Encoding은 범주형 변수의 각 범주를 고유한 정수로 변환하는 기법입니다. 각 범주는 0부터 시작하는 정수로 매핑됩니다. **장점**   * **간단하고 빠름**: 구현이 매우 간단하고 빠릅니다. * **메모리 효율성**: 비교적 적은 메모리를 사용합니다.   **단점**   * **순서 정보 부여**: 각 범주에 숫자를 할당하기 때문에, 모델이 이러한 순서 정보를 의미 있는 것으로 잘못 해석할 수 있습니다. * **고차원 데이터 처리 한계**: 많은 범주를 가진 변수에 비효율적입니다.  1-5-2. Target Encoding: Target Encoding은 범주형 변수의 각 범주를 해당 범주의 타겟 값의 평균으로 변환하는 기법입니다. 이는 특정 범주가 타겟 변수와 어떤 관계를 가지는지 파악할 수 있도록 도와줍니다. **장점**   * **정보 보존**: 범주와 타겟 변수 간의 관계를 잘 반영할 수 있습니다. * **고차원 데이터 처리**: 많은 범주를 가진 변수에 효과적입니다.   **단점**   * **데이터 누출**: 잘못 사용할 경우, 타겟 변수의 정보를 학습 데이터에 누출시킬 수 있습니다. * **오버피팅**: 특히 작은 데이터셋에서는 오버피팅의 위험이 있습니다.  Encoding 방법 적용 결과 범주형 데이터에 대해 Label Encoding과 Target Encoding을 각각 적용하여 모델의 성능을 평가한 결과는 다음과 같습니다:   |  |  |  |  | | --- | --- | --- | --- | | 인코딩 방법 | 모델 | 정확도 (Accuracy) | F1-Score | | Label Encoding | XGBoost | 0.6759 | 0.3561 | | LGBM | 0.6795 | 0.3488 | | Target Encoding | XGBoost | 0.6759 | 0.3561 | | LGBM | 0.6759 | 0.3488 |   두 가지 인코딩 방법을 적용한 결과, 성능 차이가 발생하지 않았습니다. 최종적으로 메모리 효율성과 데이터 누수를 방지하기 위해 Label Encoding을 적용하였습니다. Label Encoding을 적용함으로써 범주형 데이터의 처리가 간단해지고, 데이터 누출의 위험이 줄어들었습니다.   1. **Feature Importance: Feature importance는 머신 러닝 모델에서 각 피처(특성, 변수)가 모델의 예측 성능에 얼마나 기여하는지를 나타내는 지표입니다. 이를 통해 모델 해석 가능성을 높이고, 중요한 피처를 파악하여 모델 성능을 개선하거나 차원 축소를 수행하는 데 유용합니다.**   **다음 표는 LGBM 모델의 피처 중요도를 나타내고 있습니다.**    가장 중요한 피처는 대표업종으로 나타났으며, 그 다음으로 중요한 피처들은 제조시설면적, 용지면적, 건축면적 등 공장 면적과 관련된 변수들이 차지하였습니다. 직원 수와 관련된 피처들인 종업원합계, 남자종업원, 여자종업원도 중요한 변수로 나타났습니다. 그 외에도 지식산업센터명이 중요한 피처로 분석되었으며, 이는 산업집적지가 공장 부지 선정에 중요한 영향을 미친다는 점을 보여줍니다. 반면, 등록구분, 보유구분 및 외국인 종업원 수와 관련된 피처들은 모델 성능에 큰 영향을 미치지 않는 것으로 나타났으며, 공장 규모(소기업, 중기업, 대기업)도 중요하지 않은 변수로 분석되었습니다.  모델이 중요하게 고려하는 피처들을 분석한 결과, 대표업종, 제조시설면적, 용지면적, 건축면적, 부대시설면적, 종업원합계 순으로 '구'를 추천해주는 것으로 나타났습니다. 이는 공장 데이터에 있어서 업종 선정이 가장 중요하였으며, 면적과 종업원 수가 그 다음으로 중요한 요소임을 보여줍니다. 따라서, 공장 부지 추천 모델은 주로 대표업종, 공장 면적, 직원 수와 같은 변수들에 의해 결정된다고 할 수 있습니다. **불필요한 피처 제거** Feature Importance 분석 결과를 참고하여 모델에 크게 영향을 미치지 않는 피처들을 제거하였습니다. 제거한 피처들은 다음과 같습니다:   * **용도지역** * **등록구분** * **지목** * **보유구분**  |  |  |  | | --- | --- | --- | |  | XGBOOST | LGBM | | Accuracy | 0.6062 | 0.2708 | | F1-Score | 0.6127 | 0.2803 |   제거한 피처들이 모델 성능에 미치는 영향이 크지 않음을 확인할 수 있었습니다. Tree 모델의 특성상, 가지치기 과정에서 많은 피처들이 영향을 미칠 수 있지만, 중요하지 않은 피처들은 성능에 큰 영향을 주지 않는 것으로 나타났습니다. **추가 피처 제거 및 모델 성능 분석** Feature Importance를 통해 추가로 다음과 같은 피처들을 제거하고 모델 성능을 다시 평가하였습니다:   * **공장구분** * **설립구분** * **외국인남자종업원** * **외국인여자종업원** * **공장규모**  |  |  |  | | --- | --- | --- | |  | XGBOOST | LGBM | | Accuracy | 0.5880 | 0.2587 | | F1-Score | 0.5966 | 0.2678 |   칼럼을 제거할수록 모델의 성능이 낮아지는 것을 확인할 수 있었습니다. 이를 통해 Feature Importance가 모델에 큰 영향을 미치지 않더라도, 세부적으로 나눠지는 부분에서 중요한 역할을 한다고 판단하였습니다. 또한, 칼럼 삭제를 통해 정확도를 향상시키는 방법에는 한계가 있다는 결론에 도달하였습니다.   1. **이상치 처리**  IQR: IQR(Interquartile Range)은 사분위수 간 범위를 이용하여 이상치를 탐지하는 방법입니다. IQR은 데이터의 중앙 50%의 범위를 나타내며, 이를 통해 이상치를 효과적으로 탐지할 수 있습니다.  * **사분위수 계산**:   + Q1 (1사분위수): 데이터의 25번째 백분위수   + Q3 (3사분위수): 데이터의 75번째 백분위수 * **IQR 계산**:   + IQR = Q3 - Q1 * **이상치 경계 설정**:   + 하한 경계: Q1 - 1.5 × IQR   + 상한 경계: Q3 + 1.5 × IQR * **이상치 탐지**:   + 데이터가 하한 경계보다 작거나 상한 경계보다 큰 값은 이상치로 간주합니다.  UNIFY: UNIFY 방법은 이상치를 탐지한 후, 특정 값으로 치환하여 이상치를 처리하는 방법입니다. 이는 데이터의 일관성을 유지하고, 모델의 성능을 안정화하는 데 도움이 됩니다. 일반적으로 이상치는 평균, 중앙값 또는 특정 값으로 대체합니다. 위 방법들을 통해 모든 경우의 수에 대해 이상치를 제거 및 변경하면서 정확도와 F1 스코어를 측정하였습니다. 적용 순서는 Feature Importance의 순서로 진행하였습니다. **이상치 처리 적용 결과** 다양한 경우의 수로 IQR과 UNIFY 방법을 적용하여 정확도와 F1 스코어를 측정하였습니다. 적용 순서는 Feature Importance의 순서로 진행되었습니다.   1. 제조시설면적   라인, 도표, 스크린샷, 직사각형이(가) 표시된 사진  자동 생성된 설명  제조시설면적 데이터는 대부분이 작은 값에 몰려 있으며, 극단적인 이상치들이 존재합니다.  [IQR]   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Accuracy | | F1-Score | | | IQR 범위 | XGBOOST | LGBM | XGBOOST | LGBM | | 0.1 ~ 0.90 | 0.6628 | 0.6670 | 0.3328 | 0.3328 | | 0.1 ~ 0.95 | 0.6520 | 0.6679 | 0.3255 | 0.3269 | | 0.05 ~ 0.90 | 0.6615 | 0.6719 | 0.3317 | 0.3497 | | **0.05 ~ 0.85** | **0.6757** | **0.6767** | **0.3247** | **0.3133** | | 0.05 ~ 0.80 | 0.6397 | 0.6419 | 0.3239 | 0.3215 |   [UNIFY]   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Accuracy | | F1-Score | | | UNIFY 범위 | XGBOOST | LGBM | XGBOOST | LGBM | | 50000이상 | 0.6620 | 0.6665 | 0.3202 | 0.3234 | | 40000 이상 | 0.6620 | 0.6665 | 0.3202 | 0.3234 | | 28000이상 | 0.6620 | 0.6665 | 0.3202 | 0.3234 | | 16000이상 | 0.6604 | 0.6665 | 0.3219 | 0.3234 | | 10000이상 | 0.6594 | 0.6665 | 0.3152 | 0.3234 | | 6300이상 | 0.6660 | 0.6675 | 0.3273 | 0.3211 | | 5300이상 | 0.6620 | 0.6675 | 0.3210 | 0.3211 | | 3000이상 | 0.6630 | 0.6690 | 0.3200 | 0.3278 | | 2000이상 | 0.6645 | 0.6670 | 0.3235 | 0.3302 |  1. 용지면적   도표, 라인, 텍스트, 스크린샷이(가) 표시된 사진  자동 생성된 설명  용지면적 데이터는 대부분이 작은 값에 몰려 있으며, 극단적인 이상치들이 많이 존재하지는 않으며, 대부분 40000 이하의 데이터로 존재하는 것을 볼 수 있습니다.  [IQR]   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Accuracy | | F1-Score | | | IQR 범위 | XGBOOST | LGBM | XGBOOST | LGBM | | 0.1 ~ 0.90 | 0.6647 | 0.6641 | 0.3152 | 0.3176 | | 0.1 ~ 0.95 | 0.6460 | 0.6460 | 0.3186 | 0.3103 | | 0.1 ~ 0.85 | 0.6685 | 0.6608 | 0.3244 | 0.3149 | | 0.05 ~ 0.85 | 0.6685 | 0.6608 | 0.3244 | 0.3149 | | 0.1 ~ 0.80 | 0.6176 | 0.6244 | 0.2886 | 0.3094 |   [UNIFY]   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Accuracy | | F1-Score | | | UNIFY 범위 | XGBOOST | LGBM | XGBOOST | LGBM | | 150000이상 | 0.6757 | 0.6767 | 0.3247 | 0.3133 | | 40000 이상 | 0.6757 | 0.6767 | 0.3247 | 0.3133 | | 30000 이상 | 0.6757 | 0.6767 | 0.3247 | 0.3133 | | 16000이상 | 0.6704 | 0.6767 | 0.3132 | 0.3133 | | 12000이상 | 0.6704 | 0.6767 | 0.3132 | 0.3133 | | 8000이상 | 0.6704 | 0.6767 | 0.3132 | 0.3133 | | 6000이상 | 0.6725 | 0.6767 | 0.3139 | 0.3133 | | 5000이상 | 0.6741 | 0.6720 | 0.3212 | 0.3076 | | 4000이상 | 0.6656 | 0.6720 | 0.3165 | 0.3097 |  1. 건축면적   도표, 스크린샷, 라인, 그래프이(가) 표시된 사진  자동 생성된 설명  건축면적 데이터의 경우 극단적인 이상치가 많이 분포하지는 않으며, 20000 이하의 데이터가 많이 분포하고 있습니다.  [IQR]   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Accuracy | | F1-Score | | | IQR 범위 | XGBOOST | LGBM | XGBOOST | LGBM | | 0.05 ~ 0.85 | 0.6419 | 0.6468 | 0.3093 | 0.3180 | | 0.05 ~ 0.90 | 0.6425 | 0.6425 | 0.3199 | 0.3130 | | **0.05 ~ 0.95** | **0.6638** | **0.6776** | **0.3304** | **0.3420** | | 0.05 ~ 0.98 | 0.6600 | 0.6605 | 0.3406 | 0.3410 | | 0.10 ~ 0.95 | 0.6622 | 0.6591 | 0.3434 | 0.3213 |   [UNIFY]   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Accuracy | | F1-Score | | | UNIFY 범위 | XGBOOST | LGBM | XGBOOST | LGBM | | 70000이상 | 0.6757 | 0.6767 | 0.3247 | 0.3133 | | 3000 이상 | 0.6757 | 0.6767 | 0.3247 | 0.3133 | | 2000이상 | 0.6720 | 0.6767 | 0.3181 | 0.3133 | | 1600이상 | 0.6714 | 0.6751 | 0.3195 | 0.3098 | | 1800이상 | 0.6698 | 0.6767 | 0.3116 | 0.3133 |  1. 부대시설면적   도표, 라인, 스크린샷, 직사각형이(가) 표시된 사진  자동 생성된 설명  부대시설면적 데이터의 경우 극단적인 이상값은 많이 존재하지 않으며 13000 이하의 값이 분포하고 있습니다.  [IQR]   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Accuracy | | F1-Score | | | IQR 범위 | XGBOOST | LGBM | XGBOOST | LGBM | | 0.05 ~ 0.85 | 0.6412 | 0.6428 | 0.3324 | 0.3185 | | 0.05 ~ 0.90 | 0.6490 | 0.6527 | 0.3355 | 0.3290 | | 0.05 ~ 0.95 | 0.6571 | 0.6608 | 0.3021 | 0.2964 | | 0.05 ~ 0.98 | 0.6571 | 0.6608 | 0.3021 | 0.2964 | | 0.10 ~ 0.95 | 0.6571 | 0.6608 | 0.3021 | 0.2964 |   [UNIFY]   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Accuracy | | F1-Score | | | UNIFY 범위 | XGBOOST | LGBM | XGBOOST | LGBM | | 8000이상 | 0.6638 | 0.6776 | 0.3304 | 0.3420 | | 3000 이상 | 0.6638 | 0.6776 | 0.3304 | 0.3420 | | 2000이상 | 0.6638 | 0.6776 | 0.3304 | 0.3420 | | 1600이상 | 0.6638 | 0.6776 | 0.3304 | 0.3420 | | **1000이상** | **0.6750** | **0.6755** | **0.3472** | **0.3408** | | 800 이상 | 0.6676 | 0.6760 | 0.3245 | 0.3369 | | **900 이상** | **0.6697** | **0.6792** | **0.3311** | **0.3434** | | 850 이상 | 0.6676 | 0.6766 | 0.3386 | 0.3433 |  1. 종업원합계   도표, 라인, 스크린샷, 그래프이(가) 표시된 사진  자동 생성된 설명  종업원합계의 경우 대부분이 100명 이하로 분포하고 있으며 극단적인 이상치인  명을 넘어가는 값이 보이는데, 이 값은 대기업으로 판단하였다.  [IQR]   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Accuracy | | F1-Score | | | IQR 범위 | XGBOOST | LGBM | XGBOOST | LGBM | | 0.05 ~ 0.85 | 0.6625 | 0.6625 | 0.3212 | 0.3177 | | 0.05 ~ 0.90 | 0.6599 | 0.6663 | 0.3363 | 0.3419 | | 0.05 ~ 0.95 | 0.6528 | 0.6591 | 0.3428 | 0.3541 | | 0.05 ~ 0.98 | 0.6528 | 0.6591 | 0.3428 | 0.3541 | | 0.10 ~ 0.95 | 0.6421 | 0.6469 | 0.3195 | 0.3206 |   [UNIFY]   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Accuracy | | F1-Score | | | UNIFY 범위 | XGBOOST | LGBM | XGBOOST | LGBM | | 1000이상 | 0.6697 | 0.6792 | 0.3311 | 0.3434 | | 600이상 | 0.6697 | 0.6792 | 0.3311 | 0.3434 | | 300이상 | 0.6697 | 0.6792 | 0.3311 | 0.3434 | | 200 이상 | 0.6697 | 0.6771 | 0.3311 | 0.3391 | | 250 이상 | 0.6697 | 0.6771 | 0.3311 | 0.3391 |  1. 남자종업원   라인, 도표, 그래프, 직사각형이(가) 표시된 사진  자동 생성된 설명  남자 종업원의 분포 또한 종업원 합계와 유사한 분포를 나타났으며, 대부분의 기업에서 500명 이하의 남자 종업원을 보유하고있다.  [IQR]   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Accuracy | | F1-Score | | | IQR 범위 | XGBOOST | LGBM | XGBOOST | LGBM | | 0.05 ~ 0.85 | 0.6451 | 0.6527 | 0.3123 | 0.3061 | | 0.05 ~ 0.90 | 0.6534 | 0.6550 | 0.3147 | 0.3247 | | 0.05 ~ 0.95 | 0.6405 | 0.6427 | 0.3234 | 0.3184 | | 0.05 ~ 0.98 | 0.6405 | 0.6427 | 0.3234 | 0.3184 | | 0.10 ~ 0.95 | 0.6405 | 0.6427 | 0.3234 | 0.3184 |   [UNIFY]   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Accuracy | | F1-Score | | | UNIFY 범위 | XGBOOST | LGBM | XGBOOST | LGBM | | 1000이상 | 0.6697 | 0.6792 | 0.3311 | 0.3434 | | 500이상 | 0.6697 | 0.6792 | 0.3311 | 0.3434 | | 300이상 | 0.6697 | 0.6792 | 0.3311 | 0.3434 | | 200 이상 | 0.6697 | 0.6792 | 0.3311 | 0.3434 | | 125 이상 | 0.6697 | 0.6792 | 0.3311 | 0.3434 | | 100 이상 | 0.6676 | 0.6840 | 0.3280 | 0.3387 | | 80 이상 | 0.6702 | 0.6776 | 0.3518 | 0.3389 | | **110 이상** | **0.6697** | **0.6840** | **0.3311** | **0.3387** | | 90 이상 | 0.6676 | 0.6813 | 0.3280 | 0.3458 | | 115이상 | 0.6697 | 0.6840 | 0.3311 | 0.3387 |  1. 여자종업원   도표, 라인, 텍스트, 스크린샷이(가) 표시된 사진  자동 생성된 설명  여자종업원 데이터의 경우 종업원 합계 데이터와 유사하며 대부분 700명 이하의 여자종업원을 보유하고 있다.  [IQR]   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Accuracy | | F1-Score | | | IQR 범위 | XGBOOST | LGBM | XGBOOST | LGBM | | 0.05 ~ 0.85 | 0.6346 | 0.3086 | 0.3280 | 0.2930 | | 0.05 ~ 0.90 | 0.6502 | 0.6573 | 0.2920 | 0.3049 | | 0.05 ~ 0.95 | 0.6421 | 0.6394 | 0.3148 | 0.3131 | | 0.05 ~ 0.98 | 0.6421 | 0.6394 | 0.3148 | 0.3131 | | 0.10 ~ 0.95 | 0.6421 | 0.6394 | 0.3148 | 0.3131 |   [UNIFY]   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Accuracy | | F1-Score | | | UNIFY 범위 | XGBOOST | LGBM | XGBOOST | LGBM | | 800 이상 | 0.6697 | 0.6840 | 0.3311 | 0.3387 | | 300 이상 | 0.6697 | 0.6840 | 0.3311 | 0.3387 | | 150 이상 | 0.6697 | 0.6840 | 0.3311 | 0.3387 | | 100 이상 | 0.6697 | 0.6840 | 0.3311 | 0.3387 | | 80 이상 | 0.6697 | 0.6750 | 0.3311 | 0.3323 | | 90 이상 | 0.6697 | 0.6840 | 0.331 |  |  1. 지식산업센터명   텍스트, 스크린샷, 그래프, 라인이(가) 표시된 사진  자동 생성된 설명  지식산업센터명의 데이터의 경우 대부분은 빈 값으로 존재하였으며, 다양한 지식산업센터명이 순위를 차지하는 것을 볼 수 있다.  [UNIFY]   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Accuracy | | F1-Score | | | UNIFY 범위 | XGBOOST | LGBM | XGBOOST | LGBM | | 30개 이하 | 0.6267 | 0.6400 | 0.3106 | 0.3250 | | 10개 이하 | 0.6686 | 0.6697 | 0.3377 | 0.3298 | | 9개 이하 | 0.6607 | 0.6676 | 0.3216 | 0.3239 | | 8개 이하 | 0.6676 | 0.6723 | 0.3425 | 0.3331 | | 7개 이하 | 0.6633 | 0.6824 | 0.3310 | 0.3142 | | 6개 이하 | 0.6691 | 0.6787 | 0.3394 | 0.3370 | | 5개 이하 | 0.6644 | 0.6803 | 0.3323 | 0.3379 | | 4개 이하 | 0.6750 | 0.6766 | 0.3487 | 0.3336 | | 3개 이하 | 0.6686 | 0.6787 | 0.3282 | 0.3427 | | 2개 이하 | 0.6744 | 0.6797 | 0.3404 | 0.3420 |  1. 입주형태   텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진  자동 생성된 설명  입주형태 데이터의 경우 대부분이 ‘해당없음’이며 일반건축물, 분양, 공장임대가 그 순위를 차지했다.  [UNIFY]   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Accuracy | | F1-Score | | | UNIFY 범위 | XGBOOST | LGBM | XGBOOST | LGBM | | 400개 이하 | 0.6644 | 0.3723 | 0.3299 | 0.3235 | | 300개 이하 | 0.6670 | 0.6797 | 0.3298 | 0.3300 | | 200개 이하 | 0.6670 | 0.6797 | 0.3298 | 0.3300 | | 81개 이하 | 0.6670 | 0.6797 | 0.3298 | 0.3300 | | **80개 이하** | **0.6697** | **0.6829** | **0.3405** | **0.3448** |  1. 용도지역   텍스트, 스크린샷, 도표, 라인이(가) 표시된 사진  자동 생성된 설명  용도지역 데이터의 경우 다양한 값들이 존재하였으며, 빈 데이터가 많아고 도시지역으로 시작하는 데이터가 대부분을 차지한 것을 볼 수 있다.  [UNIFY]   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Accuracy | | F1-Score | | | UNIFY 범위 | XGBOOST | LGBM | XGBOOST | LGBM | | 41개 이하 | 0.6729 | 0.6824 | 0.3483 | 0.3371 | | 20개 이하 | 0.6670 | 0.6771 | 0.3361 | 0.3374 | | 15개 이하 | 0.6697 | 0.6840 | 0.3311 | 0.3387 | | 6개 이하 | 0.6697 | 0.6840 | 0.3311 | 0.3387 | | 2개 이하 | 0.6697 | 0.6840 | 0.3311 | 0.3387 |  1. 지목   텍스트, 스크린샷, 그래프, 라인이(가) 표시된 사진  자동 생성된 설명  지목 데이터의 경우 ‘대’, ‘공장용지’의 속성이 가장 많음을 나타내고 있다. 나머지 데이터의 경우 소규모 데이터로 존재한다.  [UNIFY]   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Accuracy | | F1-Score | | | UNIFY 범위 | XGBOOST | LGBM | XGBOOST | LGBM | | 25개 이하 | 0.6644 | 0.6720 | 0.3443 | 0.3401 | | 24개 이하 | 0.6660 | 0.6760 | 0.3451 | 0.3401 | | 21개 이하 | 0.6723 | 0.6755 | 0.3448 | 0.3397 | | 8개 이하 | 0.6628 | 0.6723 | 0.3317 | 0.3338 | | 5개 이하 | 0.6744 | 0.6797 | 0.3448 | 0.3428 | | 4개 이하 | 0.6739 | 0.6808 | 0.3386 | 0.3486 | | 3개 이하 | 0.6729 | 0.6797 | 0.3453 | 0.3422 | | 2개 이하 | 0.6729 | 0.6979 | 0.3453 | 0.3422 |   다양한 경우의 수로 EDA를 진행한 결과, 전처리를 진행하지 않은 모델의 최대 성능은 0.6829로 나타났습니다. 이는 전처리를 진행하지 않은 모델의 성능인 0.6795와 비교했을 때 0.003의 정확도 향상이 있었으나**, 유의미한 차이는 아니라고 판단되었습니다. 따라서 전처리를 적용하지 않고 모델을 생성하기로 결정하였습니다.**   1. **최적의 하이퍼파라미터 (Optuna):** XGBoost, LGBM, CatBoost, KNN 모델을 테스트한 결과, XGBoost와 LGBM의 정확도가 가장 높았습니다. 따라서 이 두 모델의 최적 하이퍼파라미터를 찾아 모델을 생성하고자 합니다. XGBClassifier와 LGBMClassifier 모델의 하이퍼파라미터 최적화는 Optuna를 사용하여 수행되었습니다. 각각의 하이퍼파라미터와 그 의미는 다음과 같습니다.  XGBClassifier  1. **max\_depth**: 트리의 최대 깊이. 트리의 복잡성을 제어하여 과적합을 방지합니다.    * 범위: 1 ~ 9 2. **learning\_rate**: 학습률. 학습 속도를 조절하며, 작은 값을 사용하면 더 세밀하게 학습할 수 있습니다.    * 범위: 1e-8 ~ 1.0 (로그 스케일) 3. **n\_estimators**: 부스팅 반복 횟수. 부스팅 트리의 개수를 의미합니다.    * 범위: 50 ~ 500 4. **subsample**: 각 트리 학습에 사용되는 데이터 샘플의 비율. 과적합을 방지하기 위해 사용됩니다.    * 범위: 0.5 ~ 1.0 5. **colsample\_bytree**: 각 트리 학습에 사용되는 피처 샘플의 비율입니다.    * 범위: 0.5 ~ 1.0 6. **gamma**: 트리 분할 시 리프 노드의 최소 손실 감소 값. 큰 값을 설정하면 모델이 더 보수적으로 분할합니다.    * 범위: 0 ~ 10 7. **min\_child\_weight**: 리프 노드가 가질 수 있는 최소 가중치 합. 과적합을 방지합니다.    * 범위: 1 ~ 10 8. **reg\_alpha**: L1 정규화 항의 가중치입니다.    * 범위: 0 ~ 10 9. **reg\_lambda**: L2 정규화 항의 가중치입니다.    * 범위: 0 ~ 10  LGBMClassifier  1. **num\_leaves**: 하나의 트리가 가질 수 있는 최대 리프 수. 트리의 복잡성을 제어합니다.    * 범위: 2 ~ 256 2. **learning\_rate**: 학습률. 학습 속도를 조절하며, 작은 값을 사용하면 더 세밀하게 학습할 수 있습니다.    * 범위: 1e-8 ~ 1.0 (로그 스케일) 3. **n\_estimators**: 부스팅 반복 횟수. 부스팅 트리의 개수를 의미합니다.    * 범위: 50 ~ 500 4. **min\_child\_samples**: 리프 노드가 가질 수 있는 최소 샘플 수. 과적합을 방지합니다.    * 범위: 5 ~ 100 5. **subsample**: 각 트리 학습에 사용되는 데이터 샘플의 비율. 과적합을 방지하기 위해 사용됩니다.    * 범위: 0.5 ~ 1.0 6. **colsample\_bytree**: 각 트리 학습에 사용되는 피처 샘플의 비율입니다.    * 범위: 0.5 ~ 1.0 7. **reg\_alpha**: L1 정규화 항의 가중치입니다.    * 범위: 0 ~ 10 8. **reg\_lambda**: L2 정규화 항의 가중치입니다.    * 범위: 0 ~ 10 9. **min\_split\_gain**: 분할 시 필요한 최소 손실 감소 값. 큰 값을 설정하면 모델이 더 보수적으로 분할합니다.    * 범위: 0 ~ 10  **하이퍼파라미터 최적화 결과** Optuna를 통해 XGBClassifier와 LGBMClassifier의 하이퍼파라미터를 최적화한 결과는 다음과 같습니다. **XGBClassifier**  * **최적 하이퍼파라미터**:   + max\_depth: 5   + learning\_rate: 0.1300918726059749   + n\_estimators: 220   + subsample: 0.8604752773818702   + colsample\_bytree: 0.947536499090015   + gamma: 0.28310148651580475   + min\_child\_weight: 2   + reg\_alpha: 0.04775304383167353   + reg\_lambda: 3.1796604958721355 * **최적 성능 (Best score)**: 0.673407482305359  **LGBMClassifier**  * **최적 하이퍼파라미터**:   + num\_leaves: 183   + learning\_rate: 0.05301025347665236   + n\_estimators: 267   + min\_child\_samples: 49   + subsample: 0.8280708153469658   + colsample\_bytree: 0.8115115502834236   + reg\_alpha: 0.24383119653567353   + reg\_lambda: 6.961964813669102   + min\_split\_gain: 0.6047767141120094 * **최적 성능 (Best score)**: 0.6830131445904954   LGBM 모델은 Optuna를 적용한 결과 정확도 0.68로 가장 높은 성능을 보였습니다. 이를 바탕으로 모델의 안정성과 정확도를 높이기 위해 K-Fold 교차 검증을 적용하였습니다.   |  |  | | --- | --- | | 모델 | Voting Ensemble Test Accuracy | | Base Model | 0.6830 | | K-Fold: 5 | 0.6990 ± 0.0039 | | K-Fold: 10 | 0.7037 ± 0.0083 | | **K-Fold: 15** | **0.7040 ± 0.0075** | | K-Fold: 20 | 0.7030 ± 0.0158 |   K-Fold 교차 검증을 통해 최종적으로 0.70의 정확도를 가진 안정적인 모델을 도출하였습니다. Cross Validation은 모델의 성능을 더 안정적이고 신뢰성 있게 평가하기 위한 기법으로, 데이터를 여러 개의 폴드로 나누어 여러 번 학습과 평가를 반복하여 모델의 일반화 성능을 평가합니다. 이를 통해 최종 모델의 성능을 최적화할 수 있었습니다.   1. **앙상블:** 모델 성능을 더욱 향상시키기 위해 앙상블 기법을 활용하려고 합니다. 앙상블 기법은 여러 개의 모델을 결합하여 하나의 모델보다 더 나은 성능을 얻고자 하는 방법입니다. 각각의 모델이 가진 약점을 보완하고 강점을 취함으로써, 예측의 정확도를 높이고, 과적합을 줄이며, 일반화 성능을 향상시키는 데 사용됩니다.  Voting (보팅): 보팅은 여러 모델의 예측 결과를 결합하여 최종 예측을 만드는 방법입니다. 각각의 개별 모델이 독립적으로 예측을 수행한 후, 이 예측들을 결합하여 최종 결과를 도출합니다. 소프트 보팅 방식을 활용하여 각각의 모델이 출력한 클래스의 확률을 평균 내어 가장 높은 확률을 가진 클래스를 최종 예측으로 선택합니다. XGBClassifier와 LGBMClassifier의 최적 하이퍼파라미터 모델을 결합하여 Voting 방식을 적용해 모델의 정확도를 테스트하였습니다. 각 모델의 최적 하이퍼파라미터는 다음과 같습니다.  **XGBClassifier의 최적 하이퍼파라미터:**   * colsample\_bytree: 0.947536499090015 * gamma: 0.28310148651580475 * learning\_rate: 0.1300918726059749 * max\_depth: 5 * min\_child\_weight: 2 * n\_estimators: 220   **LGBMClassifier의 최적 하이퍼파라미터:**   * colsample\_bytree: 0.8115115502834236 * learning\_rate: 0.05301025347665236 * min\_child\_samples: 49 * min\_split\_gain: 0.6047767141120094 * n\_estimators: 267 * num\_leaves: 183 * reg\_alpha: 0.24383119653567353 * reg\_lambda: 6.961964813669102 * subsample: 0.8280708153469658   **Voting Ensemble의 정확도는 다음과 같습니다:**   |  |  | | --- | --- | | 모델 | Voting Ensemble Test Accuracy | | Base Model | 0.6819 | | K-Fold: 5 | 0.6966 ± 0.0074 | | K-Fold: 10 | 0.7013 ± 0.0086 | | **K-Fold: 15** | **0.7033 ± 0.0100** | | K-Fold: 20 | 0.7027 ± 0.0178 |   Voting Ensemble Model의 기본 정확도는 0.68로 LGBM Optuna 모델의 결과와 동일하며, K-fold에서는 미세하게 낮은 정확도를 도출한 것으로 나타났습니다. Stacking (스태킹): 스태킹은 여러 모델의 예측 결과를 새로운 모델의 입력으로 사용하여 최종 예측을 만드는 방법입니다. 보팅과 달리 스태킹은 메타 모델을 사용하여 개별 모델들의 예측을 조합합니다.  1. **기본 모델**: 여러 개의 기본 모델을 학습시킵니다. 이 모델들은 서로 다른 알고리즘이나 하이퍼파라미터를 사용할 수 있습니다. 2. **메타 모델**: 기본 모델들이 예측한 결과를 입력으로 받아 최종 예측을 수행하는 모델입니다. 메타 모델은 기본 모델의 예측을 특징으로 사용하여 학습합니다.   LGBMClassifier의 최적 하이퍼파라미터 모델에 메타 모델을 추가하여 Stacking 모델을 테스트하였습니다. 각 모델의 최적 하이퍼파라미터 값은 다음과 같습니다.  **LGBMClassifier의 최적 하이퍼파라미터:**   * colsample\_bytree: 0.7333231436132019 * learning\_rate: 0.12693737659589577 * min\_child\_samples: 49 * min\_split\_gain: 0.27225519315678226 * n\_estimators: 193 * num\_leaves: 25 * reg\_alpha: 1.309780288948493 * reg\_lambda: 0.27599843896866716 * subsample: 0.8352629679425484   메타 모델은 Logistic Regression, RandomForest, Gradient Boosting, MLP Classifier를 사용하여 모델의 안정성을 높이기 위해 K-Fold 5회를 진행하였습니다. 결과는 다음과 같습니다.  **Stacking Model 성능:**   |  |  | | --- | --- | | Stacking Model | Test Accuracy | | Logistic Regression Stacking | 0.6593 | | Logistic Regression Stacking (k-fold 5) | 0.6632 ± 0.0083 | | **Random Forest Stacking** | **0.6605** | | Random Forest Stacking (k-fold 5) | 0.6365 ± 0.0091 | | Gradient Boosting Stacking | 0.06268 | | Gradient Boosting Stacking (k-fold 5) | 0.6189 ± 0.0074 | | MLP Classifier Stacking | 0.6575 | | MLP Classifier Stacking (K-fold 5) | 0.6480 ± 0.0047 |   Stacking Model의 경우, 오히려 성능이 떨어진 것을 확인할 수 있었습니다.   1. **Deep Learning**   딥러닝은 대규모 데이터셋을 통해 복잡한 패턴과 특성을 학습하는 기법으로, 여러 층(layer)으로 구성된 신경망을 통해 입력 데이터를 점진적으로 처리하며, 각 층은 입력 데이터의 추상화된 표현을 학습합니다. 다음과 같은 딥러닝 모델을 정의하여 사용하였습니다.  [모델 정의]   * 첫 번째 층: 입력 크기에서 512 유닛으로 변환, ReLU 활성화 함수와 드롭아웃 적용 * 두 번째 층: 512 유닛에서 256 유닛으로 변환, ReLU 활성화 함수와 드롭아웃 적용 * 세 번째 층: 256 유닛에서 128 유닛으로 변환, ReLU 활성화 함수와 드롭아웃 적용 * 네 번째 층: 128 유닛에서 클래스 개수만큼 변환 (출력층)   [모델 학습]   * **손실 함수**: 교차 엔트로피 손실 함수(nn.CrossEntropyLoss)를 사용 * **옵티마이저**: Adam 옵티마이저(optim.Adam)를 사용하여 모델의 가중치를 업데이트 * **학습 함수**: train\_model 함수는 학습(train)과 검증(validation) 데이터셋을 사용하여 모델을 학습시키고, 에포크(epoch)마다 검증 정확도를 계산하여 가장 높은 검증 정확도를 기록한 모델을 저장 * 각 에포크에서 학습 손실(train loss)과 검증 손실(val loss)을 계산하고, 검증 데이터셋에 대한 정확도(val accuracy)를 평가 * 검증 정확도가 향상될 때마다 모델의 가중치를 저장  |  |  | | --- | --- | | Model | Test Accuracy | | Deep Learning Model | 0.5524 |   최종적으로 0.5524의 정확도를 나타냈습니다. 11,000개의 적은 데이터셋으로 학습을 진행하다 보니, 복잡한 패턴을 학습하지 못하여 정확도가 높게 나오지 않았습니다.   전국 추천 모델 생성 및 인사이트  * 1. **전국 모델 생성**   지금까지는 서울시 공장 데이터로 전처리, 분석을 진행하여 최적의 ‘구’ 추천 모델 생성에 관한 것을 파악하였다. 이를 바탕으로 서울시 뿐만 아니라 전국 지자체의 모델을 생성하여 정확도 및 분석을 진행하였다.  전국 모델을 생성하는데 있어 다음과 같이 진행하였다.   * + 1. 전처리를 진행하여도 성능이 크게 좋아지지 않으며 모델의 통일성을 위해서 전처리는 진행하지 않았다.     2. Optuna를 통해서 최적의 하이퍼 파라미터를 선정하였다.     3. 모델의 안전성 및 성능 향상을 위해 Cross-Validation을 사용하였다.  1. **강원도**  * 강원도 공장 데이터의 경우 총 3812개의 데이터가 존재하였다. 총 18개의 ‘시’, ‘군’으로 분류하였다.      * 대표업종 코드 23322, 27199, 10212, 25112, 26421 레미콘 제조업, 그 외 기타 의료용 기기 제조업, 수산동물 건조 및 염장품 제조업, 구조용 금속 판제품 및 공작물 제조업, 방송장비 제조업 순서로 업종들이 차지하는 것을 볼 수 있다. * 예측 정확도  |  |  | | --- | --- | | Model | Test Accuracy | | 강원도 시/군 예측 기본 모델 | 0.4738 | | K-fold 5 | 0.4932 ± 0.0185 | | K-fold 10 | 0.4987 ± 0.0249 | | **K-fold 15** | **0.4987 ± 0.0262** | | K-fold 20 | 0.4976 ± 0.0428 |   강원도 지역 18개의 시/군 예측 모델의 정확도는 대략 50%로 볼 수 있다.   1. **경기도**  * 경기도 공장 데이터의 경우 총 70301개의 데이터가 존재하였다. 총 398개의 ‘구’, ‘동’, ‘면’, ’읍’으로 분류하였다.        * 대표업종 코드 29294, 26299, 28123, 32029, 25924 주형 및 금형 제조업, 그 외 기타 전자 부품 제조업, 배전반 및 전기 자동제어반 제조업, 기타 목재 가구 제조업, 절삭 가공 및 유사 처리업 순서로 업종들이 차지하는 것을 볼 수 있다. * 예측 정확도  |  |  | | --- | --- | | Model | Test Accuracy | | 경기도 구/동/면/읍 예측 기본 모델 | 0.3614 | | K-fold 5 | 0.3625 ± 0.0029 | | K-fold 10 | 0.3650 ± 0.0041 | | K-fold 15 | 0.3654 ± 0.0049 | | **K-fold 20** | **0.3667 ± 0.0046** |   경기도 지역 398개의 구/동/면/읍 예측 모델의 정확도는 대략 36%로 볼 수 있다.   1. **경상남도**  * 경상남도 공장 데이터의 경우 총 19719개의 데이터가 존재하였다. 총 241개의 ‘구’, ‘동’, ‘면’, ‘읍’으로 분류하였다.      * 대표업종 코드 25924, 30399, 31114, 29294, 25113 절삭 가공 및 유사 처리업, 그 외 자동차용 신품 부품 제조업, 선박 구성 부분품 제조업, 주형 및 금형 제조업, 육상 금속 골조 구조재 제조업 순서로 업종들이 차지하는 것을 볼 수 있다. * 예측 정확도  |  |  | | --- | --- | | Model | Test Accuracy | | 경상남도 구/동/면/읍 예측 기본 모델 | 0.3409 | | K-fold 5 | 0.3480 ± 0.0036 | | **K-fold 10** | **0.3534 ± 0.0079** | | K-fold 15 | 0.3532 ± 0.0120 | | K-fold 20 | 0.3529 ± 0.0132 |   경상남도 지역 241개의 구/동/면/읍 예측 모델의 정확도는 대략 35%로 볼 수 있다.   1. **경상북도**  * 경상북도 공장 데이터의 경우 총 15427개의 데이터가 존재하였다. 총 252개의 ‘구’, ‘동’, ‘면’, ‘읍’으로 분류하였다.      * 대표업종 코드 30399, 25113, 13104, 25924, 29294 그 외 자동차용 신품 부품 제조업, 육상 금속 골조 구조재 제조업, 연사 및 가공사 제조업, 절삭 가공 및 유사 처리업, 주형 및 금형 제조업 순서로 업종들이 차지하는 것을 볼 수 있다. * 예측 정확도  |  |  | | --- | --- | | Model | Test Accuracy | | **경상북도 구/동/면/읍 예측 기본 모델** | **0.3184** | | K-fold 5 | 0.3074 ± 0.0041 | | K-fold 10 | 0.3135 ± 0.0096 | | K-fold 15 | 0.3169 ± 0.0085 | | K-fold 20 | 0.3156 ± 0.0110 |   경상북도 지역 252개의 구/동/면/읍 예측 모델의 정확도는 대략 31%로 볼 수 있다.   1. **광주광역시**  * 광주광역시 공장 데이터의 경우 총 4649개의 데이터가 존재하였다. 총 111개의 ‘구’, ‘동’, ‘면’/읍으로 분류하였다.      * 대표업종 코드 29294, 28123, 25113, 30399, 25111 주형 및 금형 제조업, 배전반 및 전기 자동제어반 제조업, 육상 금속 골조 구조재 제조업, 그 외 자동차용 신품 부품 제조업, 금속 문, 창, 셔터 및 관련제품 제조업 순서로 업종들이 차지하는 것을 볼 수 있다. * 예측 정확도  |  |  | | --- | --- | | Model | Test Accuracy | | 광주광역시 구/동/면/읍 예측 기본 모델 | 0.3902 | | K-fold 5 | 0.3902 ± 0.0040 | | **K-fold 10** | **0.3987 ± 0.0080** | | K-fold 15 | 0.3922 ± 0.0251 | | K-fold 20 | 0.3930 ± 0.0249 |   광주광역시 지역 111개의 구/동/면/읍 예측 모델의 정확도는 대략 40%로 볼 수 있다.   1. **대구광역시**  * 대구광역시 공장 데이터의 경우 총 8208개의 데이터가 존재하였다. 총 95개의 ‘구’, ‘동’, ‘면’, ‘읍’으로 분류하였다.      * 대표업종 코드 30399, 29294, 25924, 13213, 28123 그 외 자동차용 신품 부품 제조업, 주형 및 금형 제조업, 절삭 가공 및 유사 처리업, 화학섬유직물 직조업, 배전반 및 전기 자동제어반 제조업 순서로 업종들이 차지하는 것을 볼 수 있다. * 예측 정확도  |  |  | | --- | --- | | Model | Test Accuracy | | 대구광역시 구/동/면/읍 예측 기본 모델 | 0.4588 | | K-fold 5 | 0.4522 ± 0.0078 | | **K-fold 10** | **0.4611 ± 0.0120** | | K-fold 15 | 0.4602 ± 0.174 | | K-fold 20 | 0.4585 ± 0.0202 |   대구광역시 지역 95개의 구/동/면/읍 예측 모델의 정확도는 대략 46%로 볼 수 있다.   1. **대전광역시**  * 대전광역시 공장 데이터의 경우 총 3225개의 데이터가 존재하였다. 총 107개의 ‘구’, ‘동’, ‘면’, ‘읍’으로 분류하였다.      * 대표업종 코드 28123, 29299, 29199, 26299, 27213 배전반 및 전기 자동제어반 제조업, 그 외 기타 특수 목적용 기계 제조업, 그 외 기타 일반 목적용 기계 제조업, 그 외 기타 전자 부품 제조업, 물질 검사, 측정 및 분석 기구 제조업 순서로 업종들이 차지하는 것을 볼 수 있다. * 예측 정확도  |  |  | | --- | --- | | Model | Test Accuracy | | 대전광역시 구/동/면/읍 예측 기본 모델 | 0.5074 | | K-fold 5 | 0.5049 ± 0.0154 | | K-fold 10 | 0.5087 ± 0.0286 | | **K-fold 15** | **0.5096 ± 0.0297** | | K-fold 20 | 0.5080 ± 0.0321 |   대전광역시 지역 107개의 구/동/면/읍 예측 모델의 정확도는 대략 50%로 볼 수 있다.   1. **부산광역시**  * 부산광역시 공장 데이터의 경우 총 10443개의 데이터가 존재하였다. 총 113개의 ‘구’, ‘동’, ‘면’, ‘읍’으로 분류하였다.      * 대표업종 코드 31114, 25924, 28123, 29294, 30399 선박 구성 부분품 제조업, 절삭 가공 및 유사 처리업, 배전반 및 전기 자동제어반 제조업, 주형 및 금형 제조업, 물질 검사, 그 외 자동차용 신품 부품 제조업 순서로 업종들이 차지하는 것을 볼 수 있다. * 예측 정확도  |  |  | | --- | --- | | Model | Test Accuracy | | **부산광역시 구/동/면/읍 예측 기본 모델** | **0.4756** | | K-fold 5 | 0.4563 ± 0.0147 | | K-fold 10 | 0.4602 ± 0.0146 | | K-fold 15 | 0.4631 ± 0.0150 | | K-fold 20 | 0.4648 ± 0.0193 |   부산광역시 지역 113개의 구/동/면/읍 예측 모델의 정확도는 대략 46%로 볼 수 있다.   1. **서울특별시**  * 서울특별시 공장 데이터의 경우 총 11633개의 데이터가 존재하였다. 총 25개의 ‘구’로 분류하였다.      * 대표업종 코드 18119, 28123, 14192, 26421, 26410 기타 인쇄업, 배전반 및 전기 자동제어반 제조업, 근무복, 작업복 및 유사 의복 제조업, 방송장비 제조업, 유선 통신장비 제조업 순서로 업종들이 차지하는 것을 볼 수 있다. * 예측 정확도  |  |  | | --- | --- | | Model | Test Accuracy | | 서울특별시 구 예측 기본 모델 | 0.6814 | | K-fold 5 | 0.6990 ± 0.0039 | | K-fold 10 | 0.7037 ± 0.0083 | | **K-fold 15** | **0.7040 ± 0.0075** | | K-fold 20 | 0.7030 ± 0.00158 |   서울특별시 지역 25개의 구 예측 모델의 정확도는 대략 70%로 볼 수 있다.   1. **세종특별자치시**  * 세종특별자치시 공장 데이터의 경우 총 816개의 데이터가 존재하였다. 총 70개의 ‘리’로 분류하였다. * 대표업종 코드 18111, 25112, 30399, 25113, 22211 경 인쇄업, 구조용 금속 판제품 및 공작물 제조업, 그 외 자동차용 신품 부품 제조업, 육상 금속 골조 구조재 제조업, 플라스틱 선, 봉, 관 및 호스 제조업 순서로 업종들이 차지하는 것을 볼 수 있다. * 예측 정확도  |  |  | | --- | --- | | Model | Test Accuracy | | 세종특별자치시 리 예측 기본 모델 | 0.1376 | | **K-fold 5** | **0.2127 ± 0.0228** | | K-fold 10 | 0.1991 ± 0.538 | | K-fold 15 | 0.1839± 0.0448 | | K-fold 20 | 0.1881 ± 0.0628 |   세종특별자치시 지역 113개의 리 예측 모델의 정확도는 대략 21%로 볼 수 있다.   1. **울산광역시**  * 울산광역시 공장 데이터의 경우 총 2895개의 데이터가 존재하였다. 총 73개의 ‘동’, ‘면’, ‘읍’으로 분류하였다.      * 대표업종 코드 30399, 28123, 31114, 25113, 30320 그 외 자동차용 신품 부품 제조업, 배전반 및 전기 자동제어반 제조업, 선박 구성 부분품 제조업, 육상 금속 골조 구조재 제조업, 자동차 차체용 신품 부품 제조업 순서로 업종들이 차지하는 것을 볼 수 있다. * 예측 정확도  |  |  | | --- | --- | | Model | Test Accuracy | | 울산광역시 동/면/읍 예측 기본 모델 | 0.4333 | | **K-fold 5** | **0.4615 ± 0.0132** | | K-fold 10 | 0.3264 ± 0.0276 | | K-fold 15 | 0.3215 ± 0.0223 | | K-fold 20 | 0.3222 ± 0.0291 |   울산광역시 지역 73개의 동/면/읍 예측 모델의 정확도는 대략 46%로 볼 수 있다.   1. **인천광역시**  * 인천광역시 공장 데이터의 경우 총 12175개의 데이터가 존재하였다. 총 73개의 ‘동’으로 분류하였다.      * 대표업종 코드 25924, 29294, 25922, 26299, 28123 절삭 가공 및 유사 처리업, 주형 및 금형 제조업, 도금업, 그 외 기타 전자 부품 제조업, 배전반 및 전기 자동제어반 제조업 순서로 업종들이 차지하는 것을 볼 수 있다. * 예측 정확도  |  |  | | --- | --- | | Model | Test Accuracy | | 인천광역시 동 예측 기본 모델 | 0.5646 | | K-fold 5 | 0.5820 ± 0.050 | | K-fold 10 | 0.5840 ± 0.0050 | | K-fold 15 | 0.5833± 0.0098 | | **K-fold 20** | **0.5858± 0.0176** |   인천광역시 지역 101개의 동 예측 모델의 정확도는 대략 58%로 볼 수 있다.   1. **전라남도**  * 전라남도 공장 데이터의 경우 총 7183개의 데이터가 존재하였다. 총 261개의 ‘동’, ‘읍’, ‘면’으로 분류하였다.      * 대표업종 코드 10220, 25113, 31114, 28123, 10611 수산식물 가공 및 저장 처리업, 육상 금속 골조 구조재 제조업, 선박 구성 부분품 제조업, 배전반 및 전기 자동제어반 제조업, 곡물 도정업 순서로 업종들이 차지하는 것을 볼 수 있다. * 예측 정확도  |  |  | | --- | --- | | Model | Test Accuracy | | 전라남도 동/읍/면 예측 기본 모델 | 0.2614 | | K-fold 5 | 0.2737 ± 0.0055 | | K-fold 10 | 0.2818 ± 0.0093 | | K-fold 15 | 0.2820 ± 0.0122 | | **K-fold 20** | **0.2841 ± 0.0175** |   전라남도 지역 261개의 동/읍/면 예측 모델의 정확도는 대략 28%로 볼 수 있다.   1. **전라북도**  * 전라북도 공장 데이터의 경우 총 7002개의 데이터가 존재하였다. 총 239개의 ‘구’, ‘동’, ‘읍’, ‘면’으로 분류하였다.      * 대표업종 코드 30399, 25113, 23911, 10611, 25112 그 외 자동차용 신품 부품 제조업, 육상 금속 골조 구조재 제조업, 건설용 석제품 제조업, 곡물 도정업, 구조용 금속 판제품 및 공작물 제조업 순서로 업종들이 차지하는 것을 볼 수 있다. * 예측 정확도  |  |  | | --- | --- | | Model | Test Accuracy | | 전라북도 구/동/읍/면 예측 기본 모델 | 0.3186 | | K-fold 5 | 0.3274 ± 0.0111 | | K-fold 10 | 0.3316 ± 0.0107 | | **K-fold 15** | **0.3358± 0.0181** | | K-fold 20 | 0.3298 ± 0.0210 |   전라북도 지역 239개의 구/동/읍/면 예측 모델의 정확도는 대략 33%로 볼 수 있다.   1. **제주특별자치도**  * 제주특별자치도 공장 데이터의 경우 총 693개의 데이터가 존재하였다. 총 37개의 ‘동’, ‘읍’, ‘면’으로 분류하였다.      * 대표업종 코드 28123, 10309, 23322, 10129, 10219 배전반 및 전기 자동제어반 제조업, 기타 과실ㆍ채소 가공 및 저장 처리업, 레미콘 제조업, 육류 기타 가공 및 저장 처리업(가금류 제외), 기타 수산동물 가공 및 저장 처리업 순서로 업종들이 차지하는 것을 볼 수 있다. * 예측 정확도  |  |  | | --- | --- | | Model | Test Accuracy | | **제주특별자치도 동/읍/면 예측 기본 모델** | **0.3434** | | K-fold 5 | 0.2833 ± 0.0103 | | K-fold 10 | 0.2939 ± 0.0289 | | K-fold 15 | 0.2818± 0.0430 | | K-fold 20 | 0.2924± 0.0815 |   전라북도 지역 239개의 구/동/읍/면 예측 모델의 정확도는 대략 34%로 볼 수 있다.   1. **충청남도**  * 충청남도 공장 데이터의 경우 총 10754의 데이터가 존재하였다. 총 193개의 ‘동’, ‘읍’, ‘면’으로 분류하였다.      * 대표업종 코드 30399, 25113, 29271, 28123, 25112 그 외 자동차용 신품 부품 제조업, 육상 금속 골조 구조재 제조업, 반도체 제조용 기계 제조업, 배전반 및 전기 자동제어반 제조업, 구조용 금속 판제품 및 공작물 제조업 순서로 업종들이 차지하는 것을 볼 수 있다. * 예측 정확도  |  |  | | --- | --- | | Model | Test Accuracy | | 충청남도 동/읍/면 예측 기본 모델 | 0.2484 | | K-fold 5 | 0.2539 ± 0.0095 | | K-fold 10 | 0.2624 ± 0.0117 | | K-fold 15 | 0.26000 ± 0.0190 | | **K-fold 20** | **0.2641± 0.0152** |   충청남도 지역 239개의 동/읍/면 예측 모델의 정확도는 대략 26%로 볼 수 있으며 높은 정확도인 것을 나타내고 있다.   1. **충청북도**  * 충청북도 공장 데이터의 경우 총 9014의 데이터가 존재하였다. 총 107개의 ‘동’, ‘읍’, ‘면’으로 분류하였다.      * 대표업종 코드 25113, 25112, 25111, 28123, 10309 육상 금속 골조 구조재 제조업, 구조용 금속 판제품 및 공작물 제조업 , 금속 문, 창, 셔터 및 관련제품 제조업, 배전반 및 전기 자동제어반 제조업, 기타 과실ㆍ채소 가공 및 저장 처리업 순서로 업종들이 차지하는 것을 볼 수 있다. * 예측 정확도  |  |  | | --- | --- | | Model | Test Accuracy | | 충청북도 동/읍/면 예측 기본 모델 | 0.2135 | | K-fold 5 | 0.2366 ± 0.0088 | | K-fold 10 | 0.2419 ± 0.0126 | | **K-fold 15** | **0.2438 ± 0.0134** | | K-fold 20 | 0.2423 ± 0.0142 |   충청북도 지역 107개의 동/읍/면 예측 모델의 정확도는 대략 24%로 볼 수 있다.  **전국 부지 추천 모델 정확도 및 인사이트**  전국 부지 추천 모델의 정확도 및 대표업종은 다음과 같다.   |  |  |  |  | | --- | --- | --- | --- | | 순위 (정확도 기준) | 지자체 | 정확도 | 대표업종 | | 1 | 서울특별시 | 0.70 | - 기타 인쇄업  - 배전반 및 전기 자동제어반 제조업  - 근무복, 작업복 및 유사 의복 제조업  - 방송장비 제조업  - 유선 통신장비 제조업 | | 2 | 인천광역시 | 0.585 | - 절삭 가공 및 유사 처리업  - 주형 및 금형 제조업  - 도금업  - 그 외 기타 전자 부품 제조업  - 배전반 및 전기 자동제어반 제조업 | | 3 | 대전광역시 | 0.509 | - 배전반 및 전기 자동제어반 제조업  - 그 외 기타 특수 목적용 기계 제조업  - 그 외 기타 일반 목적용 기계 제조업  - 그 외 기타 전자 부품 제조업  - 물질 검사, 측정 및 분석 기구 제조업 | | 4 | 강원도 | 0.5 | - 레미콘 제조업  - 기타 의료용 기기 제조업  - 수산동물 건조 및 염장품 제조업  - 구조용 금속 판제품 및 공작물 제조업  - 방송장비 제조업 | | 5 | 부산광역시 | 0.475 | - 선박 구성 부분품 제조업  - 절삭 가공 및 유사 처리업  - 배전반 및 전기 자동제어반 제조업  - 주형 및 금형 제조업  - 물질 검사  -그 외 자동차용 신품 부품 제조업 | | 6 | 울산광역시 | 0.461 | - 그 외 자동차용 신품 부품 제조업  - 배전반 및 전기 자동제어반 제조업  - 선박 구성 부분품 제조업  - 육상 금속 골조 구조재 제조업  - 자동차 차체용 신품 부품 제조업 | | 7 | 대구광역시 | 0.461 | - 그 외 자동차용 신품 부품 제조업  - 주형 및 금형 제조업  - 절삭 가공 및 유사 처리업  - 화학섬유직물 직조업  - 배전반 및 전기 자동제어반 제조업 | | 8 | 광주광역시 | 0.40 | - 주형 및 금형 제조업  - 배전반 및 전기자동제어반 제조업  - 육상 금속 골조 구조재 제조업  - 그 외 자동차용 신품 부품 제조업  - 금속 문, 창, 셔터 및 관련제품 제조업 | | 9 | 경상남도 | 0.353 | - 절삭 가공 및 유사 처리업  - 그 외 자동차용 신품 부품 제조업  - 선박 구성 부분품 제조업  - 주형 및 금형 제조업  - 육상 금속 골조 구조재 제조업 | | 10 | 제주특별자치도 | 0.343 | - 배전반 및 전기 자동제어반 제조업  - 기타 과실ㆍ채소 가공 및 저장ㅊ ㅓ리업  - 레미콘 제조업  - 육류 기타 가공 및 저장 처리업 (가금류 제외)  - 기타 수산동물 가공 및 저장 처리업 | | 11 | 경기도 | 0.336 | -주형 및 금형 제조업  - 그 외 기타 전자 부품 제조업  - 배전반 및 전기 자동제어반 제조업  - 기타 목재 가구 제조업  - 절삭 가공 및 유사 처리업 | | 12 | 전라북도 | 0.335 | - 그 외 자동차용 신품 부품 제조업  - 육상 금속 골조 구조개 제조업  - 건설용 석제품 제조업  - 곡물 도정업  - 구조용 금속 판제품 및 공작물 제조업 | | 13 | 경상북도 | 0.318 | - 그 외 자동차용 신품 부품 제조업  - 육상 금속 골조 구조재 제조업  - 연사 및 가공사 제조업  - 절삭 가공 및 유사 처리업  -주형 및 금형 제조업 | | 14 | 전라남도 | 0.284 | - 수산식물 가공 및 저장 처리업  - 육상금속 골조 구조재 제조업  - 선박 구성 부분품 제조업  - 배전반 및 전기 자동제어반 제조업  - 곡물 도정업 | | 15 | 충청남도 | 0.264 | - 그 외 자동차용 신품 부품 제조업  - 육상 금속 골조 구조재 제조업  - 반도체 제조용 기계 제조업  - 배전반 및 전기 자동제어반 제조업  - 구조용 금속 판제품 및 공작물 제조업 | | 16 | 충청북도 | 0.243 | - 육상 금속 골조 구조재 제조업  - 구조용 금속 판제품 및 공작물 제조업  - 금속 문, 창 셔터 및 관련제품 제조업  - 배전반 및 전기 자동제어반 제조업  - 과실ㆍ채소 가공 및 저장 처리업 | | 17 | 세종특별자치시 | 0.212 | - 경 인쇄업  - 구조용 금속 판제품 및 공작물 제조업  - 그 외 자동차용 신품 부품 제조업  - 육상 금속 골조 구조재 제조업  - 플라스틱 선, 봉, 관 및 호스 제조업 |  모델 성능 분석 모델의 예측 정확도는 다음과 같습니다:   * 최고 정확도: 서울특별시 (0.70) * 최저 정확도: 세종특별자치시 (0.212)   모델의 정확도는 지역별로 상이하며, 주요 도시들의 정확도는 다음과 같습니다:   * 서울특별시: 0.70 * 인천광역시: 0.585 * 대전광역시: 0.509 * 강원도: 0.5 * 부산광역시: 0.475  1. **모델의 개선된 부분**: 기존의 주관적인 방식에 비해 본 모델은 여러 측면에서 개선된 결과를 보여주고 있습니다:  * **데이터 기반의 의사 결정:** 기존 주관적 판단에 의존하던 방식에서 벗어나, 데이터를 기반으로 객관적인 부지 추천이 가능합니다. 이는 더 신뢰할 수 있는 결정을 지원합니다. * **효율성 증가:** 모델을 활용하여 공장 부지를 추천함으로써, 부지 선정 과정에서 소요되는 시간과 노력을 크게 절감할 수 있습니다.  1. **실질적인 이점과 가치:** 모델의 예측이 완벽하지 않더라도, 이를 통해 얻을 수 있는 여러 실질적인 이점과 가치를 보여주고 있습니다.  * **대표업종 파악:** 각 지역에서 존재하는 많은 대표업종 공장을 파악할 수 있었습니다. 이는 각 지역의 산업적 특성을 이해하는 데 큰 도움을 줍니다. * **데이터의 양과 정확도:** 서울특별시는 약 10,000개의 데이터로 25개의 구를 예측하여 가장 높은 정확도를 보인 반면, 세종특별자치시는 800개의 데이터로 70개의 '리' 위치를 예측하여 정확도가 낮았습니다. 이는 데이터의 양과 예측 단위의 차이가 모델 정확도에 영향을 미친다는 것을 시사합니다. * **객관적인 부지 추천:** 기존의 주관적 판단으로 공장 부지를 선정하는 것보다 데이터를 통해서 객관적인 부지 추천이 가능합니다. 특히 대표업종, 면적, 종업원의 수를 우선적으로 고려하여 부지를 추천함으로써 더 합리적인 결정을 내릴 수 있습니다. * **지속적인 개선 가능성:** 모델은 지속적으로 데이터를 학습하고 개선될 수 있습니다. 시간이 지남에 따라 더 높은 정확도의 예측을 제공할 수 있으며, 이는 장기적으로 큰 이점을 가져다줄 것입니다. * **데이터의 잠재적 가치:** 수집된 데이터는 다양한 분석과 인사이트 도출에 활용될 수 있습니다. 이는 부지 선정뿐만 아니라 다른 경영 의사 결정에도 유용한 정보를 제공합니다.  1. **4공장 부지 추천의 유연성:** 공장 부지 추천의 경우 정답이 존재하지 않으므로, 다양한 사용자 의견(면적, 종업원 수, 지자체)을 통해 가장 적합한 지역을 추천합니다. 공장 부지 추천은 단순히 정확도로만 판단할 수 없으며, 다양한 요인을 고려해야 합니다. 이로 인해 데이터 기반의 추천 시스템은 다양한 조건을 반영하여 사용자 맞춤형 추천을 제공합니다.  3. 공장 부지 매물 데이터 수집 전국적으로 공장 부지 추천을 위해 다양한 공장 매물 데이터를 수집하였습니다. 주요 데이터 출처는 다음 두 곳으로, 각각의 특징과 수집된 데이터를 아래와 같이 정리하였습니다. 1. 공장 통 닷컴 (http://www.okfactory.com/)  * **총 매물 수:** 5,212개 * **수집된 칼럼:**   + **매물위치:** 매물이 위치한 주소   + **종류:** 매물의 유형 (예: 공장, 창고 등)   + **제목:** 매물의 제목   + **상태:** 매물의 현재 상태 (예: 매매 중, 임대 중 등)   + **가격:** 매물의 가격 정보   + **면적:** 매물의 면적   + **방식:** 거래 방식 (예: 매매, 월세)   + **참고사항:** 추가적인 정보나 설명   + **조회수:** 매물의 조회 수   + **매물위치(URL):** 매물 상세 페이지 URL  2. 부동산 119 (https://www.bd119.com/)  * **총 매물 수:** 542개 * **수집된 칼럼:**   + **종류:** 매물의 유형 (예: 공장, 창고 등)   + **매물위치:** 매물이 위치한 주소   + **면적:** 매물의 면적   + **가격:** 매물의 가격 정보   + **층수(방 수):** 매물이 위치한 층수 및 방 수   + **진입도로:** 매물로 진입할 수 있는 도로 정보   + **주차 여부:** 매물의 주차 가능 여부   + **조회수:** 매물의 조회 수   + **제목:** 매물의 제목   + **장점:** 매물의 장점이나 특징   + **매물위치(URL):** 매물 상세 페이지 URL  제외된 데이터 출처 데이터 수집 과정에서 몇몇 부동산 사이트는 다음과 같은 이유로 크롤링 대상에서 제외되었습니다:   1. **부동산 114:** 공장 매물에 대한 정보가 부족하고, 공장 전문 매물 사이트가 아닌 일반 부동산 매물이 많아 제외하였습니다. 2. **공장 다나와:** 특정 지역(예: 경기, 화성)에 편중된 매물들이 많아 전국적인 데이터를 수집하기 어려워 제외하였습니다. 3. **부동산 뱅크:** 매물 설명이 불명확하여 제외하였습니다.  데이터 수집 방법 위의 두 사이트에서 데이터를 크롤링하여 수집하였습니다. 수집된 데이터는 공장 부지 추천 모델에 입력되어, 사용자에게 최적의 공장 부지를 추천하는 데 활용됩니다. 각 매물의 다양한 속성 정보는 모델이 공장 부지를 추천하는 데 있어 중요한 역할을 하며, 특히 매물 위치, 면적, 가격 등의 정보는 매물 선택에 큰 기여를 할 것이라고 판단됩니다. 데이터 활용 수집된 공장 부지 매물 데이터는 다음과 같은 방식으로 활용됩니다:   * **매물 추천:** 사용자가 특정 지역에 공장 부지를 필요로 할 때, 해당 지역의 매물 데이터를 바탕으로 여러 매물을 추천합니다. * **데이터 업데이트:** 지속적으로 새로운 매물 데이터를 수집하여 모델을 최신 상태로 유지하고, 추천의 품질을 개선합니다.  1. 업종 코드 기반 GPT 4.0 API 활용 설명   사용자로부터 특정 업종 코드를 입력 받아 해당 업종의 장점, 단점, 전망에 대한 정보를 제공하는 기능을 포함하고 있습니다. 이를 위해 GPT 4.0 API를 활용하여 업종에 대한 상세한 분석과 설명을 제공합니다. 이 방법은 업종별 특성에 대한 심도 있는 정보를 사용자에게 제공함으로써, 사용자들이 보다 나은 의사결정을 할 수 있도록 지원합니다.   **사용자 입력 받기**   * 사용자로부터 업종 코드를 입력받습니다. 업종 코드는 산업 분류에 따라 각 업종을 식별할 수 있는 고유한 코드입니다.    **GPT 4.0 API 호출 준비**   * 업종 코드가 입력되면, 해당 코드를 기반으로 GPT 4.0 API에 요청을 준비합니다. 요청에는 업종 코드에 대한 정보를 포함시키며, 이 정보를 바탕으로 GPT 4.0이 적절한 응답을 생성하도록 합니다.    **API 호출**   * 준비된 요청을 GPT 4.0 API에 보냅니다. 이 때, API는 업종 코드에 대응하는 업종에 대한 장점, 단점, 전망을 분석하고 응답으로 제공합니다.    **응답 처리 및 사용자에게 정보 제공**   * GPT 4.0 API로부터 받은 응답을 처리하여 사용자에게 전달합니다. 응답에는 해당 업종의 장점, 단점, 그리고 미래 전망에 대한 상세한 설명이 포함됩니다.   텍스트, 스크린샷, 폰트이(가) 표시된 사진  자동 생성된 설명  텍스트, 스크린샷, 폰트이(가) 표시된 사진  자동 생성된 설명   1. 중소벤처 지역 특화 산업 데이터 활용   중소벤처 지역 특화 데이터를 활용하여 사용자들에게 최적의 공장 부지를 추천하는 시스템을 추가하였습니다. 이를 통해 각 지역의 주력 산업 정보를 사용자들에게 제공함으로써, 적합한 산업에 맞는 최적의 공장 부지를 선택할 수 있는 정보를 제공합니다. 아래는 중소벤처 지역 특화 산업 데이터의 활용 방안에 대한 상세 설명입니다. 전국적인 중소벤처 지역 특화 데이터 제공 먼저, 사용자들에게 전국적으로 중소벤처 지역 특화 데이터가 있다는 사실을 알려줍니다. 이를 통해 각 지역이 어떤 산업에 주력하고 있는지를 명확히 인식할 수 있으며, 사용자는 자신의 업종에 맞는 최적의 지역을 쉽게 파악할 수 있습니다. 아래는 전국 각 지역의 주력 산업을 정리한 데이터입니다.   |  |  |  | | --- | --- | --- | | 연번 | 지역 | 주력산업 | | 1 | 부산 | 초정밀 소재부품, 저온고압 에너지저장 공급시스템, 실버케어테크 | | 2 | 대구 | 전기·자율 모빌리티부품, 기계요소 소재부품, 디지털의료기기 | | 3 | 광주 | 스마트홈 부품, 모빌리티 의장·전장부품, 생체의료 소재·부품 | | 4 | 대전 | 나노반도체, 물류·국방 서비스 로봇, 정밀의료 바이오헬스 | | 5 | 울산 | 전기 자동차부품, 가스연료 선박기자재, 기능성 화학소재 | | 6 | 강원 | 천연물 바이오소재, 세라믹원료소재, 디지털 헬스케어 | | 7 | 충북 | 첨단반도체, 융합바이오, 친환경 모빌리티부품 | | 8 | 충남 | 디스플레이 부품·장비, 탄소저감 자동차부품, 고기능성 그린바이오 | | 9 | 전북 | 농생명 바이오, 특수목적용 지능형 기계부품, 탄소융복합소재 | | 10 | 전남 | 환경에너지 소재부품, 친환경에너지 설비기자재, 자연유래 헬스케어 | | 11 | 경북 | 첨단 디지털부품, 신소재부품가공, 라이프케어소재 | | 12 | 경남 | 첨단 정밀기계, 첨단 항공부품, 항노화메디컬 | | 13 | 제주 | 지능형 관광서비스, 청정바이오, 그린에너지 솔루션 | | 14 | 세종 | 지능형 모빌리티 부품, 기능성 바이오 소재 |  지역 선택 시 상세 정보 제공 사용자가 특정 지역을 선택했을 때, 해당 지역의 주력 산업에 대한 상세 정보를 제공합니다. 이를 통해 사용자는 선택한 지역이 어떤 산업에 주력하고 있으며, 해당 지역에서 어떤 사업을 추진할 수 있는지에 대해 구체적으로 알 수 있습니다. 추천 시스템 활용 방안 최종적으로 사용자에게 최적의 공장 부지를 추천하기 위해 다음과 같은 절차를 거칩니다. 이를 통해 사용자는 자신이 원하는 지역과 조건에 맞는 공장 부지를 쉽게 찾을 수 있으며, 선택한 업종에 대한 상세 정보도 제공받을 수 있습니다. 지역 선택 먼저, 사용자는 아래의 17개 지역 중에서 원하는 지역을 선택합니다.   * 강원도 * 경기도 * 경상남도 * 경상북도 * 광주광역시 * 대구광역시   대전광역시   * 부산광역시 * 서울특별시 * 세종특별자치시 * 울산광역시 * 인천광역시 * 전라남도 * 전라북도 * 제주특별자치도 * 충청남도 * 충청북도  중소벤처 지역 특화 데이터 제공 선택한 지역에 대한 중소벤처 지역 특화 데이터를 사용자에게 제공합니다. 이를 통해 사용자는 해당 지역의 주력 산업을 확인하고, 자신이 원하는 업종과의 연관성을 파악할 수 있습니다. 사용자 정보 수집 및 지역 추천 사용자로부터 다음 8개의 질문에 대한 답변을 받습니다.   1. **어떤 공장을 보유하고 싶으신가요? (임대: 1 / 자가: 2)** 2. **공장 등록구분을 선택해주세요. (1: 등록변경 / 2: 부분등록 / 3: 신규등록 / 4: 완료신고)** 3. **공장 규모를 선택해주세요. (1: 소기업 / 2: 중기업 / 3: 대기업)** 4. **남자종업원 수를 입력해주세요.** 5. **여자종업원 수를 입력해주세요.** 6. **외국인 남자종업원 수를 입력해주세요.** 7. **외국인 여자종업원 수를 입력해주세요.** 8. **대표업종 코드를 입력해주세요. (예: 18111)**   사용자로부터 수집한 정보를 바탕으로 최적의 지역을 추천합니다. 대표업종 설명 제공 사용자가 입력한 대표업종 코드에 대해 GPT-4.0 API를 사용하여 해당 산업의 장점, 단점, 전망을 상세히 설명해줍니다. 이를 통해 사용자는 자신이 선택한 업종에 대한 깊은 이해를 얻을 수 있으며, 이를 바탕으로 더 나은 의사 결정을 할 수 있습니다. 매물 추천 추천받은 지역과 매물이 존재하는 공장 데이터를 매칭시켜 사용자가 찾고자 하는 공장 부지 매물을 추천합니다. 이를 통해 사용자는 추천된 지역 내에서 자신이 원하는 조건에 맞는 공장 부지를 쉽게 찾을 수 있습니다. **전체 절차 요약**  1. **지역 선택**: 사용자는 17개 지역 중 원하는 지역을 선택합니다. 2. **중소벤처 지역 특화 데이터 제공**: 선택한 지역의 주력 산업 정보를 제공합니다. 3. **사용자 정보 수집 및 지역 추천**: 8개의 질문을 통해 사용자의 요구 사항을 수집하고, 이를 바탕으로 최적의 지역을 추천합니다. 4. **대표업종 설명 제공**: 사용자가 입력한 업종 코드에 대해 GPT-4.0 API를 통해 해당 업종의 장점, 단점, 전망을 설명합니다. 5. **매물 추천**: 추천받은 지역과 매물을 매칭시켜 사용자가 원하는 공장 부지를 추천합니다.   최종적으로, 사용자는 공장 매물 추천과 선택한 업종에 대한 장점, 단점, 전망에 대한 정보를 얻음으로써, 공장 부지 추천에 대해 종합적이고 유익한 정보를 얻을 수 있습니다. 이를 통해 사용자들은 더 나은 의사 결정을 할 수 있으며, 최적의 공장 부지를 찾는 데 있어 많은 도움을 받을 수 있습니다. 결론  1. **산업 집적지의 중요성**: 산업단지와 같은 공장 집적지는 기업들이 상호 협력하고 기술을 공유하며 인재를 유치하는 데 중요한 역할을 합니다. 한국에서 성공적으로 발생한 산업 집적지의 사례로는 홍릉 바이오 클러스터, 창원 국가산업단지, 여의도 증권가 등이 있습니다. 이러한 집적지는 시너지 효과를 극대화하여 기업의 경쟁력을 높이고 생산성을 향상시킬 수 있는 최적의 환경을 제공합니다. 2. **부지 선택의 복잡성 해결: 공장 부지 선택은 다양한 요소를 고려해야 하는 복잡한 과정입니다. 본 프로젝트는 이러한 복잡성을 줄이고 기업들이 최적의 부지를 선택할 수 있도록 돕기 위해 빅데이터 분석을 활용한 공장 부지 추천 모델을 개발하였습니다. 이를 통해 기업들은 보다 정확한 의사 결정을 내릴 수 있으며, 이는 기업의 성공과 직결됩니다.** 3. **모델 개발 및 성능 분석: LightGBM 모델이 안정적이고 높은 정확도를 나타내는 것으로 밝혀졌습니다. 최적의 하이퍼파라미터를 찾기 위해 Optuna를 사용했으며, Cross-Validation을 통해 모델의 안정성과 정확도를 더욱 향상시켰습니다. 각 지역별로 최적화된 모델을 적용한 결과, 서울특별시에서는 0.7의 정확도를 기록했습니다. 공장 부지 추천 모델은 다양한 요인을 고려해야 하므로, 정답이 존재하지 않습니다. 사용자 맞춤형 추천을 위해 대표업종, 면적, 종업원 수 등을 고려했습니다.** 4. **전국 부지 추천 모델: 서울특별시를 포함한 전국 17개 지역에 대해 공장 부지 추천 모델을 생성하였습니다. 각 지역별 데이터를 분석하고 모델을 최적화한 결과, 일부 지역에서 상대적으로 낮은 정확도가 나타났습니다. 이는 해당 지역들의 산업 활동이 균일하지 않거나 데이터의 양이 부족했기 때문입니다. 그러나 이를 통해 지역별 산업적 특성을 반영한 맞춤형 부지 추천이 가능해졌습니다.** 5. **공장 부지 매물 데이터 수집: 공장 통 닷컴과 부동산 119에서 공장 매물 데이터를 수집하여 모델에 활용하였습니다. 수집된 매물 데이터는 공장 부지 추천 모델에 입력되어 사용자에게 최적의 공장 부지를 추천하는 데 사용됩니다.** 6. **업종 코드 기반 GPT 4.0 API 활용: 사용자가 특정 업종 코드를 입력하면 해당 업종의 장점, 단점, 전망에 대한 정보를 제공하는 기능을 추가하였습니다. 이를 통해 사용자는 자신이 선택한 업종에 대한 깊은 이해를 얻을 수 있습니다.** 7. **중소벤처 지역 특화 산업 데이터 활용: 중소벤처 지역 특화 데이터를 활용하여 각 지역의 주력 산업 정보를 사용자에게 제공함으로써, 적합한 산업에 맞는 최적의 공장 부지를 선택할 수 있도록 지원합니다.** 8. **추천 시스템: 최종적으로 사용자에게 최적의 공장 부지를 추천하기 위해 지역 선택, 사용자 정보 수집, 중소벤처 지역 특화 데이터 제공, 대표업종 설명 제공, 매물 추천 등의 절차를 통해 사용자에게 종합적이고 유익한 정보를 제공합니다.**   결론적으로, 본 프로젝트에서 개발한 공장 부지 추천 모델은 다양한 데이터를 분석하여 기업들이 최적의 부지를 선택할 수 있도록 돕습니다. 이를 통해 기업들은 경쟁력을 강화하고, 생산성을 높이며, 장기적인 성공을 이룰 수 있을 것입니다. |

**활용데이터**

|  |
| --- |
| [제공 데이터]  1. **공장등록 필지 정보 조회 서비스(API)**    * **데이터 설명**: 공장 등록 필지 정보 조회 서비스를 통해 전국 공장등록 현황 데이터의 빈 값이 존재할 경우 추가적인 데이터를 얻어 보완하였습니다.    * **활용 이유**: 공장 등록 데이터의 완전성을 확보하고, 분석의 정확도를 높이기 위해 사용되었습니다. 2. **공장 등록 생산 정보 조회 서비스(API)**    * **데이터 설명**: 공장 등록 필지 정보 조회 서비스를 통해 전국 공장등록 현황 데이터의 빈 값이 존재할 경우 추가적인 데이터를 얻어 보완하였습니다.    * **활용 이유**: 공장 등록 데이터의 완전성을 확보하고, 분석의 정확도를 높이기 위해 사용되었습니다.  [외부 데이터]  1. **전국 공장등록현황 데이터 (https://www.data.go.kr/data/3041646/fileData.do)**    * **설명**: 한국산업단지공단에서 제공하는 전국 공장등록현황 데이터를 사용하였습니다. 이 데이터는 공장의 위치, 대표업종, 직원 수, 생산품, 시설 면적 등의 다양한 정보를 포함합니다.    * **활용 이유**: 공장 부지 추천 모델을 개발하기 위해 필수적인 정보로, 공장들의 입지와 관련된 다양한 변수를 포함하고 있어 최적의 부지를 추천하는 데 중요한 역할을 합니다. 2. **중소벤처 지역 특화 산업 데이터 (https://www.data.go.kr/data/15069529/fileData.do)**    * **설명**: 중소벤처 지역 특화 데이터를 활용하여 각 지역의 주력 산업 정보를 분석하였습니다. 이 데이터는 각 지역별로 주력 산업군을 포함하고 있습니다.    * **활용 이유**: 특정 지역의 산업 특성을 고려하여 기업들이 최적의 공장 부지를 선택할 수 있도록 지원하기 위해 사용되었습니다. 이는 지역별 특화 산업에 맞춘 공장 부지 추천에 큰 도움이 됩니다. 3. **한국표준산업분류표**    * **설명**: 한국표준산업분류표를 참조하여 업종 코드를 분석하고, 각 업종에 대한 상세 정보를 제공하였습니다.    * **활용 이유**: 업종별 특성과 필요 요건을 정확히 이해하고, 기업들이 자신에게 맞는 부지를 선택할 수 있도록 지원하기 위해 사용되었습니다. 이를 통해 업종별로 최적의 부지 추천이 가능합니다.  데이터 활용 이유  1. **데이터 완전성 및 정확도 확보**    * 다양한 출처의 데이터를 통합하고 빈 값을 보완하여, 공장 부지 추천 모델의 정확도를 높였습니다. 2. **지역별 특화 산업 분석**    * 중소벤처 지역 특화 데이터를 활용하여 각 지역의 산업 특성을 파악하고, 이를 기반으로 기업들이 최적의 부지를 선택할 수 있도록 지원합니다. 3. **업종별 맞춤 추천**    * 한국표준산업분류표를 활용하여 업종별 특성과 필요 요건을 정확히 분석하고, 기업들이 자신에게 맞는 부지를 선택할 수 있도록 돕습니다. 4. **종합적 데이터 분석**    * 공장 등록 현황, 필지 정보, 생산 정보 등을 종합적으로 분석하여 기업들이 최적의 부지를 선택할 수 있도록 지원합니다. |

**사업화방안 및 기대효과**

|  |
| --- |
| 1. 사업화 방안  1. **플랫폼 개발 및 운영**    * **플랫폼 구축**: 공장 부지 추천 서비스를 제공하는 웹/모바일 플랫폼을 개발합니다. 사용자 친화적인 인터페이스를 통해 사용자들이 쉽게 접근하고 사용할 수 있도록 합니다.    * **데이터 통합**: 공장등록 필지 정보 조회 서비스, 공장 등록 생산 정보 조회 서비스, 전국 공장등록현황 데이터, 중소벤처 지역 특화 산업 데이터, 한국표준산업분류표 등 다양한 데이터 소스를 통합하여 플랫폼에 반영합니다.    * **실시간 업데이트**: 공장 매물 데이터와 관련 정보를 실시간으로 업데이트하여 최신 정보를 제공합니다.    * **사용자 맞춤 추천**: 사용자 입력 데이터를 기반으로 맞춤형 공장 부지 추천을 제공하는 알고리즘을 개발합니다. 2. **마케팅 및 홍보**    * **디지털 마케팅**: 검색 엔진 최적화(SEO), 소셜 미디어 광고, 콘텐츠 마케팅 등을 통해 플랫폼을 홍보하고 사용자 유입을 촉진합니다.    * **산업 박람회 및 세미나**: 산업 관련 박람회, 세미나, 컨퍼런스 등에 참여하여 플랫폼을 홍보하고 잠재 고객을 확보합니다.    * **파트너십 구축**: 부동산 중개업체, 산업 단지 관리 기관, 중소벤처기업부 등과의 파트너십을 통해 플랫폼의 신뢰성과 접근성을 높입니다. 3. **고객 지원 및 서비스 개선**    * **고객 지원 센터**: 사용자의 문의와 피드백을 처리할 수 있는 고객 지원 센터를 운영합니다.    * **데이터 분석 및 피드백 반영**: 사용자의 피드백을 분석하여 플랫폼을 지속적으로 개선합니다.    * **교육 및 컨설팅 서비스**: 공장 부지 선택과 관련된 교육 및 컨설팅 서비스를 제공하여 사용자가 최적의 결정을 내릴 수 있도록 지원합니다.  2. 기대효과  1. **경제적 파급 효과**    * **산업 경쟁력 강화**: 최적의 공장 부지 추천을 통해 기업들이 효율적인 운영을 할 수 있게 도와줍니다. 이는 생산성 향상과 비용 절감으로 이어져 기업의 경쟁력을 강화합니다.    * **중소기업 지원**: 중소기업들이 적합한 부지를 선택하여 성공적으로 사업을 운영할 수 있도록 도와줍니다. 이는 중소기업의 성장과 고용 창출로 이어져 지역 경제에 긍정적인 영향을 미칩니다.    * **부동산 시장 활성화**: 공장 부지 매매 및 임대가 활발해지면서 부동산 시장의 활성화에 기여할 수 있습니다. 2. **사회적 파급 효과**    * **지역 균형 발전**: 중소벤처 지역 특화 데이터를 활용하여 각 지역의 산업 특성을 고려함으로써 지역 균형 발전을 촉진할 수 있습니다. 이는 지역 간 경제 격차를 줄이는 데 기여합니다.    * **환경적 지속 가능성**: 최적의 부지 선택을 통해 교통, 물류, 에너지 사용 등을 최적화하여 환경적인 영향을 최소화할 수 있습니다. 이는 지속 가능한 산업 발전에 기여합니다.    * **일자리 창출**: 공장 설립과 운영을 통한 일자리 창출로 지역 사회의 고용 안정성과 생활 수준 향상에 기여할 수 있습니다. 3. **데이터 기반 의사 결정 문화 확산**    * **근거 기반 의사 결정**: 데이터에 기반한 의사 결정 문화의 확산을 촉진하여 기업들이 보다 신뢰성 있는 정보를 바탕으로 전략을 수립하고 실행할 수 있도록 지원합니다.    * **시간 및 비용 절감**: 기업들이 지역을 선정하는 데 소요되는 시간과 비용을 절감할 수 있습니다. 이는 기업들이 빠르고 효율적으로 공장 부지를 선정하는 데 큰 도움을 줄 것입니다.    * **복잡한 절차 간소화**: 데이터 기반의 접근을 통해 복잡한 절차를 간소화하고, 기업들이 지역 선정 과정에서 겪는 어려움을 줄일 수 있습니다. |