
독거노인 인구수 예측 머신 러닝 결과 보고서

2025년 2월

소속
KG_IT뱅크학원
이름
남정한

목차

목차	2
1. 배경 및 목적	3
2. 문제 정의 및 목표	3
3. 데이터 설명	4
4. 모델 선택 및 설명	4
5. 모델 학습 및 평가	6
6. 결과 분석	8
7. 결론	9

1. 배경 및 목적

최근 몇 년간 독거노인의 인구수가 급격히 증가하고 있습니다. 통계청에 따르면, 독거노인의 수는 매년 증가 추세에 있으며, 2020년에는 독거노인이 전체 노인 인구의 20%를 넘어선 것으로 나타났습니다. 이는 우리 사회가 고령화 사회로 접어들면서 나타나는 현상으로, 앞으로도 독거노인의 수는 더욱 증가할 것으로 예상됩니다. 독거노인의 급증은 단순한 인구 변화에 그치지 않고, 다양한 사회적 문제를 야기하고 있습니다. 독거노인들은 신체적, 정신적 건강 문제가 많고, 사회적 고립으로 인한 외로움과 우울증을 겪는 경우가 많습니다. 또한, 경제적 어려움과 생활 수준의 차이로 인해 복지 시스템에 대한 의존도가 높아지는 상황이 지속되고 있습니다.

이와 같은 문제를 해결하기 위해서는 독거노인의 인구 변화를 예측하고 이에 대응하는 정책을 마련하는 것이 중요합니다. 본 연구의 목적은 머신 러닝 모델을 활용하여 독거노인의 인구수를 예측하는 것입니다. 독거노인 수의 변화를 정확히 예측할 수 있다면, 사회복지 시스템 및 정책적 대응을 보다 효과적으로 수립할 수 있으며, 고립된 노인들에게 필요한 지원을 시기 적절하게 제공할 수 있을 것입니다. 또한, 이러한 예측 모델은 향후 독거노인과 관련된 다양한 사회문제들을 해결하는 데 중요한 역할을 할 수 있습니다. 본 연구에서는 데이터 분석과 머신 러닝을 통해 독거노인의 인구 변화를 예측하고, 그에 대한 대응 전략을 마련하는 데 기여하고자 합니다.

2. 문제 정의 및 목표

본 연구의 주요 문제는 독거노인 수의 예측입니다. 현재 독거노인의 수는 급격히 증가하고 있으며, 이는 우리 사회의 고령화 문제와 밀접하게 연관되어 있습니다. 독거노인의 수를 정확하게 예측하는 것은 관련 정책의 수립 및 복지 서비스의 효율적인 제공을 위해 매우 중요합니다. 하지만 독거노인 수는 단순히 과거 데이터를 기반으로 한 예측에 그치지 않고, 다양한 사회적, 경제적 요인들이 복합적으로 작용하는 문제이기 때문에 예측 모델을 개발하는 데 어려움이 따릅니다.

본 연구에서는 독거노인 수의 예측을 목표로 하여, 여러 독립 변수들을 기반으로 종속 변수인 독거노인 수를 예측하는 모델을 구축하려고 합니다. 이를 위해 사용될 독립 변수는 총 인구수, 인구성장률, 노인인구, 노령화지수, 독거 노인수 등으로, 각 변수는 독거노인의 수와 밀접한 관계를 맺고 있습니다. 예를 들어, 총 인구수와 인구성장률은 사회의 전체적인 변화 추이를 반영하며, 노인인구와 노령화지수는 고령화 사회로 접어드는 속도와 그로 인한 영향을 보여줍니다. 마지막으로, 독거노인 수는 가장 직접적인 변수로, 이전의 독거노인 수를 바탕으로 미래의 독거노인 수를 예측할 수 있는 중요한 지표로 작용합니다.

따라서, 본 연구의 목표는 이러한 독립 변수들을 통해 독거노인 수를 예측하는 것입니다. 이를 통해 예측된 독거노인 수는 향후 사회적, 정책적 대응을 위한 중요한 지표가 될 수 있으며, 복지 시스템의 개선, 자원 배분의 효율화 및 고립된 노인들을 위한 맞춤형 지원 정책 수립에 큰 도움이 될 것입니다. 또한, 본 연구에서는 다양한 머신 러닝 기법을 활용하여 예측 모델의 정확도를 높이고, 실제 데이터에 대한 분석을 통해 예측 모델의 신뢰성을 검증하는 것을 목표로 합니다.

3. 데이터 설명

본 연구에서는 독거노인 수를 예측하기 위해 여러 가지 독립 변수를 포함한 데이터를 사용하였습니다. 데이터셋은 통계청에서 제공된 통계 데이터를 기반으로 합니다. 주요 변수로는 총 인구수, 인구성장률, 노인인구, 노령화지수, 독거노인 수가 포함됩니다.

- 총 인구수:** 한국의 전체 인구 수를 나타내며, 사회 전반적인 인구 변화를 반영합니다. (2000-2070 년)
- 인구성장률:** 1 년 동안 인구의 증감률을 나타내는 지표로, 인구 변화의 속도를 측정합니다. (2000-2070 년)
- 노인인구:** 전체 인구 중 65 세 이상의 인구 비율로, 고령화 정도를 나타냅니다. (2000-2070 년)
- 노령화지수:** 노인 인구와 14 세 이하의 유소년총 인구의 비율을 통해 사회의 고령화 정도를 나타내는 지표입니다. (2000-2070 년)
- 독거노인 수:** 독거노인의 수로, 본 연구에서 예측하려는 주요 대상입니다. (2000-2024 년)

데이터는 훈련 데이터와 테스트 데이터로 분할하여 모델 학습 및 성능 평가에 사용되었습니다. 초기에는 남녀 성비도 변수로 포함하였지만 독거노인 수와 연관관계가 적다고 판단하여 독립변수에서 제거하였습니다. 독립 변수 간의 상관관계를 분석하여 모델에 적합한 형태로 변환되었습니다.

4. 모델 선택 및 설명

본 연구에서는 독거노인 수를 예측하기 위해 다중 선형 회귀(Multiple Linear Regression) 모델을 사용했습니다. 다중 선형 회귀는 여러 독립 변수들이 종속 변수에 미치는 영향을 선형적으로 모델링하는 기법으로, 다양한 특성이 종속 변수에 어떻게 영향을 미치는지 분석할 수 있어 예측에 유용한 방법입니다. 본

연구에서는 독거노인 수를 예측하는 모델을 구축하기 위해, 총 인구수, 인구 성장률, 노인 인구, 노령화 지수와 같은 사회적 요인들을 변수로 사용했습니다.

초기 접근 방법: 변화율을 활용한 예측

초기에는 2000년부터 2024년까지의 데이터를 바탕으로, 다중 선형 회귀 모델을 만들어 독거노인 수를 예측했습니다. 이 모델에서는 2000년부터 2024년까지의 평균 변화율을 기반으로 2025년 이후부터 2070년까지의 예측값을 산출했습니다. 예를 들어, 총 인구수나 노인 인구, 노령화 지수 등의 변화를 비율로 계산한 후, 이를 미래에 적용하는 방식으로 예측을 진행했습니다.

이 방법은 상대적으로 단순하고 직관적이지만, 모든 변수에 대해 일정한 변화율을 적용했기 때문에 사회적 변화나 예외적인 사건들을 반영하기 어려운 한계가 있었습니다. 즉, 예측된 값은 과거 데이터의 평균적인 변화에 근거하므로 급격한 변화나 특정 연도의 특이사항을 반영하는 데 한계가 있었습니다. 실제로 머신 러닝을 돌려본 결과 독거노인 수는 2070년에 25,000,000 명을 넘어섰습니다. 결과 값이 너무 크게 나온 것을 보고 모델을 개선해야 된다는 것을 알았습니다. 따라서 개선된 접근 방법으로 다시 예측을 시도했습니다.

개선된 접근 방법: 실제 데이터를 활용한 예측

초기 방법에서 발생한 한계를 보완하기 위해, 이후에는 실제 미래 데이터를 활용하는 방식으로 모델을 개선했습니다. 2025년부터 2070년까지의 예측을 위해, 통계청에서 예측한 값(2025-2070년)을 토대로 총 인구수, 노인 인구, 노령화 지수 등을 가져와 모델에 반영했습니다. 이렇게 함으로써 과거의 평균 변화율 대신 실제 사회적 변화를 반영한 데이터를 사용하게 되었습니다.

이 방식은 과거의 데이터를 기반으로 예측을 수행하는 대신, 사회적, 경제적 요인들이 반영된 실제 데이터를 사용하여 미래를 예측할 수 있게 해 주었습니다. 예를 들어, 총 인구수나 노령화 지수는 단순한 변화율로 예측하기보다, 연도별로 실제 변화를 반영하는 것이 더 정확한 예측을 가능하게 만들었습니다.

또한, 이 과정에서는 훈련 세트와 테스트 세트의 R^2 값을 계산하여 모델의 성능을 평가했습니다. 훈련 세트와 테스트 세트의 R^2 값은 모델이 데이터에 얼마나 잘 적합되는지, 그리고 예측 정확도를 점검하는 중요한 지표로 활용되었습니다.

모델 개선 후 예측 및 성능 평가

모델을 개선한 후에는, 예측 결과를 시각화하여 2025년부터 2070년까지의 미래 독거노인 수를 보여주었습니다. 예측된 값을 선 그래프 형태로 표시하여, 미래 변화를 보다 직관적으로 확인할 수 있도록 했습니다. 이를 통해, 예측된 독거노인 수의 변화를 명확히 시각화하고, 데이터에서 나타나는 경향성을 파악할 수 있었습니다. 또한, y 축을 백만 단위로 표시하여 예측 결과의 크기를 비교할

수 있도록 했습니다. 2025년부터 2050년까지는 가파르게 증가하다가 2050년부터 기울기가 급격하게 감소하여 2070년까지 이어집니다.

이러한 이유는 총 인구수의 감소와 인구 성장을 때문입니다. 노령화 지수는 계속해서 증가하지만 총 인구수가 감소하여 전체 독거노인수의 영향을 끼쳤습니다. 회귀 계수를 이용하여 분석하였을 때, 인구성장률이 가장 큰 영향을 끼쳤고, 그 다음으로는 노령화지수가 독거노인 수 예측에 영향을 미쳤습니다.

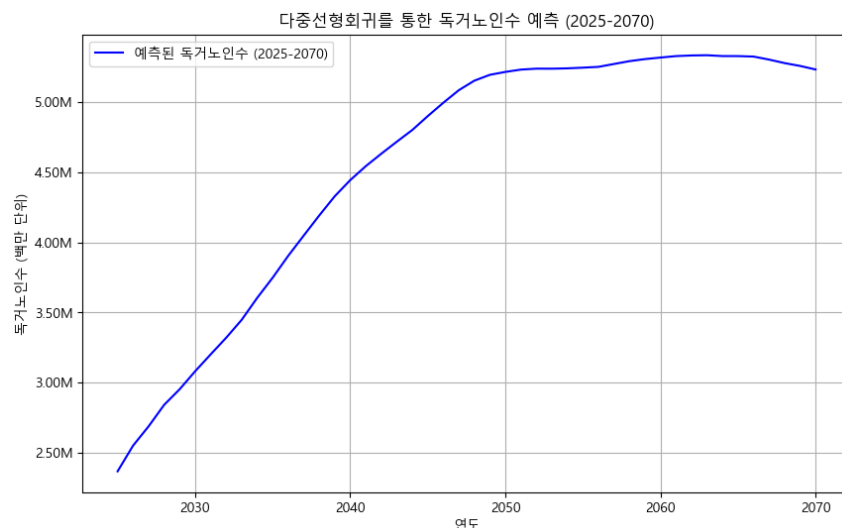


그림 1. 다중선형회귀를 통한 독거노인 수 예측(2025-2070)

모델 개선의 이유 및 결과

초기의 예측 방법에서는 변화율을 단순히 적용하여 예측값을 생성했지만, 이는 급격한 사회적 변화나 예외적인 사건들을 반영하기 어려운 점이 있었습니다. 이후에는 실제 미래 데이터(통계청에서 예측한 데이터)를 활용함으로써, 예측 모델이 사회적 변화와 현실적인 요소들을 보다 정확히 반영하도록 개선되었습니다. 또한, 변화율을 실시간 데이터로 반영하는 방식으로 예측의 신뢰성을 높였고, 그 결과 더 정교하고 현실적인 예측이 가능 해졌습니다.

따라서, 초기의 간단한 예측 방법에서 실제 데이터를 반영한 예측 방법으로 전환하면서, 독거노인 수 예측의 정확도와 신뢰도가 현저히 향상되었습니다. 이를 통해 보다 정확한 예측을 제공하고, 사회적 정책에 중요한 기초 자료를 제공할 수 있게 되었습니다.

5. 모델 학습 및 평가

본 연구에서는 독거노인 수를 예측하기 위해 다중 선형 회귀(Multiple Linear Regression) 모델을 사용하여 데이터를 학습시키고 평가했습니다. 모델 학습

과정에서는 훈련 데이터와 테스트 데이터를 나누어, 모델이 실제 데이터를 기반으로 학습하도록 했으며, 평가 단계에서는 모델이 얼마나 정확하게 예측했는지를 다양한 지표를 통해 확인했습니다.

모델 학습 과정

모델 학습을 위해, 2000년부터 2024년까지의 데이터를 사용하여 훈련 세트와 테스트 세트로 데이터를 분할했습니다. 이를 위해 `train_test_split` 함수를 사용해 데이터의 80%를 훈련 세트로, 나머지 20%를 테스트 세트로 할당했습니다. 훈련 세트는 모델이 실제 데이터를 학습하는 데 사용되며, 테스트 세트는 학습된 모델이 새로운 데이터에 대해 어떻게 예측하는지를 평가하는 데 사용되었습니다.

이후, `LinearRegression()` 클래스를 사용하여 다중 선형 회귀 모델을 생성하고, 훈련 데이터(`X_train`, `y_train`)를 사용하여 모델을 학습시켰습니다. 이 과정에서 모델은 총 인구수, 인구 성장률, 노인 인구, 노령화 지수와 같은 독립 변수들이 독거노인 수에 어떻게 영향을 미치는지를 학습하게 됩니다.

모델 평가

모델 학습 후, 평가 단계에서는 훈련 세트와 테스트 세트에 대해 각각 모델의 성능을 평가했습니다. 이를 위해 `model.score()` 함수를 사용하여 각 세트에 대한 R^2 값을 계산했습니다. R^2 값은 모델이 데이터를 얼마나 잘 설명하고 예측하는지를 나타내는 지표로, 1에 가까울수록 모델이 데이터를 잘 설명한다고 볼 수 있습니다.

- **훈련 세트 R^2 값:** 모델이 훈련 데이터에 얼마나 잘 맞는지를 나타냅니다.
- **테스트 세트 R^2 값:** 모델이 새로운 데이터(테스트 데이터)에 대해 얼마나 잘 예측하는지를 평가합니다.

이 값을 통해 모델의 과적합(overfitting) 여부를 확인할 수 있습니다. 훈련 세트에서 높은 R^2 값을 얻으면서 테스트 세트에서 낮은 R^2 값을 얻는 경우, 모델이 훈련 데이터에만 과도하게 맞춰져 새로운 데이터에서는 성능이 떨어질 가능성이 있음을 시사합니다. 본 연구에서는 훈련 세트(0.9995)와 테스트 세트(0.9989)에서 모두 높은 R^2 값을 얻어 모델이 잘 일반화되었음을 확인했습니다.

모델 개선 후 평가

초기 모델에서는 평균 변화율을 기반으로 예측을 진행했으나, 이후에는 실제 데이터를 반영하는 방식으로 예측을 개선했습니다. 이를 통해 훈련 세트와 테스트 세트에서 예측 정확도가 개선되었음을 확인할 수 있었습니다.

모델 개선 후, 2025년부터 2070년까지의 예측을 진행했으며, 예측값을 시각화하여 미래 독거노인 수의 변화를 명확히 확인할 수 있었습니다. 또한, 모델 학습 과정에서 예측값과 실제값을 비교함으로써 R^2 값을 기준으로 예측 모델의 성능을 지속적으로 평가할 수 있었습니다.

6. 결과 분석

본 연구에서는 다중 선형 회귀 모델을 사용하여 독거노인 수를 예측하였으며, 예측 결과를 실제 데이터와 비교하여 모델의 성능을 분석했습니다. 결과적으로, 모델은 훈련 세트와 테스트 세트에서 모두 높은 평가를 받았으며, 예측값은 독거노인 문제에 대한 중요한 시사점을 제공할 수 있었습니다.

모델 예측 성능

모델은 2000년부터 2024년까지의 독거노인 수에 대해 훈련되었고, 테스트 세트에서 예측을 진행한 결과 높은 R^2 값을 기록했습니다. 이는 모델이 실제 데이터를 잘 학습하고, 독거노인 수를 예측하는 데 효과적인 성능을 보였다는 것을 의미합니다. 훈련 세트와 테스트 세트에서 모두 높은 R^2 값을 기록했으므로, 모델은 과적합(overfitting) 없이 일반화된 예측을 수행할 수 있었습니다.

특히, 2025년 이후의 예측 결과는 모델이 사회적 변화를 잘 반영하고 있다는 점에서 의미가 있습니다. 예측된 독거노인 수는 예상보다 증가하는 경향을 보였으며, 이는 고령화 사회와 독거노인 문제가 심각해질 것이라는 사회적 현실을 반영한 결과로 해석할 수 있습니다. 예측값은 2025년부터 2070년까지의 데이터를 기반으로 제공되었으며, 이를 통해 향후 독거노인 증가 추세를 명확히 확인할 수 있었습니다.

미래 예측 결과

모델을 통해 예측된 2025년부터 2050년까지의 독거노인 수는 꾸준히 증가하는 경향을 보였으며, 2050년부터 2070년까지는 거의 변화가 없었습니다. 특히 노령화지수와 인구 성장률이 독거노인 수에 미치는 영향을 명확히 확인할 수 있었습니다. 예측된 독거노인 수는 2025년 이후 지속적으로 증가할 것으로 예상되었으며, 이는 고령화 사회로의 진입과 관련된 다양한 사회적 문제들을 암시합니다. 예를 들어, 사회적 고립이나 경제적 어려움 등 독거노인들이 겪을 수 있는 문제들이 더 심각해질 가능성이 높다는 점에서, 이는 향후 정책적 대응의 필요성을 강조합니다.

7. 결론

본 연구에서는 독거노인 수를 예측하기 위해 다중 선형 회귀 모델을 사용하여, 2000년부터 2024년까지의 데이터를 학습한 후, 2025년부터 2070년까지의 독거노인 수를 예측하였습니다.(독거노인 수를 제외한 나머지 독립변수들은 2070년까지 데이터로 학습) 예측 결과, 독거노인 수는 지속적으로 증가하는 추세를 보였지만, 총 인구의 감소로 인하여 2050년부터는 독거노인의 인구수가 정체되었습니다. 하지만 총 인구 수 대비 독거노인 수는 점점 증가하였으므로 이는 고령화 사회와 함께 독거노인 문제가 심각해질 것이라는 중요한 시사점을 제공합니다.

모델은 훈련 세트와 테스트 세트에서 모두 높은 예측 성능을 보였고, 특히 독거노인 수 예측에서 실질적인 사회적 의미를 도출할 수 있었습니다. 예측된 데이터는 독거노인 문제를 해결하기 위한 정책적 대응의 필요성을 강조하며, 향후 사회적 고립, 경제적 어려움 등 다양한 문제를 해결하기 위한 대비책을 마련하는 데 기여할 수 있을 것입니다.

또한, 본 연구에서 사용된 다중 선형 회귀 모델은 상대적으로 간단한 모델이지만, 향후 더 복잡한 비선형 모델이나 딥러닝 모델을 통해 예측의 정확성을 한층 개선할 수 있을 것으로 기대됩니다. 또한, 사회적 요인들에 대한 추가적인 데이터 수집과 분석이 이루어진다면, 더욱 정교한 예측 모델을 구축할 수 있을 것입니다.

이 연구의 결과는 독거노인 문제에 대한 사회적 인식 제고뿐만 아니라, 정책 결정자들에게 유용한 정보를 제공할 수 있으며, 향후 독거노인 문제를 해결하기 위한 다양한 정책 개발에 기여할 수 있을 것입니다. 예측 모델을 기반으로 한 정책적 접근이 이루어질 경우, 독거노인들의 삶의 질 향상과 사회적 안전망 구축에 실질적인 도움이 될 수 있을 것입니다.