**빅데이터 분석 결과 보고서**

**생활 환경에 따른 노인 기대 수명 및 질병 정보 제공 서비스**

**DLT (Desired of Lifetime)**

유정민

목차

[목차 2](#_Toc154068413)

[**1. 분석 배경 및 기획 의도** 3](#_Toc154068414)

[**1.1 분석 배경** 3](#_Toc154068415)

[**1.2 기획의도** 3](#_Toc154068416)

[**2. 분석 목표** 3](#_Toc154068417)

[**3. 분석 데이터** 4](#_Toc154068418)

[**3.1** **데이터 수집** 4](#_Toc154068419)

[**3.2** **데이터 정합성 검정** 8](#_Toc154068420)

[**3.3** **데이터 전처리** 9](#_Toc154068421)

[**4. 분석 과정 및 결과** 9](#_Toc154068422)

[**4.1 기대 수명 예측 모델 훈련 및 시각화** 10](#_Toc154068423)

[**5. 활용 방안** 15](#_Toc154068424)

[**6. 참고문헌** 15](#_Toc154068425)

# **1. 분석 배경 및 기획 의도**

## **1.1 분석 배경**

현재 우리 사회는 고령화 문제를 지속적으로 겪고 있다. 지난해 기준 전국 65세 이상 고령인구는 901만 8000명으로 전체 인구의 17.5%의 비율로 1970년대부터 지속적으로 고령 인구가 증가하고 앞으로도 계속해서 증가할 것이다. 고령 인구 증가의 원인 중 하나로 의료 수준 증가로 인한 기대 수명의 지속적인 증가가 있다.

기대 수명은 사회, 경제 및 관련 의사 결정에 중대한 지표로 자리 잡고 있다. 인구의 건강 수준을 가늠하는 척도로, 기대 수명은 공공 보건 정책, 의료 자원 배분, 보험 및 연금 계획과 같은 중요한 영역에 영향을 미친다. 또한, 기대 수명은 개인의 삶의 질과 밀집하게 연결되어 있으며, 인구 구조와 경제 발전에 중요한 정보를 제공한다.

## **1.2 기획의도**

고령 인구들을 위한 기대 수명 예측 모델을 개발하여 기대 수명 예측 모델을 제공 함으로써 사용자가 기대 수명에 대한 정보를 얻어, 앞으로의 노후 계획, 건강 관리 계획을 세울 수 있다. 이 모델 훈련은 서비스 사용자의 건강 상태를 기반으로, 개인별로 기대 수명을 예측할 수 있는 모델을 훈련하여, 기대 수명에 대한 예측 데이터를 제공하도록 한다.

# **2. 분석 목표**

이 연구의 분석 목표는 고령화가 진행 중인 현대 사회에서 기대 수명을 정확하게 예측하는 것이다. 이를 통해 우리는 공공 보건과 사회 보장 체계에 중대한 영향을 끼치는 다양한 요인들을 이해하며 개인의 특성을 반영한 맞춤형 예측 모델을 개발하는 것이다. 전국의 고유한 환경적, 경제적, 사회적 조건을 모델에 사용함으로써 지역 공동체와 관련 기관들이 보다 정밀한 데이터를 기반으로 의사결정을 내릴 수 있도록 지원한다.

# **3. 분석 데이터**

개인별 건강 상태에 따른 수명 데이터를 활용하면 기대 수명에 대한 예측 모델이 정확하지만, 개인의 프라이버시 문제로 데이터를 일반적인 상황에서는 구할 수 없다. 이로 인해, 전국의 연령대, 성별로 나눈 연평균 데이터를 사용하기로 했다.

## **데이터 수집**

모든 데이터는 연령대, 성별을 기준으로 수집한 특성이다.

- 기대 여명

인구 데이터로는 기대 수명 데이터, 사망 인구 데이터가 있다.

해당 데이터는 타겟 특성으로, 모델이 정확히 훈련되었는지 판단하는 기준점이 된다.

- 지역별 생활 습관

지역별 생활 습관으로는 흡연 비율, 스트레스 인지율, 운동 실천 비율, 음주 비율을 사용하였다.

스트레스 인지율 데이터와 운동 실천 비율 데이터는 데이터의 겉측값이 너무 많고, 기대 수명 간의 상관관계가 너무 약하기에 제외하고, 흡연 비율, 음주 비율을 특성으로 모델에 사용하기로 했다.

흡연 비율 데이터와 스트레스 인지율, 운동 실천 비율, 음주 비율 데이터는 지역 총 인구의 일부인 모집단에서 조사한 데이터다.

질병 관리청\_지역사회건강조사(흡연 비율)(2013-2022)

국민 건강 보험\_지역별 의료이용 통계연보(음주 비율)(2013-2022)

-지역별 질병 데이터

질병 데이터는 만성 질환인 고혈압과 당뇨병을 사용하였다.

많은 질병 데이터 중 만성 질환을 선택한 이유는 장기적인 영향력이 강하기 때문에 개인의 건강에 영향을 미치기에 다른 질병에 비해 기대 수명과의 연관성이 높다고 판단하여 사용하였다.

고혈압 데이터와 당뇨병 데이터는 전체 인구 수에서 의료 이용률에 대한 퍼센트 데이터를 사용하였다.

국민 건강 보험공단\_고혈압의료이용률(2013-2022), 국민 건강 보험공단\_당뇨병의료이용률(2013-2022)

### **3.1.1 사용 데이터 목록 및 출처**

<표 1> 지역별 기대수명, 2013-2022

단위: 연령

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 |
| 서울 | 83.56 | 83.88 | 84.23 | 83.59 | 84.86 | 85.09 | 85.38 | 85.64 | 85.77 |  |
| 전남 | 80.78 | 81.16 | 81.61 | 82.06 | 82.47 | 82.72 | 83.05 | 83.33 | 83.57 |  |

* 통계청(KOSIS)\_2022년 생명표 보도자료

<표 2> 지역별 흡연 비율, 2013-2022

단위: %

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 |
| 서울 | 21.7 | 20.6 | 19.4 | 19.5 | 18.8 | 18.6 | 16.7 | 15.5 | 14.9 | 15.3 |
| 전남 | 20.8 | 20.9 | 19.2 | 20.2 | 18.8 | 18.8 | 18.5 | 16.9 | 16.5 | 17.7 |

* 질병 관리청\_지역사회건강조사(흡연 비율)(2013-2022)

<표 3> 지역별 음주 비율, 2013-2022

단위: %

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 |
| 서울 | 60.7 | 60.3 | 61.6 | 61.5 | 61.0 | 59.7 | 58.7 | 53.0 | 50.9 | 55.3 |
| 전남 | 50.4 | 52.2 | 52.4 | 51.8 | 52.2 | 51.3 | 50.6 | 45.4 | 45.7 | 47.5 |

* 국민 건강 보험\_지역별 의료이용 통계연보(음주 비율)(2013-2022)

<표 4> 지역별 고혈압 유병률, 2013-2022

단위: %

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 |
| 서울 | 14.73 | 14.95 | 15.27 | 15.85 | 16.45 | 17.03 | 17.65 | 18.26 | 19.03 | 21.03 |
| 전남 | 18.88 | 19.36 | 19.83 | 20.48 | 21.04 | 21.78 | 22.89 | 23.87 | 25.13 | 26.91 |

* 국민 건강 보험공단\_고혈압의료이용률(2013-2022)

<표 5> 지역별 당뇨병 유병률, 2013-2022

단위: %

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 |
| 서울 | 5.26 | 5.42 | 5.64 | 5.98 | 6.33 | 6.67 | 7.03 | 7.42 | 7.88 | 8.01 |
| 전남 | 7.09 | 7.4 | 7.71 | 8.08 | 8.53 | 8.96 | 9.54 | 10.08 | 10.72 | 11.13 |

* 국민 건강 보험공단\_당뇨병의료이용률(2013-2022)

## **데이터 정합성 검정**

연령대별, 성별을 기준점으로 데이터를 조사하면 데이터 샘플의 숫자가 부족하여 모델이 과소적합되는 현상이 발생했다. 모델의 샘플 수는 20대부터 80대까지 5세를 기준으로 남성, 여성의 데이터 샘플들을 수집했기 때문에, 최대 24개까지 구할 수 있었으며 8개의 특성을 가지고 있다.

해당 문제를 해결하기 위해 ChatGPT와 엑셀을 사용하여 임의적으로 데이터 샘플의 수를 늘려 모델 훈련을 진행했다. 해당 데이터 샘플의 개수는 200개에 8개의 특성을 가지게 되었다.

## **데이터 전처리**

데이터 전처리는 python을 사용하여 진행했으며, 데이터 값의 평균에서 얼마나 떨어져 있는지를 의미하는 표준 점수를 사용하여 데이터의 스케일을 동일하게 설정했다.

# **4. 분석 과정 및 결과**

이번 프로젝트에서 사용하는 모델은 다중 선형 회귀 모델을 사용했다. 개인별로 기대 수명을 예측하기 위한 모델을 훈련시킨다. 모델의 정확도를 판단하기 위한 객관적인 지표는 결정 계수()을 사용했다.

모델이 훈련 샘플에만 과하게 훈련되어 새롭게 입력한 값들에 대한 예측 값의 오차를 줄이기 위해 훈련 샘플과 테스트 샘플을 분리하여 모델 훈련을 진행했다.

사용 프로그램은 python을 사용해서 데이터 전처리, 모델 훈련 및 예측을 진행했다.  
공공데이터를 사용하여 데이터를 분석하고 시각화하여 분석을 진행하였다. 사용한 특성은 연령대, 성별을 기준으로 흡연 비율, 음주 비율, 비만 비율, 당뇨 비율, 고혈압 비율 있으며 타겟 변수는 기대 수명이다. 모델을 훈련하기에 앞서, 훈련 세트와 테스트 세트로 샘플을 나누고, 전처리를 진행하고 모델 훈련을 진행했다.

## **4.1 기대 수명 예측 모델 훈련 및 시각화**

제일 간단한 다중 선형 회귀 모델을 훈련한 다음, 훈련 세트와 테스트 세트에 대한 결정 계수() 값을 보고, 모델이 제대로 훈련되었는지 판단했다.

Google colab의 python을 사용하여 모델을 훈련시켰으며, sklearn 패키지에 포함된 클래스들과 메서드를 사용하였다.

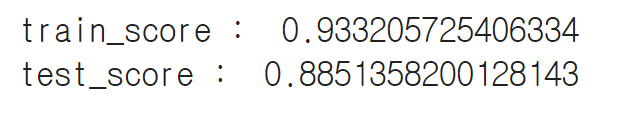
모델의 시각화는 matplotlib 라이브러리를 사용하였다.

모델의 결정 계수 값은 0.95를 기준으로 모델 훈련을 진행하겠다.

(김보람 외 1명 (2017) 시군구별 사회경제적 결핍 수준과 기대수명, 건강기대수명의 연관성: 해당 논문을 참고했습니다)

### **개인별 기대 수명 예측 모델 훈련 결과**

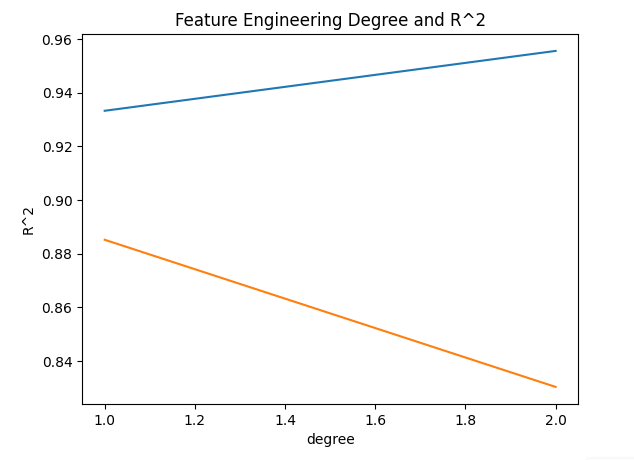
<그림 1> 다중 선형 회귀 모델 결정 계수()



해당 모델은 훈련 세트와 테스트 세트의 점수 모두 낮은 점수를 보여주고 있다. 훈련 세트는 0.9를 넘기고 있지만, 0.95보다 낮은 값이며, 훈련 세트는 이보다 낮은 0.88의 점수를 가지고 있다. 현재 모델은 과소적합이 의심되며, 해결하기 위해 특성 공학(Feature Engineering)을 통해 모델의 결정 계수 값을 높이도록 하겠다.

특성 공학(Feature Engineering)은 특정 샘플을 기준으로 모델의 특성을 늘려, 모델이 훈련 샘플에 대한 적합도를 올리는 방식이다. 모델의 특성 개수를 늘리기 위해 degree 파라미터 값을 통해 생성할 특성의 차수를 지정하고, 이 degree 값을 사용하여 훈련 샘플이 과적합(Overfitting)이 되지 않을 정도로 훈련 샘플에 대한 적합도를 올린다.

<그림 2> degree 값과 결정 계수 값의 상관 관계



위의 그래프를 봤을 때, 1부터 2까지의 그래프를 보여주고 있다. 훈련 세트에 대한 점수가 파란색, 테스트 세트에 대한 점수가 노란색이다.

<그림 3> degree 값과 결정 계수 값의 상관 관계 수치



degree가 2 이상으로 높아지면, 기하급수 적으로 테스트 샘플에 대한 적합도가 감소하기에 1과 2로 범위를 조절한 그래프를 시각화 하였다.

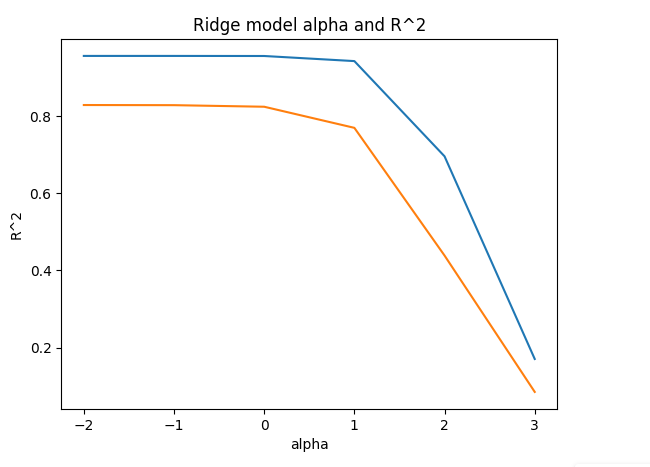
Degree 2일 때, 훈련 세트에 대한 점수가 0.95를 넘기지만, 그만큼 테스트 샘플에 대한 점수가 떨어져 과적합 문제가 발생하고 있다.

과적합 문제를 해결하기 위해 규제(Regularzation)을 사용하였다. 규제는 모델의 훈련 강도를 조절하여, 과적합을 어느정도 완화해주는 기법이다.

규제는 릿지(Ridge) 회귀 모델과 라쏘(Lasso) 회귀 모델이 있으며, 릿지(Ridge) 회귀 모델을 먼저 사용하였다.

릿지 회귀 모델은 alpha 값을 사용하여 모델에 대한 규제 강도를 조절할 수 있다. alpha 값에 따른 모델의 규제 강도의 차이를 보자.

<그림 4> alpha 값과 결정 계수 값의 상관 관계(릿지 회귀 모델)



릿지 회귀 모델의 결정 계수 값을 봤을 때, 오히려 모델이 악화되는 현상을 볼 수 있다.

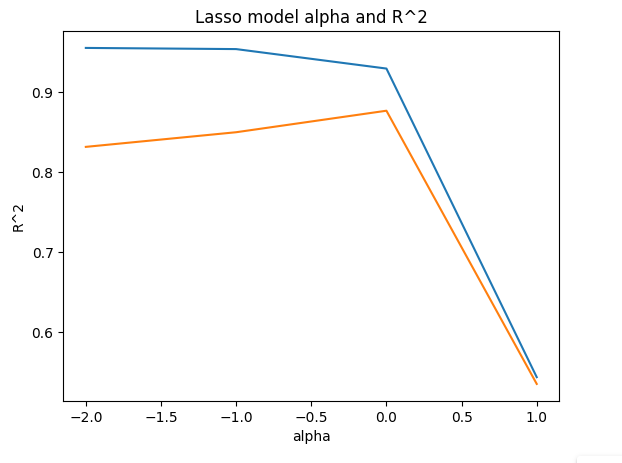
<그림 5> 릿지 회귀 모델 alpha = 0.01



릿지 회귀 모델을 사용하여 훈련 샘플에 대한 점수도 거의 바뀌지 않았으며, 테스트 샘플에 대한 점수는 오히려 떨어진 것으로 보인다.

다른 규제인 라쏘(Lasso) 회귀 모델을 사용해보자.

<그림 5> alpha 값과 결정 계수 값의 상관 관계(라쏘 회귀 모델)



라쏘 회귀 모델의 그래프는 릿지 회귀 모델과 비교해서 alpha 값을 조정하면 눈에 띄게 테스트 샘플의 점수가 올랐다.

<그림 6> 릿지 회귀 모델 alpha = 0.1



훈련 샘플의 점수가 0.95일 때, 테스트 샘플의 점수가 0.85라는 처음 다중 선형 회귀 모델에 비해서 어느정도 개선된 모습을 보인다. 다만, 테스트 샘플에 대한 점수가 비교적 감소하였으며 여전히 낮은 점수를 보여주고 있기 때문에 모델에 대한 과소적합 문제는 여전히 존재하고 있다.

이를 해결하기 위해 경사 하강법(Gradient Descent)와 손실 함수(Loss function)을 활용하여 모델을 여러 번 훈련하며 모델에 대한 정확도를 높이며, 모델의 정확도를 판단하기 위한 평균 제곱 오차(MSE/ Mean Squared Error)와 같은 다양한 기법들을 학습하여 모델의 정확도를 높이겠다.

# **5. 활용 방안**

이번 프로젝트의 모델을 활용하여 지자체와 고령 인구, 기업들을 위한 기대 수명 예측 모델을 통해 지자체에서는 공공 보건 정책과 같은 정책에 대한 참고와 기업에서는 전략적인 사업을 구상할 수 있다.

기업은 연령별 인구 분포 데이터와 연계하여 타깃 마케팅 전략을 수립하고, 연령별 소비자의 요구에 맞는 제품 개발 및 서비스를 제공할 수 있다.

정부는 기대 수명 예측을 바탕으로 노령화 대비 계획을 수립하여, 연금제도 개선, 노인 복지 프로그램의 확장 등 사회 안정을 강화할 수 있다.

# **6. 참고문헌**

박경애 and 김수영. (2007). 시도별 기대수명 차이: 성, 연령 및 사망원인. 통계 연구, 12(2), 274-302.

오영환. (2020). 다중선형회귀분석 기법을 이용한 고교야구투수의 방어율 예측. 한국지식정보기술학회 논문지, 15(4), 497-506.

김순영. (2020). 우리나라 사망력 개선 시각화와 기대수명 증가에 대한 연령별 기여도 분석. 통계연구, 25(3), 1-31.

김보람 and 윤태호. (2017). 시군구별 사회경제적 결핍 수준과 기대수명, 건강기대수명의 연관성. 보건교육건강증진학회지, 34(5), 29-40.