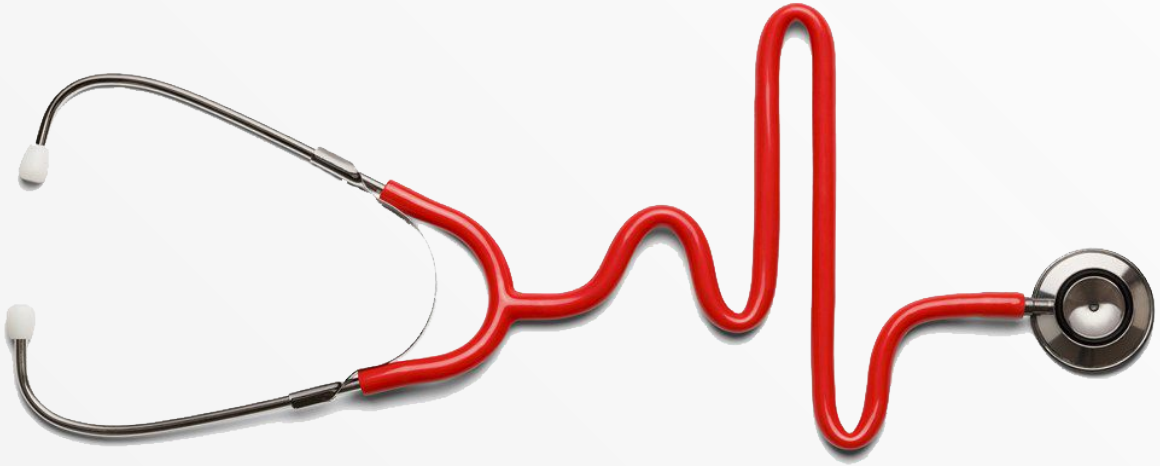


공공 의료 데이터 기반 지역별 회피가능 사망자수 예측 모델링

4망자예측조

팀장:신재호 팀원:박현주,이영서,정소현

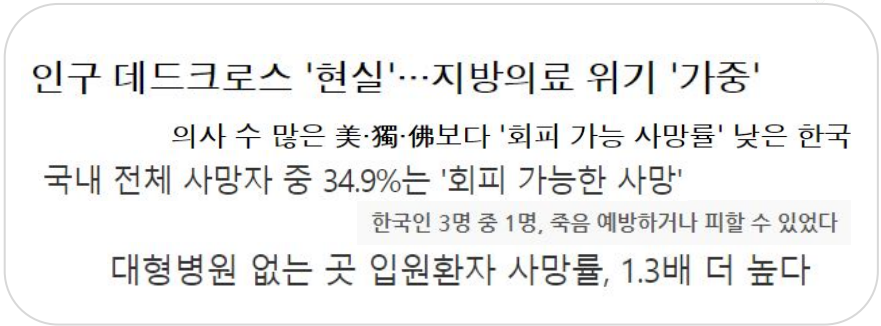
INDEX

I. 프로젝트 배경

프로젝트 주제 선정 이유

프로젝트 주제

공공 의료 데이터 기반
‘지역별 회피가능 사망자수’ 예측 모델링



〈‘지방의료위기’와 ‘회피가능사망’ 관련 보도자료〉

주제 선정 이유

지난해 사상 처음 인구 데드크로스가 현실화된 이후로
수도권과 지방 간 의료 격차가 점점 심화되고 있다.
우리는 이 위기를 가장 잘 보여주는 지표 중 하나인 ‘회피가능사망’에 주목하게 되었다.
의료비용, 의료시설, 의료환경 등 측정 가능한 다양한 요인들이
어떻게 회피가능사망에 영향을 미치는지 분석하고,
지역별 회피가능사망수를 예측하여 지방 의료 위기를 현실적으로 바라보고자 한다.

회피가능사망(Avoidable Mortality)이란?
의료적 지식과 기술을 고려했을 때 조기 검진, 시의적절한 치료 등과 같은
양질의 보건 의료서비스를 통해 피할 수 있거나 예방이 가능한 사망으로 정의됨

I. 프로젝트 배경

프로젝트 개요

프로젝트 개요

- 프로젝트 수행기간: 2021/9/27 ~ 2021/10/6 (8일간)
- 프로젝트 내용: 공공 의료데이터를 활용하여 지역별 회피가능사망수 예측
- 프로젝트 수행방향: 회피가능사망수에 영향을 미치는
  Feature를 수집하고 분석하여 적절한 예측 모델 개발
- 프로젝트 수행 도구
  - 수행 환경: Colab, Zoom, Google Drive
  - 분석기술 패키지: Pandas, Numpy, Sklearn,
  - 시각화 기술 패키지: Matplotlib, Seaborn, Folium
- 기대효과: 지역별 회피가능사망수 예측 모델을 통해
  지방의 보건의료 현실을 지적하여 더 나은 의료 복지 환경을 기대

Ⅱ. 팀 구성 및 역할

팀 구성 및 역할

“4망자예측조”

프로젝트 주제인 ‘사망자예측’과
‘4조’의 의미를 담아 작명

멘토
전준걸

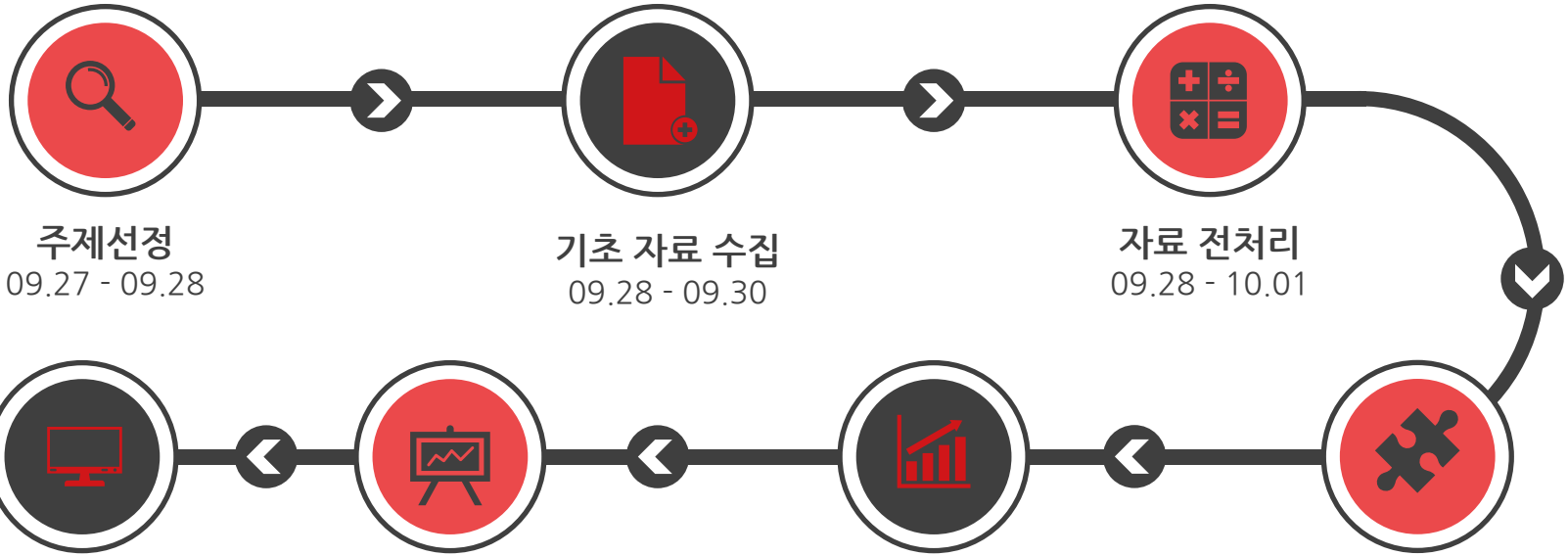
상시 피드백

팀장	팀원	팀원	팀원
신재호	박현주	이영서	정소현

기획 → 데이터 수집 → 데이터 전처리 → 시각화 → 데이터 분석&검증

프로젝트 전 과정 공동참여

III. 수행 절차 및 방법


주제선정
09.27 - 09.28
기초 자료 수집
09.28 - 09.30
자료 전처리
09.28 - 10.01
데이터 분석 및 파악
09.29 - 10.04
데이터 시각화
모델 분석 및 구현, 검증
10.04 -10.05
데이터 결과 해석
10. 05
프로젝트 발표 준비
10. 05

III. 수행 절차 및 방법 - 데이터 명세

통계청을 비롯 여러 기관의 CVS 자료를 수집

데이터 이름	출처	요약	제공형태
건강보험 청구자료	공공데이터 포털	요양급여비, 입내원일수를 시도코드 기준으로 추출	csv
행정구역별 인구수	국가통계포털(KOSIS)	시도코드 기준 총인구수 추출	csv
요양기관수 현황	보건의료빅데이터개방시스템	인구 및 면적데이터 활용하여 각 요양기관 수 산출	csv
전국 의료인력 현황	국가통계포털(KOSIS)	의사, 간호사로 구분되는 데이터를 시도코드 기준으로 추출	csv
중증도 보정입원 의료 사망비	건강보험의료지도	2019년 기준으로 데이터 전처리 후, 지역별 인구 데이터와 결합하여 '회피가능 사망수'계산	csv
지역별 면적	국가통계포털(KOSIS)	시도별 면적 정리후, 의료인수, 의료기관 자료와 결합하여 '면적당 의료인수, 의료기관'등을 계산	csv
전국 의료기관 병상수	국가통계포털(KOSIS)	필요없는 컬럼을 제외한 전처리, 시도코드 추가	csv
지역종별 의료인력	국가통계포털(KOSIS)	상급종합병원, 종합병원, 병원, 의원의 의료인력 데이터	csv

III. 수행 절차 및 방법 - 데이터 전처리 및 탐색

전처리 과정도

1. 결측치 제거
2. 2019년도 데이터만 선택하여 기간 통일
3. 데이터 별로 사용한 시도코드 버전이 다른 부분 통일
4. 인구대비 / 면적대비 등 기존의 자료를 가공하여 새 요소 생성
5. 시도 코드를 기준으로 데이터 통합

III. 수행 절차 및 방법 - 전처리

시도코드 변경 / 2019년 자료 추출 및 변형

A. 시도코드 전처리

```
# 국민건강보험 raw_data 공단 시도코드 (구버전 사용)
raw_data['시도코드'].unique()
```

```
array([11, 26, 27, 28, 29, 30, 31, 36, 41, 42, 43, 44, 45, 46, 47, 48, 49])
```

비교

```
# 중증도보정입원의료 사망률 data 시도코드
die_raw['\t시도코드'].unique()
```

```
array([11, 21, 22, 23, 24, 25, 26, 29, 31, 32, 33, 34, 35, 36, 37, 38, 39])
```

```
# 시도코드 전처리후
healthcare_total['시도코드'].unique()
```

전처리

```
array([11, 21, 22, 23, 24, 25, 26, 29, 31, 32, 33, 34, 35, 36, 37, 38, 39])
```

B. 2019년 해당 자료 추출

```
# 중증도보정입원의료 사망율 raw data 지표연도
die_raw['지표연도'].unique()
```

```
array([2017, 2019, 2018, 2016, 2015, 2011, 2012, 2013, 2014])
```

```
# 중증도보정입원의료 사망율 raw data 지표연도(samwple)
die_raw.sample(5)
```

	지표연도	\t시도코드	\t시도명	\t구분	\t성별	\t연령대	\t보험료구분	\t이벤트건수	\t대상자건수	\t지표비율	\t표준화율
296	2013	34	\t충청남도	\t권내	\t전체	\t전체	\t전체	3080	2689	115.0	115.0
248	2017	32	\t강원도	\t권내	\t전체	\t전체	\t전체	3802	3305	115.0	115.0
279	2019	33	\t충청북도	\t권내	\t전체	\t전체	\t전체	997	819	122.0	122.0
54	2015	22	\t대구광역시	\t권내	\t전체	\t전체	\t전체	3596	3568	101.0	101.0
308	2017	34	\t충청남도	\t전체	\t전체	\t전체	\t전체	6410	6403	100.0	100.0

```
# 2019년도 데이터만 출력
df1 = df1[df1.지표연도 == 2019]
df1.head()
```

	지표연도	\t시도코드	\t시도명	\t구분	\t성별	\t연령대	\t보험료구분	\t이벤트건수	\t대상자건수	\t지표비율	\t표준화율
1	2019	11	서울특별시	권외	전체	전체	전체	2599	2263	115.0	115.0
2	2019	11	서울특별시	권내	전체	전체	전체	3792	4358	87.0	87.0
3	2019	11	서울특별시	전체	전체	전체	전체	6391	6410	100.0	100.0
36	2019	21	부산광역시	전체	전체	전체	전체	2418	2913	83.0	83.0
37	2019	21	부산광역시	권내	전체	전체	전체	1215	1663	73.0	73.0

III. 수행 절차 및 방법 - 데이터 가공

다양한 데이터를 미리 정한 표준 규격에 맞추기 위한 수치 계산 작업

각 지역별로 데이터 그룹화

지역별로 1차 분류 한 이후에 다른 기준으로 세분화된 데이터들을 시도별로 한 개의 데이터가 되도록 합해준다.

```
df = total.groupby('시도코드')[['심결요양급여비용총액']].sum()
```

지역별 특성으로 인한 차이 제거

각 지역의 인구 수,면적 등의 차이를 수치에서 제거하여, 보다 정확한 비교를 도모한다.

```
dt['면적당 요양기관수'] = dt['요양기관총합'] / dt['면적(2019)'] *10000000
dt['면적당 의사수'] = dt['의사'] / dt['면적(2019)'] *10000000
dt['면적당 간호사수'] = dt['간호사'] / dt['면적(2019)'] *10000000
```

가독성을 위해 새로운 수치 산출

면적당, 인당 수치로 변환하면서 다음과 같은 문제를 해결한다.

1. 지나치게 작은 값으로 변하는 문제
2. 소수점이 길어지는 문제

```
dt['면적당 요양기관수'] = dt['면적당 요양기관수'].round(2)
dt['면적당 의사수'] = dt['면적당 의사수'].round(2)
dt['면적당 간호사수'] = dt['면적당 간호사수'].round(2)
```

Raw Data

	시도코드	심결요양급여비용총액
0	36	19240
1	36	145680
2	36	13240
3	36	861590
4	36	48440
...	...	...
13178153	36	13170
13178154	36	11210
13178155	36	13230
13178156	36	11210
13178157	36	11210

III. 수행 절차 및 방법 - 데이터 가공

다양한 데이터를 통합

사용 코드

```
total = pd.merge(compare, death)
total.columns
```

통합된 데이터, 데이터 내 컬럼값

시도코드	시도명	총 인구	주민등록인구	면적(2019)	인구당 요양급여비	인구당 입내원일수	천명당 의료인수	천명당 의사수
11	서울특별시	66.496942	9639541	9729107	605237001.6	19486.007369	0.434774	11.805914
21	부산광역시	86.370176	3372692	3413841	770073413.2	25404.979491	0.567675	10.846434
22	대구광역시	61.359540	2429940	2438031	883517307.5	20611.235100	0.458401	10.700028

...

..

면적당 요양기관수	면적당 요양기관수(상급)	면적당 요양기관수(상급+종합)	면적당 요양기관수(종합)	면적당 요양기관수(병원)	면적당 요양기관수(의원)	십만명당 회피가능사망수
0.895063	0.000015	2.147919e-08	9.417798e-08	7.269879e-08	3.767119e-07	0.000014
0.706854	0.000003	5.194310e-09	3.765875e-08	3.246444e-08	1.843980e-07	0.000003
0.741994	0.000002	5.659199e-09	1.810944e-08	1.245024e-08	1.233705e-07	0.000002

```
Index(['시도코드', '시도명', '총 인구', '주민등록인구', '면적(2019)', '인구당 요양급여비', '인구당 입내원일수',
       '천명당 의료인수', '천명당 의사수', '천명당 간호사수', '면적당 의료인수', '면적당 의료인수(상급)',
       '면적당 의료인수(상급+종합)', '면적당 의료인수(종합)', '면적당 의료인수(병원)', '면적당 의료인수(의원)',
       '천명당 병상수', '천명당 요양기관수', '천명당 요양기관수(상급)', '천명당 요양기관수(상급+종합)',
       '천명당 요양기관수(종합)', '천명당 요양기관수(병원)', '천명당 요양기관수(의원)', '면적당 요양기관수',
       '면적당 요양기관수(상급)', '면적당 요양기관수(상급+종합)', '면적당 요양기관수(종합)', '면적당 요양기관수(병원)',
       '면적당 요양기관수(의원)', '십만명당 회피가능사망수'],
      dtype='object')
```

Ⅳ.분석 및 시각화-문제 제기

지역인구대비 사망자수 분포

중증도보정의료사망비 데이터를 사용하여 지역별 **'회피가능사망자수 (Avoidable Mortality)'** 비교하였다. **1차분석**에서 표로 확인할 수 있는 사실이 한정적일 수 있고, 지역간 차이를 분류기준으로 삼은 만큼 지도 시각화를 통해 추가적으로 얻을 수 있는 정보가 있다고 판단하여, 지도를 통한 **2차 분석**을 진행하였다.

시도코드	시도명	십만명당 회피가능사망자수	
36	전라남도	max	152.8
32	강원도		120.8
35	전라북도		107.1
39	제주특별자치도		99.1
37	경상북도		97.0
38	경상남도		90.7
34	충청남도		90.5
21	부산광역시		86.4
33	충청북도		84.9
11	서울특별시		66.5
23	인천광역시		65.6
22	대구광역시		61.4
24	광주광역시		60.7
31	경기도		59.5
25	대전광역시		55.6
26	울산광역시	min	52.7
29	세종특별자치시		37.9

사망자수 내림차순

최대값과 최소값을 확인

지도상 분포

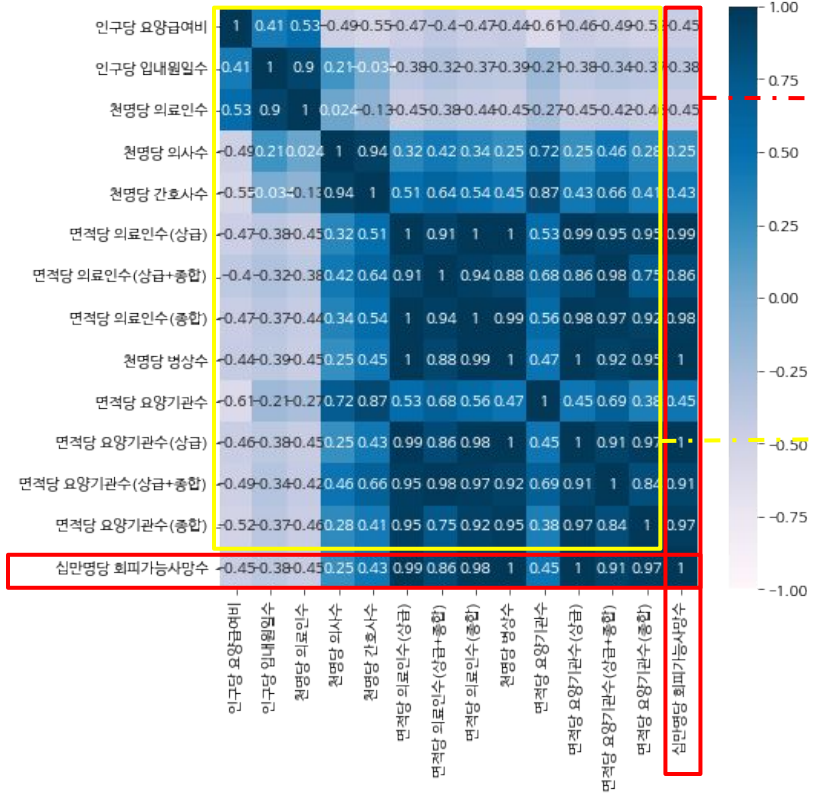
시각적 요소 분석

최대/최소 약 3배 차이
분석이 필요한
유의미한 차이가
있다고 판단

'지역간 접근성'이 연관
수도권,광역시에 낮은
사망자수 분포 확인
>>변수 선정시 고려

Ⅳ. 분석 및 시각화 - 상관관계 확인

다양한 데이터의 상관관계 확인(pearson계수)



1차 분석

예측하고자 하는 결과 값
'회피가능 사망자수'와
원인 변수들의 상관관계 분석

2차 분석

원인 변수들 간의 상관관계 분석

지역별 회피가능 사망자수 예측
선형 회귀 모델 (Linear Regression)

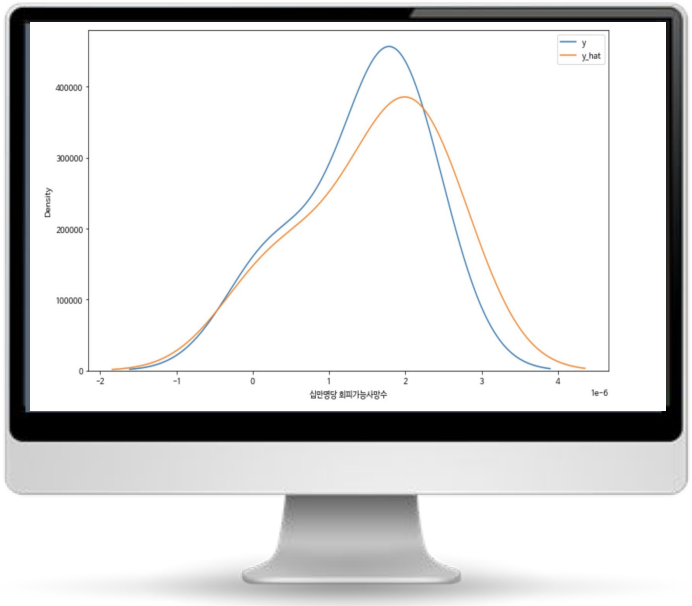
전처리

- **Feature(x)** : 상관관계 분석을 통해 선택 (28개 ➡️ 12개)
- **Target(y)** : 십만명당 회피가능 사망자 수
- **Feature Scaling** : StandardScaler
- Train 과 Test 데이터 **8 : 2**

성능평가 결과

Polymal_r제곱	Linear_r제곱	MAE	MSE	RMSE
0.517978	0.919159	1.503154e-07	4.039039e-14	2.009736e-07

성능평가 결과와 KDE그래프를 그려봤을 때, 실제값과 예측값은 모두 오른쪽으로 편중되어 있으며, 독립변수와 종속변수 간의 선형관계가 성립됨을 확인 가능



seaborn - KDE(커널밀도 그래프)

V. 결론 및 향후 과제

수행 결과와 해석

수행 결과

종속변수인 '회피가능사망수'와 여러 개의 독립변수를 선형 회귀 분석으로 모델링 후
R^2값 0.91을 기록하며 높은 정확도의 예측 모델을 개발함

결과 해석

1) 사망수에 영향을 미치는 것은 특정 변수가 아닌 전체 의료환경이다.

독립변수들의 상관계수에 따르면 어느 특정 변수가 큰 영향을 미치는 것이 아닌,
의료비용, 의료시설, 의료환경 등 다양한 요인들이 사망 수와 상관관계가 있다는 것을 확인했다.
이는 어느 특정 요인이 아닌 의료 환경을 구성하는 구조적, 지리적인 모든 종합적 요소들이 상관관계가 있다는 것을 의미한다.

2) 사망에 이르게 하는 요인은 환자 중에서도 중환자의 의료환경과 상관관계가 있다.

상급종합병원, 종합병원, 병원, 의원으로 분류되는 요양기관 수가 사망수에 영향을 미칠 것이라는 가설을 세웠고,
실제로는 전체 요양기관 수가 아닌 상급종합병원, 종합병원의 수가
전체 요양기관의 수 보다 높은 상관계수를 기록한 것을 확인할 수 있었다.
이는 주로 생사를 다루는 중환자의 비중이 높은 상급의료기관의 환경이 상관관계가 크다는 것을 의미한다.

V. 결론 및 향후 과제

기대효과 및 향후 과제

기대효과

1. 의료 취약 지역 예측 및 의료 인프라 개선
+
2. 국토 균형 발전을 위한 지표 제시
+
3. 실제와 예측 비교를 통해 지역별 로컬 변수 분석

향후 과제

본 프로젝트에서는 면적 크기만을 고려했으나,
향후에는 접근성(교통편, 병원과의 이동거리 등) 데이터를 세밀하게 분석하여
방문 빈도가 높은 특정 질병과 회피가능사망수의 상관관계를 바탕으로
접근성과 치료 여건을 개선할 수 있는 예측 모델을 통해 여러가지 지표를 개선할 수 있을 것이다.

Team Comment

“죽음과의 사투"
죽음에 대한 데이터를 분석하면서 느낀
‘4망자 예측'팀 팀원들의 후기입니다.

신재호
Programmer

“이 프로젝트는 저에게
이론적으로 알고 있다고
생각하는 것과 실제로
수행하는 것의 차이를 체감할
수 있는 시간이었습니다.
또한 팀원들의 도움을 너무도
많이 받아 미안하고 감사하며
의지가 넘치는 모습들이 많은
자극이 되었습니다."

박현주
Programmer

“이 프로젝트는 나에게 처음엔
혼란만을 안겨주었지만,
각자의 강점을 가진 팀원들
덕분에 무사히 프로젝트를
끝낼 수 있었다.
특히 내가 부족한 점이
무엇인지를 파악할 수 있는
자기성찰의 시간이었던 것
같다."

이영서
Programmer

“이 프로젝트는 나에게 좋은
동기부여가 됐다. 팀원들과
함께 협력하고 , 무엇보다
팀원들의 열정에 많은 자극을
받았다. 또한 그간 수업시간에
배웠던 내용을 활용하여 ,
프로젝트를 진행하며 어느새
데이터 수집, 전처리, 시각화
및 분석 등에
익숙해진 모습을 발견할 수
있는 뜻 깊은 시간이었다.”

정소현
Programmer

“이 프로젝트는 나에게 그
어떤 공부보다도 기억에 남는
공부였다. 그간 많은 양을
공부했고 실습해왔는데,
시간이 지나면 코드들이
잊혀졌다. 그런데 내 필요에
의해 코드들을 찾아쓰고, 나의
상황에 맞게 코드들을
적용하다보니 이번에 사용한
코드들만큼은 절대
잊어버리지 않는다.”

Thank you

4망자예측조