머신러닝 결과보고서

채수병

목차

1. 프로젝트 기획

- 프로젝트 기획의도, 개발 목표 및 계획 수립(WBS)

- 개발 계획서

1. 데이터 수집

- 필요한 데이터 정의(데이터 정의서, 획득 계획서)

- 사용할 수 있는 데이터 유형과 소스

- 데이터 수집 및 정합성 점검

1. 데이터 분석

* 데이터 준비(데이터 전처리)
* 데이터 분석
* 모델 평가 및 검증

1. 시스템 구현

* 설계 및 구현

1. 평가

* 프로젝트 평가
* 모델 발전 계획 수립 및 향후 계획

1. 프로젝트 기획

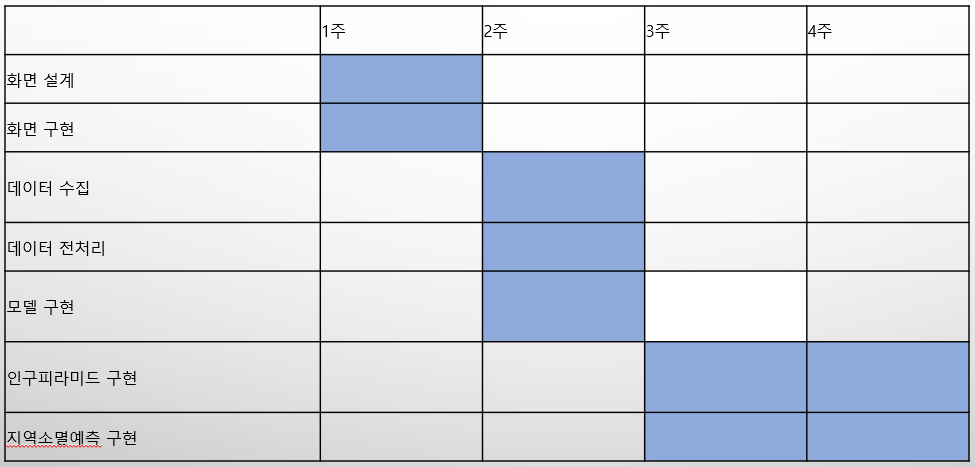
- 프로젝트 기획의도

저출산 고령화 문제가 심각해지고 있다. 하지만 소수만 이를 체감할 뿐 대다수는 인지하지 않고 살아간다. 이를 해결하기 위해 도표로 보여주어 시각적으로 사람들에게 각인 시키고자 한다.

- 개발 목표 및 계획 수립(WBS)

지역별, 연령별 인구수 예측 머신러닝 모델을 만든다. 계획은 아래와 같다.

- 개발 계획서



2. 데이터 수집

- 필요한 데이터 정의(데이터 정의서, 획득 계획서)

출생률 예측 – 가임 여성 수, 평균 임금, 사교육비, 집값, 혼인 건수 등

사망률 예측 – 노인 인구, 암 환자 수

- 사용할 수 있는 데이터 유형과 소스

KOSIS csv 파일, 한국 부동산원 부동산 통계 정보 시스템 csv 파일

- 데이터 수집 및 정합성 점검

KOSIS, 한국 부동산원 부동산 통계 정보 시스템 등 신뢰할 수 있는 사이트에서 데이터를 수집했고 데이터가 일관적이고 중복이 없어 정합성에 문제가 없었다.

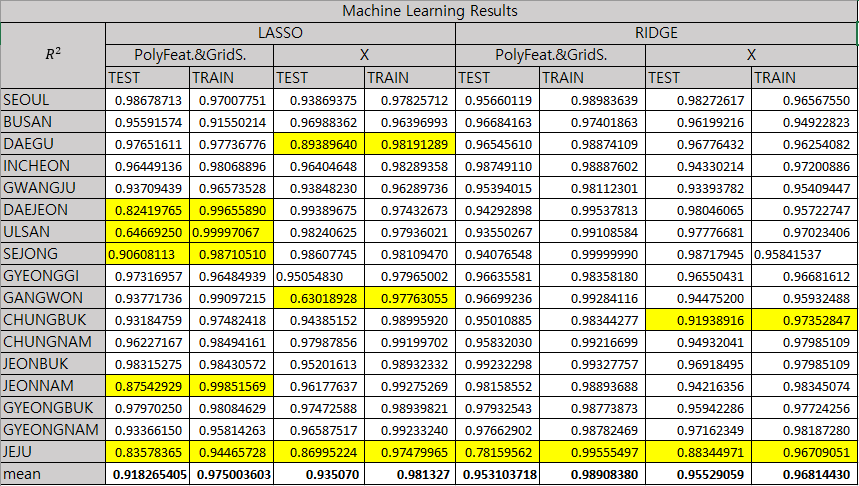
3. 데이터 분석

- 데이터 준비(데이터 전처리)

다양한 범위에 분포되어 있던 연도를 2007~2023년으로 통일하여 추렸다. 지역은 SEOUL, BUSAN, DAEGU, INCHEON, GWANGJU, DAEJEON, ULSAN, SEJONG, GYEONGGI, GANGWON, CHUNGBUK, CHUNGNAM, JEONBUK, JEONNAM, GYEONGBUK, GYEONGNAM, JEJU로 통일하였다.

- 데이터 분석

지역 소멸 예측과 인구 피라미드 구현을 위해 2025~2075년 사망률, 출생률 데이터가 필요했다. 이를 얻기 위해 사망률과 출생률을 예측해야 했다. 사망률의 input 값으로 노인 인구, 암환자 수를 사용했고 출생률의 input 값으로 가임 여성 수, 평균 임금, 사교육비, 집값, 혼인 건수 등을 사용했다. 예측 값 (사망률, 출생률)을 구하기 위해 LASSO회귀와 RIDGE회귀 두가지 방법이 고려되었다.



정확도 분석 결과 RIDGE회귀가 LASSO회귀보다 나은 결과를 보였고 RIDGE회귀를 사용해 사망률과 출생률을 예측했다. 사망률은 전체 인구가 줄어들어도 증가하는 추세를 보였고 출생률은 감소하는 추세를 보였다. 이로 인해 지역 소멸 예측에서 모든 지역의 소멸도가 0에 가까워져 소멸되는 모습을 보였고 인구 피라미드는 시간이 지날수록 역 삼각형 형태를 띄었다.

- 모델 평가 및 검증

오버 피팅이 발생했지만 오차가 적어 큰 문제가 없었다.

4. 시스템 구현

- 설계 및 구현

Numpy와 Pandas를 활용하여 머신 러닝 모델을 구축하였으며 linear regression과 RIDGE회귀를 사용하여 한국 인구를 예측하도록 하였다.

5. 평가

- 프로젝트 평가

- 모델 발전 계획 수립 및 향후 계획