**빅데이터 분석 결과 보고서**

**노인 건강 분석 및 정책 제공 서비스**

**DLT(Desired Of Lifetime)**

유정민, 박의범

목차

[목차 2](#_Toc153982419)

[**1. 분석 배경 및 기획 의도** 3](#_Toc153982420)

[**1.1 분석 배경** 3](#_Toc153982421)

[**1.2 기획의도** 3](#_Toc153982422)

[**2. 분석 목표** 3](#_Toc153982423)

[**3. 분석 데이터** 4](#_Toc153982424)

[**3.1** **데이터 수집** 4](#_Toc153982425)

[**3.2** **데이터 정합성 검정** 10](#_Toc153982426)

[**3.3** **데이터 전처리** 10](#_Toc153982427)

[**4. 분석 과정 및 결과** 11](#_Toc153982428)

[**4.1 기대 수명 예측 모델 훈련 및 시각화** 12](#_Toc153982429)

[**5. 활용 방안** 18](#_Toc153982430)

[**6. 참고문헌** 18](#_Toc153982431)

# **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**

지역별 지자체와 고령 인구들을 위한 기대 수명 예측 모델을 개발하여 지역별 기대 수명 예측 데이터를 제공 함으로써 공공 보건 정책이나 전략적인 사업을 구상할 수 있다. 가장 인프라와 경제 수준이 좋은 서울과 환경은 좋지만 경제 수준이 낮은 전남을 집중적으로 지역별 건강 상태, 생활 수준, 여러 질병의 유병률 등 다양한 데이터를 분석하여 예측 모델을 만들어 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 지역별 특성을 반영한 맞춤형 예측 모델을 개발하는 것이다. 서울과 전남 지역을 중심으로 각 지역의 고유한 환경적, 경제적, 사회적 조건을 모델에 사용함으로써 지역 공동체와 관련 기관들이 보다 정밀한 데이터를 기반으로 의사결정을 내릴 수 있도록 지원한다.

# **3. 분석 데이터**

지역은 서울과 전남을 선정하여 데이터 분석과 모델 훈련을 진행하였다.

서울과 전남으로 지역을 정한 이유는 서울 지역은 경제적으로 가장 우수한 지역으로 경제적으로 매우 눈에 띄는 지표를 보이고 있으며, 인프라 측면에서도 매우 우수하다. 전남 지역은 서울과 비교하여 경제적으로도 좋지 않고, 인프라 측면에서도 부실한 지역이 많기에 전남 지역을 선정하였다.

## **데이터 수집**

- 인구 데이터

인구 데이터로는 기대 수명 데이터, 사망 인구 데이터가 있다.

기대 수명 데이터는 타겟 변수로 사용하기 위해서 수집하는 데이터로써 사용하였다.

사망 인구 데이터는 인구 10만명을 기준으로 잡은 데이터를 사용하였으며 서울과 전남 지역의 기대 수명과 연관이 깊은 데이터라고 판단하여 사용하였다.

- 경제 수준 데이터

각 지역 별 총 소득에 해당하는 GRDP 데이터를 사용하였다.

지역별 경제 상황을 보기에 가장 객관적인 데이터는 GRDP 데이터라고 판단하여 사용하였다.

- 지역별 생활 습관

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

생활 습관 데이터 전부 건강적인 측면과 간접적으로 연관성이 있기 때문에 사용하였다.

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

-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.

### **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 사망 인구, 2013-2022

단위: 10만명

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 |
| 서울 | 420.1 | 422.8 | 434.6 | 443.8 | 441.5 | 468.9 | 457.6 | 477.5 | 516.2 | 589.6 |
| 전남 | 862 | 847.9 | 873.3 | 874.8 | 892.7 | 917.3 | 900.9 | 943.6 | 958.4 | 1147 |

* 통계청(KOSIS)\_2022년 사망자수, 사망률

<표 3> 지역별 1인당 지역 총생산(GRDP), 2013-2022

단위: 10만원

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 |
| 서울 | 340,712 | 350,258 | 370,168 | 387,364 | 404,080 | 423,742 | 435,927 | 444,545 | 472,040 |  |
| 전남 | 66,476 | 66,215 | 68,878 | 71,615 | 73,732 | 75,425 | 76,948 | 78,819 | 88,384 |  |

* 통계청\_지역 총소득(2013-2022)

<표 4> 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 5> 지역별 스트레스 인지율, 2013-2022

단위: %

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 |
| 서울 | 29.4 | 30.6 | 31.5 | 30.1 | 28.1 | 27.7 | 27.2 | 29.5 | 28 | 25.7 |
| 전남 | 25 | 26.6 | 25.7 | 25.5 | 25.4 | 25.4 | 24.2 | 21.7 | 23.7 | 21.5 |

* 국민 건강 보험\_지역별 의료이용 통계연보(스트레스 인지율)(2013-2022)

<표 6> 지역별 운동 실천 비율, 2013-2022

단위: %

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2013 | 2014 | 2015 | 2016 | 2017 | 2018 | 2019 | 2020 | 2021 | 2022 |
| 서울 | 55.9 | 55.5 | 57.5 | 55.8 | 61.5 | 67 | 61 | 53.1 | 55.5 | 62.3 |
| 전남 | 40.2 | 42 | 43.4 | 38.8 | 35.3 | 39.5 | 38.8 | 36.7 | 35.9 | 44.5 |

* 질병 관리청\_지역사회건강조사(운동 실천 비율)(2013-2022)

<표 7> 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 8> 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\_당뇨병의료이용률(2013-2022)

<표 9> 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**

각 데이터의 정합성을 검사한 결과, GRDP 데이터와 기대 수명 데이터에서 문제가 발생하였다.

-기대 수명 데이터와 GRDP 데이터는 2022년 데이터를 발표하지 않았기 때문에 모두 결측치를 가지고 있다.

-결측치 데이터를 일단 NULL 값으로 처리하고, 데이터 분석 시 NULL 값을 가지는 변수들은 제외하고 분석을 진행했다.

또, 사망률 데이터와 GRDP 데이터는 10만명, 10만원 단위를 가지고 있기 때문에

공식을 사용해서 비율로 변환해준 뒤에 사용하였다. 데이터는 단위 수를 맞추기 위해 표준화를 진행했으며, 비율의 수치는 소수점 두번째 자리까지 통일하였다.

## **데이터 전처리**

조사한 모든 데이터들을 matrix 형태로 정리하고 모든 데이터들의 종속변수에 대한 상관관계를 분석하였다.

운동 실천 비율 데이터는 기대 수명과의 상관 관계가 약하기 때문에 1가지 변수는 모든 matrix에서 삭제하였다.

<그림 1>은 각 변수들이 만나는 지점이 상관계수를 나타내며 색상에 의해 상관관계를 표현한다.

정합성 검정에서 발견된 2022년 기대 수명 데이터와 GRDP 데이터는 연도와 지역으로 간단한 선형 회귀 모델을 만들어서 값을 대체하였다.

# **4. 분석 과정 및 결과**

이번 프로젝트에서 사용하는 모델은 다중 선형 회귀 모델로 지역별로 기대 수명을 예측하기 위한 모델을 훈련시킨다. 사용하는 설명 변수들을 최적화 하기 위한 변수 선택법(variable selection)은 단계적 선택법(stepwise selection)이며 단계적 선택법은 후진 소거법(Backward Elimination)을 주로 사용하였다.

사용 프로그램은 R을 사용해서 데이터 전처리, 분석, 모델 훈련 및 예측을 진행했다.

서울 지역과 전남 지역을 중점적으로 기대 수명을 예측하기 위해서 공공데이터를 사용하여 데이터를 분석하고 시각화하여 분석을 진행하였다. 사용한 독립 변수는 총 인구수, 기대 수명, 사망률, GRDP, 흡연 비율, 음주 비율, 운동 실천 비율, 스트레스 인지율이 있으며 이 데이터들의 단위가 다르기 때문에 표준화를 진행하고 모델을 단계적으로 훈련하였다.

## **4.1 기대 수명 예측 모델 훈련 및 시각화**

### **지역별 기대 수명 예측 모델 훈련 결과 및 시각화**

* 서울 모델 훈련

**(서울 기대 수명 예측 모델)**

**<표 10> 서울 기대 수명 회귀분석**

|  |  |  |  |
| --- | --- | --- | --- |
| **독립 변수** | **W** | **S** | **P** |
| **Intercept** | 2.250e+03 | 2.610e+03 | 0.547 |
| **연도** | -1.069e+00 | 1.318e+00 | 0.566 |
| **GRDP** | 1.058e-03 | 3.520e-04 | 0.205 |
| **흡연 비율** | 1.804e+00 | 8.691e-01 | 0.286 |
| **스트레스 인지율** | 6.337e-01 | 4.972e-01 | 0.424 |
| **음주율** | -3.044e+00 | 9.079e-01 | 0.185 |
| **고혈압** | 1.923e+01 | 6.617e+00 | 0.211 |
| **당뇨병** | -7.490e+01 | 2.258e+01 | 0.186 |
| **사망률** | -3.250e-01 | 1.114e-01 | 0.210 |
| **DF** | 1 | | |
| **Adjusted** | 0.9857 | | |
| **P-value** | 0.1475 | | |

W: 가중치(Weight), S: 표준 오차, p: 유의수준,

P-value: 유의수준, Adjusted : 조정된 결정 계수, DF(Degree of Freedom): 자유도(p-n)

-P-value 값이 0.1475로 0.05보다 높게 나왔기에 통계적으로 유의미한 모델이 아니라고 의심된다.

-DF(자유도) 수치가 1로 매우 낮은 것으로 측정되었다.

-각 변수에 대한 P-value 수치가 너무 높게 나왔다.

* 전남 모델 훈련

**(전남 기대 수명 예측 모델)**

**<표 11> 전남 기대 수명 회귀분석**

|  |  |  |  |
| --- | --- | --- | --- |
| **독립 변수** | **W** | **S** | **P** |
| **Intercept** | -1.112e+03 | 2.904e+02 | 0.163 |
| **연도** | 5.947e-01 | 1.455e-01 | 0.153 |
| **GRDP** | 2.375e-05 | 1.798e-05 | 0.413 |
| **흡연 비율** | 1.267e-02 | 4.797e-02 | 0.836 |
| **스트레스 인지율** | -6.553e-02 | 4.921e-02 | 0.410 |
| **음주율** | 2.506e-02 | 3.129e-02 | 0.570 |
| **고혈압** | -1.397e-01 | 2.453e-01 | 0.670 |
| **당뇨병** | -3.966e-01 | 6.791e-01 | 0.663 |
| **사망률** | -9.908e-04 | 1.451e-03 | 0.619 |
| **DF** | 1 | | |
| **Adjusted** | 0.997 | | |
| **P-value** | 0.03967 | | |

W: 가중치(Weight), S: 표준 오차, p: 유의수준,

P-value: 유의수준

Adjusted : 조정된 결정 계수 DF(Degree of Freedom): 자유도(p-n)

-p-value 값이 0.03967로 0.05보다 낮은 수치로 유의한 모델이라고 볼 수 있다.

-DF 수치가 1로 매우 낮은 수치로 측정되었다.

- Adjusted 값이 매우 높아 과적합(Overfitting) 문제가 의심된다.

-두 모델 모두 DF 수치가 1로 수치가 매우 낮은 문제가 있다. 이 문제는 특성의 수에 비해 샘플의 수가 부족하다는 의미이다. 또, 전남 모델은 과적합 문제가 의심되는데 이 문제는 모델의 샘플 수가 절대적으로 부족한 이유가 있다. 해당 문제를 해결하기 위해서는 모델의 샘플 수를 더 높이는 방법이 있다. 샘플의 수를 높이기 위해서 지역을 특성으로 두고 두 모델을 하나의 모델로 만들고 모델 훈련을 진행해보자.

### **통합 기대 수명 예측 모델 훈련 결과 및 시각화**

**(통합 기대 수명 예측 모델)**

**<표 12> 기대 수명 회귀분석**

|  |  |  |  |
| --- | --- | --- | --- |
| **독립 변수** | **W** | **S** | **P** |
| **Intercept** | -9.633e+03 | 6.749e+02 | 0.1840 |
| **지역** | 5.260e-01 | 3.393e-01 | 0.1521 |
| **연도** | -4.059e+00 | 1.998e-01 | 0.0697 |
| **GRDP** | -8.226e-06 | 7.897e-06 | 0.3221 |
| **흡연 비율** | -2.025e-02 | 1.579e-01 | 0.9005 |
| **스트레스 인지율** | -4.860e-02 | 7.148e-02 | 0.5119 |
| **음주 비율** | -7.208e-03 | 5.463e-02 | 0.8977 |
| **고혈압** | -9.726e-02 | 2.804e-01 | 0.7359 |
| **당뇨병** | -2.257e-01 | 8.339e-01 | 0.7921 |
| **사망률** | -6.208e-04 | 3.038e-03 | 0.8422 |
| **DF** | 10 | | |
| **RSE** | 0.2716 | | |
| **Adjusted** | 0.9687 | | |
| **P-value** | 9.818e-08 | | |

W: 가중치(Weight), S: 표준 오차, p: 유의수준

DF(Degree of Freedom): 자유도(p-n), RSE(Residual Standard Error): 잔차표준오차

P-value: 유의수준, Adjusted : 조정된 결정 계수

-지역 특성을 추가하여 지역별로 기대 수명 값에 기여하도록 설정했다.

-P-value 값은 0.05보다 낮은 수치로 이 모델이 통계적으로 유의함을 알 수 있다.

-Adjusted 값은 0.9687로 96.87%로 신뢰할 수 있는 모델이며 전남 모델의 과적합 문제가 해결된 것을 알 수 있다.

-모델의 DF 수치는 10으로 충분한 데이터와 특성을 가지고 훈련된 것으로 보인다.

-RSE 값은 0.2716로 매우 낮은 값으로 예측 값이 실제 값에서 벗어난 정도를 의미한다.

<표 12>의 회귀분석 결과를 단계적 선택법인 StepAIC(model, direction = “both”)를 사용해서 변수를 최적화 하였다.

**(통합 기대 수명 예측 모델)**

**<표 12-2> 최적화한 기대 수명 회귀분석**

|  |  |  |  |
| --- | --- | --- | --- |
| **독립 변수** | **W** | **S** | **P** |
| **Intercept** | -8.406e+02 | 1.861e+02 | 0.000410 |
| **지역** | 4.627e-01 | 9.358e-02 | 0.000176 |
| **연도** | -3.601e+00 | 8.793e-01 | 0.000955 |
| **GRDP** | -5.633e-06 | 2.973e-06 | 0.077547 |
| **고혈압** | -1.269e-01 | 1.003e-01 | 0.224996 |
| **DF** | 15 | | |
| **RSE** | 0.2282 | | |
| **Adjusted** | 0.9779 | | |
| **P-value** | 5.394e-13 | | |

W: 가중치(Weight), S: 표준 오차, p: 유의수준

DF(Degree of Freedom): 자유도(p-n), RSE(Residual Standard Error): 잔차표준오차

P-value: 유의수준, Adjusted : 조정된 결정 계수

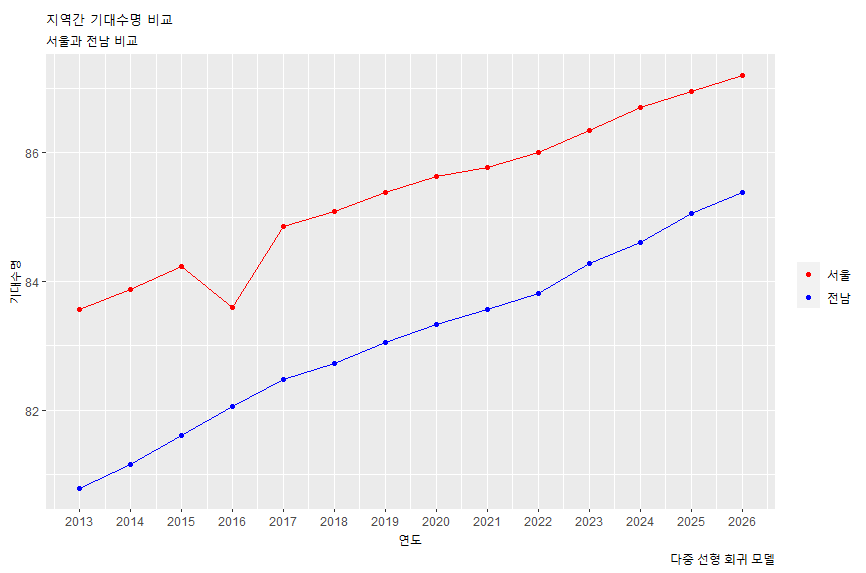
-각 변수의 p-value값도 0.05보다 높은 변수가 눈에 띄게 줄어들었다.

-p-value 값은 9.818e-08 -> 5.394e-13으로 크게 감소하였다.

-RSE 값이 0.2716 -> 0.2282로 감소했다.

이 모델을 사용한 지역별 기대 수명 예측 결과는 <그림 1>과 같다.

**<그림 1> 기대 수명 예측 결과**



분석 결과, 지역별 기대 수명 변화 추이는 계속해서 증가할 것으로 보인다.

# **5. 활용 방안**

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

# **6. 참고문헌**

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