**사망보험에서의 건강나이 빅데이터 플랫폼 제안**

**팀 명 : Kwkw**

**팀 원 : 김영빈 , 이현경**

**목 차**

**Ⅰ. 요약(abstract) ----------------------------------------4p**

**Ⅱ. 서론**

**1. 보험업계 현황 분석 ------------------------------------------ 5p**

**(1) 생명보험의 종류**

**(2) 우리나라의 생명보험 회사의 보유 계약 건수**

**(3) 국민들의 생명보험 인식**

**2. 보험의 문제점 인식 및 개선 방향 -----------------------------10p**

**(1) 현행 사망보험 문제점 인식**

**(2) 현행 사망보험의 개선 방안**

**Ⅲ. 본론**

**1. 전 처리를 통한 건강나이 산출 대상자 선정 ---------------------14p**

**(1) 사망원인 및 과거 데이터 전 처리**

**1) 하위 6항목 제거**

**2) 달리 분류되지 않는 증상∙징후’&‘질병이환 및 사망의 외인’항목축소**

**3) 과거 데이터 제거**

**4) 이상치(outlier) 제거**

**5) 자료 신뢰성 검토**

**(2) 세부 사망 원인 순위**

**2. 본격적인 건강나이 산출 --------------------------------------32p**

**(1) 실제 데이터 소개(건강검진 데이터)**

**(2) 분석 방법론 소개**

**1) 군집 분석0**

**2) 회귀 분석**

**① 선형 회귀**

**② 다항 회귀**

**③ 그래디언트 부스팅 회귀**

**④ 랜덤 포레스트 회귀**

**⑤ 보험료 산출 방법**

**(3) 실제 예시를 통해 예측하기**

**1) 군집화 모델로 예측**

**2) 회귀 모델로 예측**

**Ⅳ. 결론**

**1. 보험계약자가 얻을 수 있는 이점 ----------------------------------47p**

**2. 보험회사가 얻을 수 있는 이점 --------------------------------------48p**

**3. 국가적 차원에서 얻을 수 있는 이점 ---------------------------------49p**

**4. 논문의 한계점 ------------------------------------------------ 49p**

**Ⅴ. 참고 문헌**

**표 목 차**

**[표 - 01] 생명보험회사의 보유 계약 건수**

**[표 – 02] 2015년 ~ 2019년 보험 가입률**

**[표 - 03] 2017년 ~ 2019년 3개년도 개인 연령별 생명보험 가입률**

**[표 - 04] 생명보험에 대한 인식 조사**

**[표 - 05] 생명보험의 소비자 이해도 수준**

**[표 - 06] 생명보험의 가입 의향이 없는 이유**

**[표 - 07] 1983년 ~ 2018년 남/년 전 연령 사망자 수 및 사망자 비율**

**[표 - 08] 하위 6항목 사망자 수 및 사망자 비율**

**[표 - 09] ‘달리 분류되지 않은 증상 및 징후’의 세부 중 상위 2항목**

**[표 - 10] 남성 전 연령 시대별 사망원인 분석**

**[표 - 11] 여성 전 연령 시대별 사망원인 분석**

**[표 - 12] 시대별 남녀 전 연령 사망원인조사 순위**

**[표 - 13] 5년 단위의 사망 통계 요약(summary)**

**[표 - 14] 2014년~ 2018년 30세 이상 남성의 사망원인 별 사망자 수**

**[표 - 15] 2014년~ 2018년 30세 이상 여성의 사망원인 별 사망자 수**

**[표 - 16] 연령대 그룹 구분**

**[표 - 17] 예시 – 2014년 ~ 2018년도 남성 데이터**

**[표 - 18] 예시 – 2014년 ~ 2018년도 여성 데이터**

**[표 - 19] 각 특성의 중요도**

**[표 - 20] 60세 만기 전기납, 주 계약 1억원 기준**

**그림 목차**

**[그림 - 01] 1983년 ~ 2018년도 남녀 사망 원인 변화**

**[그림 - 02] 5년 단위 남성의 연령별 사망자수**

**[그림 - 03] 5년 단위 여성의 연령별 사망자수**

**[그림 - 04] 2004년 ~ 2018년 남성 전체의 boxplot과 분포표**

**[그림 - 05] 2004년 ~ 2018년 여성 전체의 boxplot과 분포표**

**[그림 - 06] 예시-독립변수(트리글리세라이드)의 군집점수**

**[그림 - 07] 선형 회귀 모델링 코드**

**[그림 - 08] 다중 회귀 모델링 코드**

**[그림 - 09] 그래디언트 부스팅 모델링 코드**

**[그림 - 10] 랜덤 포레스트 모델링 코드**

**[그림 – 11] 남성 17항목을 통한 군집화 점수 산출**

**[그림 – 12] 남성 17항목을 통한 군집화 점수 상위자 나열**

**Ⅰ. 요약**

2020년도를 살아가는 우리는 예상치 못한 사고를 대비하기 위한 보험의 필요성을 인식하고 있습니다. 이를 인식한 각 보험회사는 각종 보험을 개발하여 고객들의 위험을 보장해주고 있습니다.

여러 종류의 보험회사 중 생명보험회사가 판매하는 보험은 약80%정도가 사망을 보장해주는 사망보장형 상품으로 나타납니다. 하지만, 여러 가지의 이유로 보험소비자들에게 관심이 멀어지고 있습니다.

개선 방안 첫 번째로, 보험계약자들의 형평성(제고)로 개선방안을 생각하였습니다. 우리는 통계청에서 발표하는 통계 자료를 통해 1) 군집 분류를 통한 보험료 할인 시스템, 2) 머신러닝 기법을 사용한 건강나이 예측시스템을 구축했습니다.

두 번째는, 보험금 이외의 개별계약자들이 보험가입을 통하여 얻는 이점을 만들어 줘야 한다고 생각하였습니다. 그에 대한 방안으로 헬스케어 서비스를 제시합니다.

이를 통하여 얻을 수 있는 이점은

◼ 보험계약자 : 병원과 보험회사에서 이중적으로 자신의 건강을 관리(care)해주며 의식을 함양시켜 준다는 점

◼ 보험회사 : 보험계약초기 이후 개별 보험계약자들의 건강정보를 얻을 수 있어 타 보험에서의 위험율 산출 및 개발 시 데이터가 다양해진다는 점

◼ 국가 차원 : 보험회사에서도 국민건강의식수준을 함양시켜주기 때문에 사회적 비용이 줄어든다는 점

이 있습니다.

**Ⅱ. 서론**

**1. 보험업계 현황 분석**

**(1) 생명보험의 종류**

생명보험이란, 피보험인을 대상으로 하여 사고나 재해를 보상해주는 것이고, 좁은 의미로는 피보험인이 사망하였을 때에 소정의 금액을 보상해주는 보험입니다. 생명보험은 크게 3가지의 종류로 분류할 수 있습니다.

* 사망보험 : 피보험자의 사망을 보험사고로 인식하는 보험계약으로 종신기간 동안 사망을 보험사고로 인식하는 종신보험과 일정기간 동안의 사망을 보험사고로 인식하는 정기보험이 존재합니다.
* 생존보험 : 피보험자가 일정한 연령에 도달하는 것을 보험사고로 인식하는 보험계약입니다.
* 생사혼합보험 : 생존보험과 사망보험을 합친 보험계약으로서, 피보험자가 일정한 연령에 도달하는 것과 일정한 연령 도달 전 사망하는 것을 모두 보험사고로 인식하는 결합보험계약입니다.

**(2) 우리나라의 생명보험회사의 보유계약건수**

[표 – 01] 생명보험회사의 보유 계약 건수[[1]](#footnote-1)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 보유계약건수  (건)  기업명 | 개인보험 | | | 단체보험 |
| 생존보험 | 사망보험 | 생사혼합보험 |
| 한화생명 | 1,498,691 | 7,878,655 | 372,928 | 984,924 |
| 삼성생명 | 252,415 | 1,382,591 | 227,249 | 793,597 |
| ABL생명 | 2,763,981 | 12,762,544 | 312,816 | 1,363,319 |
| 흥국생명 | 197,926 | 2,088,685 | 312,968 | 18,045 |
| 교보생명 | 1,596,568 | 5,976,746 | 146,252 | 398,860 |
| DGB생명 | 31,533 | 440,891 | 120,678 | 981,130 |
| 미래에셋생명 | 358,959 | 1,905,960 | 254,959 | 39,585 |
| KDB생명 | 311,185 | 1,421,635 | 251,166 | 302,808 |
| DB생명 | 65,734 | 1,220,565 | 76,273 | 8,282 |
| 동양생명 | 197,220 | 2,876,457 | 696,752 | 22,336 |
| 메트라이프생명 | 147,266 | 1,153,466 | 247,184 | 180 |
| 푸르덴셜생명 | 136,699 | 1,049,207 | 24,695 | - |
| 신한생명 | 410,589 | 4,420,050 | 330,340 | 40,215 |
| 처브라이프생명 | 12,851 | 119,320 | 23,430 | 164 |
| 오렌지라이프생명 | 269,673 | 1,465,989 | 126,009 | - |
| 하나생명 | 48,096 | 126,328 | 55,422 | 319,992 |
| KB생명 | 183,889 | 202,512 | 78,316 | 23,815 |
| 총 보험가입금액(백만원)[[2]](#footnote-2) | 287,867,683 | 1,559,430,265 | 116,381,580 | 53,365,642 |

\* 생명보험협회 자료 참조

위 표는 ‘우리나라 생명보험회사의 보유계약건수’입니다. 이를 통하여, 생명보험을 크게 3종류로 나누었을 때 대부분의 생명보험회사는 사망보험에 치중하고 있음을 알 수 있습니다[[3]](#footnote-3). (사망보험계약건수가 약79.293%, 계약건수는 약4649만계약의 비중을 차지함.) 또한, 총 보험가입금액을 보면, 약1560조원의 금액을 지급할 의무를 보험회사가 지니고 있습니다. 이는 보험회사에서 관리하는 자산이며, 이러한 큰 금액을 위험률, 할인율, 사업비율 등 기초요율을 토대로 안전하게 관리해야 한다는 것을 알 수 있습니다.

만일 기초요율의 안정성이 유지되지 않는다면, 요율의 불안정성이 보험회사의 리스크로 발생하게 되고 보험손익의 불안정성으로 확대될 것입니다.

**(3) 국민들의 생명보험 인식**

위에 나타난 [표 –01]을 보면, 현재 우리나라에서는 생명보험만으로도 많은 보험계약이 유지되고 있으며, 손해보험∙연금보험∙재보험 등 모든 보험을 합치면 현재 삶을 살아가며 보험을 빼고는 살아갈 수 없다라고 해도 과언이 아닙니다.

하지만, 다른 금융상품들과 다르게 부합계약이기 때문에 단순히 보험회사의 현행 보험유지계약건수만으로‘보험회사의 신뢰도 및 공정성이 높다’라고 평가할 수 없습니다. 소비자가 보험에 대해 잘 알고 꼭 필요한 보험을 가입하고 있는지 조사할 필요가 있습니다. 이에 따라, 보험소비자의 보험회사 만족도 및 신뢰도를 파악하고, 보험계약자들의 보험인식현황에 대해 알아봅니다.

[표 - 02] 2015년 ~2019년 보험 가입률 (단위 : %)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 2015년 | 2016년 | 2017년 | 2018년 | 2019년 |
| 생명보험 | 78.9 | 73.4 | 78.2 | 79.5 | 72.7 |
| 손해보험 | 79.7 | 76.2 | 78.4 | 80.0 | 76.2 |

\* 2019 보험소비자 설문조사 참조

[표 - 03] 2017년 ~2019년 3개년도 개인연령별 생명보험 가입률 (단위 : %)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 20대 | 30대 | 40대 | 50대 | 60대 |
| 2017년 | 63.0 | 81.7 | 86.6 | 87.7 | 71.5 |
| 2018년 | 77.1 | 87.6 | 88.8 | 83.8 | 66.6 |
| 2019년 | 66.5 | 82.6 | 85.8 | 83.8 | 64.5 |

\* 2019 보험소비자 설문조사 참조

위 [표 - 02]을 보면, 2019년도 보험가입률이 최근 5년에 대해서 최저를 기록하고 있다는 것을 알 수 있습니다. 이는 저금리시대의 도달 및 경기성장률의 감소에도 간접적인 영향이 있을 수 있지만 실질적으로 보험계약에 대한 보험계약자들의 불만 또는 리스크는 없는지 확인하여 개선할 필요가 있습니다.

[표 - 04] 생명보험에 대한 인식조사

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 아주 좋다 | 좋다 | 예전과 동일하다 | 나빠졌다 | 아주 나빠졌다 | 모르겠다 |
| 2015년 | 3.8 | 29.1 | 48.9 | 2.9 | 0.2 | 15.0 |
| 2018년 | 2.5 | 33.2 | 49.3 | 4.8 | 0.5 | 9.8 |

\* 2018 생명보험 성향조사 참조

[표 - 05] 생명보험의 소비자 이해도수준[[4]](#footnote-4) (단위 : %)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 0~3개 | 4~6개 | 7~9개 | 10개 이상 |
| 2015년 | 22.1 | 39.9 | 24.9 | 13.3 |
| 2018년 | 7.1 | 24.1 | 36 | 32.9 |

\* 2018 생명보험 성향조사 참조

[표 -05]를 보면, 과거에 비해 높은 이해도 수준이 됐음을 알 수 있습니다. [표 -04]에서는 ‘좋다/나빠졌다’의 증가 및 ‘모르겠다’의 감소가 눈에 띄는 항목입니다. 이는 보험계약자 본인이 보험에 대해서 올바르게 판단할 만큼 수준이 향상되었음을 의미합니다.

즉, [표 – 01, 02, 03, 04, 05]를 토대로 도출할 수 있는 내용은

1) 현행 생명보험회사들의 주요 상품은 사망보험상품 이라는 것.

2) 보험 가입률의 최근 5개년 추이를 보면 2019년이 역대 최저를 기록했다는 점.

3) 생명보험의 인식조사 결과가 명확하게 긍정 또는 부정이 나뉘어지고, 고객들의 보험 이해도 수준이 향상되었다는 것.

입니다.

이 세 가지의 결과를 토대로, 보험가입자(보험계약자)는 부합계약성을 띤 보험계약을 전보다 더 자세히 이해하고 자신에게 알맞은 상품을 가입하고 있음을 알 수 있습니다. 따라서, 보험상품을 개발할 시, 고객들의 긍정적 신호를 얻을 수 있는 방향으로 제시하려 합니다. 현재 보험계약의 가입률 추세가 감소하고 있는 것이 무리한 부합계약으로 인한 탈퇴가 아닌지 되돌아보고, 이를 개선할 수 있는 방향으로 보험요율의 산출방식을 변경할 필요가 있습니다.

**2. 보험의 문제점 인식 및 개선 방향**

**(1)현행 사망보험의 문제점**

첫 번째 현행 사망보험의 문제점은 ‘보험료 액수’입니다.

현행 사망보험은 정기보험과 종신보험으로 나뉘어지며, 사망을 보험사고로 인식하여 사망 발생시 보험회사가 보험가입금액을 주는 보험계약입니다. 이 때, 30세 남자 1억 사망종신보험 가입 시 주계약 기준으로 보험료는 대부분 약 30만원입니다[[5]](#footnote-5). 이 금액은 다른 보험의 보험료에 비해 매우 높은 금액입니다. 종신보험은 사망 했을 시 보험가입금액을 보장하는데, 사람의 사망률은 100%이므로 보험 회사는 보험리스크를 감당하기 위해 보험료를 올릴 수 밖에 없습니다.

[표 - 06] 생명보험의 가입 의향이 없는 이유 (단위 : %, 복수응답가능)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 경제적으로 여유가 없어서 | 국가의 보장을 기대해서 | 보험기간이 너무 길어서 | 다른 투자나 저축이 유리해서 |
| 2015년 | 80.3 | 11.0 | 17.1 | 10.5 |
| 2018년 | 70.4 | 16.5 | 15.5 | 13.5 |

\* 2018 생명보험 성향조사 참조

[표 - 06]에서도 나타나듯, 경제적 이유로 보험료납입이 무리가 있다는 응답이 예전보다는 감소하였으나 그래도 70%나 차지합니다.

두 번째로는 ‘보험기간’입니다. 보험기간이란 보험가입시점부터 보장시점까지의 보험보장기간을 뜻합니다. 사망이라는 보험사고는 먼 미래에 발생을 하고 보험료납입시점은 보험가입시점부터 발생합니다. 그러다 보니, 자신이 사망할 수 있다는 생각을 하고 있지 않을 때부터 보험료를 납입할 수밖에 없는 시스템은 보험가입률을 낮추는 요인이 됩니다.

셋째로는 ‘보험료의 형평성’입니다. 현행 보험가입방식은 일반체/건강체를 구분하여 보험료를 달리 하지만, 각 개인에게 알맞은 보험료를 산출하여 개인별 형평성에 맞는 보험가입이 진행되고 있지는 않습니다. 따라서, 몸 관리, 운동 습관 등 개인의 건강 및 생활 습관에 따라 보장기간이 짧거나 길어질 수 있습니다. 예를 들어, 30세의 두 남성이 동일한 1억 정액보장 상품을 가입하였는데, 한 사람은 60세에, 다른 사람은 100세에 사망했다고 가정해 봅시다. 이 때, 60세에 사망하는 사람이 보험금액입장에서는 유리합니다. 그 이유는 정액보장형상품 대부분이 물가연동형으로 이루어 지지 않기 때문입니다. 60세에 받는 1억과 40년 후에 받는 1억은 물가 차이로 인해 그 가치가 달라집니다. 보험료를 내는 시점은 동일하지만, 보험가입금액을 받는 시점에서의 차이가 발생하면서 개인별로 수지상등의원칙을 이루지 못하게 됩니다. 이에 따라 소비자가 ‘보험이 비싸다’, ‘건강하면 사망보험을 들 필요가 없다’라는 인식이 생깁니다.

**(2) 현행 사망보험의 개선방안**

현행 사망보험에서의 문제점은 아주 기본적인 보험료산출방식에 원인이 있음을 알 수 있습니다.

제안하는 개선방안은 ‘할인◦할증을 통해 형평성을 맞춘 보험료 산출’입니다. 형평성의 문제는 “보험업법 제129조(보험요율 산출의 원칙) 3. 보험요율이 보험계약자 간에 부당하게 차별적이지 아니할 것”이라고 작성되어 있을 만큼 보험요율 산출 시 중요하게 보는 대목입니다. 이를 해석하면, 보험료가 이유 없이 차별적이면 안되지만, 합리적인 이유가 발생할 시에는 차별을 둘 수 있다는 것으로 해석할 수 있습니다.

또한, 현행 자동차보험 제도를 보면, 계약자들의 공정한 보험료 산출을 위하여 피보험자의 연령, 운전자 범위, 운행거리 등 다양한 차등화 요소와 사고 경력에 따른 개별적인 위험을 반영하여 산출[[6]](#footnote-6)하고 있습니다. 다시 말하면, 운전자의 위험수준에 부합하는 적정보험료를 부과하기 위해 無사고자에게는 보험료를 할인하고 有사고자에게는 보험료를 할증하는 방법을 적용한다는 것입니다. 이로 인해, 민원을 감소시키고 보험료를 자신의 사고빈도 및 심도와 연결 지어 적정보험료를 개별적으로 부과하여 민원을 감소시키는 기대효과를 가져옵니다.

예를 들어[[7]](#footnote-7), 사고건수요율을 적용하여 할인◦할증 폭을 확인해보면,

◼ 직전 1년간 무사고이면서, 직전 3년간 사고가 1건 이하인 경우 : 보험료할인 약 3%~11%할인

◼ 직전 1년간 사고가 1건 이상이거나, 직전 3년간 사고가 2건 이상인 경우 : 보험료할증 약 6% ~ 30%

인 것을 보아, 할인◦할증 변동 폭을 쉽게 알 수 있습니다.

이처럼, 현행 할인할증방식은 자동차보험에서 가장 크게 자리잡았으며, 자동차보험 이외의 여러 보험에서도 개인별 할인할증방식을 도입하려는 움직임을 보이고 있습니다.

따라서, 우리는 할인◦할증 제도를 사망보험에 도입하여, 보험료를 개인별 상황에 맞게끔 다시 산출하려 합니다. 그 방법은 1) 군집 분류를 통해 건강점수를 산출하고 이를 토대로 보험료를 할인하는 방법, 2) 머신러닝을 이용해 건강나이를 예측하여 현행 보험나이를 재 산출시켜 보험료를 변동시키는 방법입니다.

이를 통해, 1) 보험계약자의 민원 감소 2) 건강나이(점수) 산출이라는 헬스케어 서비스 제공을 통한 사망보험의 가입률 증진 3) 보험회사의 보험금 지출 유동성문제를 안정화시킴으로써 투자수익 극대화 등의 효과를 기대합니다.

**Ⅱ. 본론**

우선, 우리는 통계청에서 발표하는 사망원인분석자료를 사용하여 우리나라의 사망원인 및 사망에 대한 여러 가지 자료를 뽑아내려 합니다. 통계청에서는 1983년도부터 2018년까지의 36년의 데이터를 공개하고 있으며, 이를 우리가 원하는 방향으로 맞추기 위하여 데이터를 전처리 하고자 합니다.

이에 앞서, 통계청에서 발표하는 사망원인은 크게 19가지의 대 분류가 가능하며 분류 내용은 다음과 같습니다.

◼ 특정 감염성 및 기생충성 질환(A00-B99)

◼ 신생물(C00-D48)

◼ 혈액 및 조혈기관질환과 면역메커니즘을 침범하는 특정장애(D50-D89)

◼ 내분비, 영양 및 대사 질환(E00-E88)

◼ 정신 및 행동장애(F01-F99)

◼ 신경계통의 질환(G00-G98)

◼ 눈 및 눈부속기의 질환 (H00-H57)

◼ 귀 및 유돌의 질환 (H60-H93)

◼ 순환계통 질환 (I00-I99)

◼ 호흡계통의 질환 (J00-J98, U04)

◼ 소화계통의 질환 (K00-K92)

◼ 피부 및 피하조직의 질환 (L00-L98)

◼ 근골격계통 및 결합 조직의 질환 (M00-M99)

◼ 비뇨생식계통의 질환 (N00-N98)

◼ 임신, 출산 및 산후기 (O00-O99)

◼ 출생전후기에 기원한 특정병태 (P00-P96)

◼ 선천 기형, 변형 및 염색체 이상 (Q00-Q99)

◼ 달리 분류되지 않은 증상, 징후 (R00-R99)

◼ 질병이환 및 사망의 외인 (V01-Y89)

**1. 전 처리를 통한 건강나이 산출 대상자 선정**

**(1) 사망원인 및 과거 데이터 전 처리 [[8]](#footnote-8)**

**1) 하위 6항목 제거**

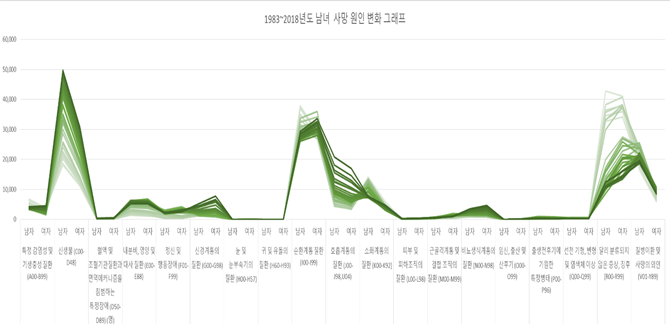
36년의 사망원인분석자료를 통합해 나타내 보았습니다.

[표 - 07] 1983년 ~ 2018년 남/녀 전(全) 연령 사망자 수 및 사망자 비율

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 남/녀 전(全)연령 | 남성 | | 여성 | |
| 사망자수 | 사망자 비율 | 사망자수 | 사망자 비율 |
| 특정 감염성 및 기생충성 질환 (A00-B99) | 144,797 | 2.863% | 88,990 | 2.244% |
| 신생물 (C00-D48) | 1,303,224 | 25.772% | 763,184 | 19.242% |
| 혈액 및 조혈기관질환과 면역메커니즘을 침범하는 특정장애  (D50-D89) | 7,793 | 0.154% | 9,208 | 0.232% |
| 내분비, 영양 및 대사 질환 (E00-E88) | 165,057 | 3.264% | 160,800 | 4.054% |
| 정신 및 행동장애 (F01-F99) | 68,687 | 1.358% | 83,168 | 2.097% |
| 신경계통의 질환 (G00-G98) | 78,820 | 1.559% | 90,785 | 2.289% |
| 눈 및 눈부속기의 질환 (H00-H57) | 136 | 0.003% | 205 | 0.005% |
| 귀 및 유돌의 질환 (H60-H93) | 78 | 0.002% | 72 | 0.002% |
| 순환계통 질환 (I00-I99) | 1,074,965 | 21.258% | 1,084,746 | 27.349% |
| 호흡계통의 질환 (J00-J98, U04) | 316,993 | 6.269% | 242,286 | 6.109% |
| 소화계통의 질환 (K00-K92) | 371,431 | 7.345% | 147,785 | 3.726% |
| 피부 및 피하조직의 질환 (L00-L98) | 3,912 | 0.077% | 7,170 | 0.181% |
| 근골격계통 및 결합 조직의 질환 (M00-M99) | 18,653 | 0.369% | 38,832 | 0.979% |
| 비뇨생식계통의 질환 (N00-N98) | 62,085 | 1.228% | 64,501 | 1.626% |
| 임신, 출산 및 산후기 (O00-O99) | 0 | 0.000% | 2,625 | 0.066% |
| 출생전후기에 기원한 특정병태 (P00-P96) | 11,851 | 0.234% | 9,286 | 0.234% |
| 선천 기형, 변형 및 염색체 이상 (Q00-Q99) | 15,011 | 0.297% | 13,312 | 0.336% |
| 달리 분류되지 않은 증상, 징후 (R00-R99) | 652,514 | 12.904% | 845,215 | 21.310% |
| 질병이환 및 사망의 외인 (V01-Y89) | 760,822 | 15.045% | 314,147 | 7.920% |

[표 - 07]를 각 연도 별로 그래프에 나타내면, 다음과 같습니다.

[그림 - 01] 1983년 ~ 2018년도 남녀 사망 원인 변화[[9]](#footnote-9)



[표 - 07]와 [그림 – 01]를 통해 알 수 있는 특징이 세 가지 있습니다.

◼ 남성 및 여성 모두 크게 4가지의 질병 분류에 의해 사망한다는 것을 알 수 있습니다. (남성의 경우, 상위 4항목 누적합산비율 = 74.979%, 여성의 경우, 상위 4항목 누적합산비율 = 75.821%)

◼ [그림 – 01]에서, 연도의 흐름에 따라 사망자 수가 증가하는 추세가 (신생물, 내분비 영양 및 대사질환 등) 있는 반면, 사망자수가 감소하는 추세(달리 분류되지 않는 증상 및 징후)를 보이는 것도 존재합니다. 이를 통해, 과거와 현재의 사망원인이 달라지고 있음을 알 수 있습니다.

◼ 첫 번째 특징에서 언급했듯이, 우리나라 사망 인구는 특정 사망원인에 몰려 있습니다. 하위 6항목의 사망자 수 및 사망자 비율이 1% 미만입니다. (남성의 경우, 하위 6항목 누적합산비율 = 0.470%, 여성의 경우, 하위 6항목 누적합산비율 = 0.72%)

[표 - 08] 하위 6항목 사망자 수 및 사망자 비율

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 남/녀 전(全)연령 | 남성 | | 여성 | |
| 사망자수 | 사망자 비율 | 사망자수 | 사망자 비율 |
| |  | | --- | | 혈액 및 조혈기관질환과 면역메커니즘을 침범하는 특정장애 (D50-D89) | | 7,793 | 0.154% | 9,208 | 0.232% |
| 눈 및 눈부속기의 질환 (H00-H57) | 136 | 0.003% | 205 | 0.005% |
| 귀 및 유돌의 질환 (H60-H93) | 78 | 0.002% | 72 | 0.002% |
| 피부 및 피하조직의 질환 (L00-L98) | 3,912 | 0.077% | 7,170 | 0.181% |
| 임신, 출산 및 산후기 (O00-O99) | 0 | 0.000% | 2,625 | 0.066% |
| 출생전후기에 기원한 특정병태 (P00-P96) | 11,851 | 0.234% | 9,286 | 0.234% |

위의 자료를 바탕으로, 건강나이 산출 시 사용할 데이터 중 하위 6항목을 제거할 수 있습니다.

**2) ‘달리 분류되지 않는 증상∙징후’ & ‘질병이환 및 사망의 외인’항목 축소**

① 질병이환 및 사망의 외인(V01-Y89)이란?

‘질병이환 및 사망의 외인(V01-Y89)’을 중분류로 나누어보면 8가지로 분류됩니다.

◼ 운수사고 (V01-V99)

◼ 낙상(추락) (W00-W19)

◼ 불의의 익사 및 익수 (W65-W74)

◼ 연기, 불 및 불꽃에 노출 (X00-X09)

◼ 유독성 물질에 의한 불의의 중독 및 노출 (X40-X49)

◼ 고의적 자해(자살) (X60-X84)

◼ 가해(타살) (X85-Y09)

◼ 모든 기타 외인 (Re. V01-Y89)

‘질병이환 및 사망의 외인’데이터는 우리가 산출할 “사망보험가입자들의 보험 가입 시 보험료 형평성을 맞추기 위한 건강나이 산출”의 목적으로 사용하기엔 문제점이 존재합니다.

[표 – 07]과 [그림 - 01]에서 ‘질병이환 및 사망의 외인’의 비율을 보면, 총 사망자중 남성 = 15.045%, 여성 = 7.920%로 비교적 높은 비율을 나타내고 있습니다. 그러나, 건강을 유지하는 보험계약자에게 보험료를 할인해주고자 하는 취지에서 볼 때, 건강과는 연관성이 먼 데이터(자살, 타살, 운수사고, 추락 등)를 포함하게 되면 오차가 크게 발생합니다.

따라서, 전체 데이터에서 차지하는 비중이 크지만, 앞으로의 논의에서 ‘질병이환 및 사망의 외인’으로 인한 사망자는 제거합니다.

② 달리 분류되지 않은 증상 및 징후(R00-R99)란?

[표 - 07] 및 [그림 - 01]에서 나타난‘달리 분류되지 않는 증상 및 징후’의 특징을 보겠습니다.

◼ 통계청에서 사망원인분석표를 발표한 1983년부터 2018년까지 총 사망자중 남성 = 12.904%, 여성 = 21.310%로서 높은 비율을 나타냅니다. 또한, 여성이 일반적으로 남성보다 약 2배 높은 사망원인을 차지하고 있습니다.

◼‘달리 분류되지 않은 증상 및 징후’의 경우, 시간이 흐름에 따라 전체 사망원인에서의 비율이 대폭 감소함을 알 수 있습니다.

남성의 경우 전(全)연령 : 24.617% ▷ 7.013%

여성의 경우 전(全)연령 : 36.002% ▷ 11.022%

[표 - 09]‘달리 분류되지 않은 증상 및 징후’의 세부 중 상위 2항목

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 달리 분류되지 않은 증상 및 징후 | 2015년 | 2014년 | 2013년 | 2012년 |
| 전체 사망자수 | 24,396 | 24,098 | 24,935 | 25,606 |
| R54 (노쇠) | 14,054 | 15,058 | 14,631 | 15,054 |
| R98-R99 (지켜본 사람이 없는 사망) | 6,427 | 6,456 | 6,551 | 6,640 |

[표 - 09]에 따르면, 대부분의 ‘달리 분류되지 않은 증상 및 징후’는 ‘노쇠’ 및 ‘지켜본 사람이 없는 사망’으로 분류가 됩니다.

따라서, 앞으로의 분석 과정에서 ‘달리 분류되지 않은 증상 및 징후’를 ‘노쇠로 인한 사망’으로 축소합니다.

**3) 과거 데이터 제거**

① 1983년 ~ 2018년

[표 – 10] 남성 전(全)연령 시대별 사망원인 분석[[10]](#footnote-10)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 전(全)연령 남성 | 1985년 | 1995년 | 2005년 | 2015년 |
| 특정 감염성 및 기생충성 질환 (A00-B99) | 4.034% | 2.752% | 2.541% | 2.495% |
| 신생물 (C00-D48) | 14.435% | 23.770% | 30.805% | 32.193% |
| 혈액 및 조혈기관질환과 면역메커니즘을 침범하는 특정장애  (D50-D89) | 0.096% | 0.161% | 0.153% | 0.214% |
| 내분비, 영양 및 대사 질환 (E00-E88) | 1.253% | 3.138% | 4.664% | 3.864% |
| 정신 및 행동장애 (F01-F99) | 0.663% | 1.764% | 1.446% | 1.490% |
| 신경계통의 질환 (G00-G98) | 0.966% | 0.943% | 1.476% | 2.870% |
| 눈 및 눈부속기의 질환 (H00-H57) | 0.002% | 0.004% | 0.000% | 0.000% |
| 귀 및 유돌의 질환 (H60-H93) | 0.004% | 0.003% | 0.001% | 0.001% |
| 순환계통 질환 (I00-I99) | 26.307% | 22.487% | 20.221% | 18.745% |
| 호흡계통의 질환 (J00-J98, U04) | 3.787% | 4.616% | 6.238% | 10.309% |
| 소화계통의 질환 (K00-K92) | 9.737% | 9.753% | 6.046% | 4.938% |
| 피부 및 피하조직의 질환 (L00-L98) | 0.014% | 0.055% | 0.112% | 0.121% |
| 근골격계통 및 결합 조직의 질환 (M00-M99) | 0.254% | 0.480% | 0.307% | 0.334% |
| 비뇨생식계통의 질환 (N00-N98) | 0.732% | 0.876% | 1.134% | 2.005% |
| 임신, 출산 및 산후기 (O00-O99) | 0.000% | 0.000% | 0.000% | 0.000% |
| 출생전후기에 기원한 특정병태 (P00-P96) | 0.028% | 0.114% | 0.353% | 0.235% |
| 선천 기형, 변형 및 염색체 이상 (Q00-Q99) | 0.366% | 0.407% | 0.225% | 0.148% |
| 달리 분류되지 않은 증상, 징후 (R00-R99) | 24.617% | 10.371% | 8.657% | 7.013% |
| 질병이환 및 사망의 외인 (V01-Y89) | 12.705% | 18.304% | 15.620% | 13.026% |

1985 / 1995 / 2005 / 2015년도의 각 10년마다 전(全) 연령 남성의 사망원인을 분석해보면 크게 3가지의 특징이 나타납니다.

◼ 신생물의 경우, 연도가 최신에 가까울수록 사망원인비중이 크게 증가합니다. (14% ▷ 32%)

◼ ‘호흡계통의 질환’ 또한 연도가 최신에 가까울수록 사망원인비중이 증가함을 알 수 있으며, 순위도 상위 5개 항목에 포함되었습니다. (4% ▷ 10%)

◼ ‘달리 분류되지 않은 증상 및 증후’의 경우, 최신에 가까울수록 사망원인비중이 크게 감소함을 알 수 있습니다. (24% ▷ 7%)

[표 - 11] 여성 전(全)연령 시대별 사망원인 분석

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 전(全)연령 여성 | 1985년 | 1995년 | 2005년 | 2015년 |
| 특정 감염성 및 기생충성 질환 (A00-B99) | 2.887% | 1.509% | 1.935% | 2.822% |
| 신생물 (C00-D48) | 11.800% | 17.466% | 22.242% | 23.800% |
| 혈액 및 조혈기관질환과 면역메커니즘을 침범하는 특정장애 (D50-D89) | 0.166% | 0.184% | 0.213% | 0.321% |
| 내분비, 영양 및 대사 질환 (E00-E88) | 1.284% | 4.026% | 5.584% | 4.660% |
| 정신 및 행동장애 (F01-F99) | 0.452% | 2.681% | 2.334% | 2.477% |
| 신경계통의 질환 (G00-G98) | 1.069% | 0.926% | 1.945% | 5.114% |
| 눈 및 눈부속기의 질환 (H00-H57) | 0.005% | 0.011% | 0.000% | 0.000% |
| 귀 및 유돌의 질환 (H60-H93) | 0.002% | 0.002% | 0.002% | 0.002% |
| 순환계통 질환 (I00-I99) | 28.884% | 30.779% | 26.555% | 24.988% |
| 호흡계통의 질환 (J00-J98, U04) | 4.262% | 4.533% | 5.308% | 9.806% |
| 소화계통의 질환 (K00-K92) | 4.923% | 4.293% | 2.814% | 3.413% |
| 피부 및 피하조직의 질환 (L00-L98) | 0.014% | 0.117% | 0.232% | 0.230% |
| 근골격계통 및 결합 조직의 질환 (M00-M99) | 0.417% | 1.349% | 1.010% | 0.789% |
| 비뇨생식계통의 질환 (N00-N98) | 0.909% | 0.983% | 1.409% | 2.814% |
| 임신, 출산 및 산후기 (O00-O99) | 0.127% | 0.085% | 0.050% | 0.033% |
| 출생전후기에 기원한 특정병태 (P00-P96) | 0.041% | 0.093% | 0.387% | 0.222% |
| 선천 기형, 변형 및 염색체 이상 (Q00-Q99) | 0.411% | 0.430% | 0.277% | 0.167% |
| 달리 분류되지 않은 증상, 징후 (R00-R99) | 36.002% | 21.520% | 18.675% | 11.022% |
| 질병이환 및 사망의 외인 (V01-Y89) | 6.347% | 9.012% | 9.027% | 7.321% |

1985 / 1995 / 2005 / 2015년도의 각 10년마다 전체 연령 여성의 사망원인을 분석해보면 크게 3가지의 특징이 나타납니다.

◼ ‘신생물’의 경우, 30년 동안 사망원인비중이 약 두 배 증가했습니다. (12% ▷ 24%)

◼ ‘달리 분류되지 않은 증상 및 증후’의 경우, 최근에 가까울수록 사망원인비중이 크게 감소했습니다. (36% ▷ 11%)

◼ 2015년 최근에 들어 ‘질병이환 및 사망의 외인’보다 ‘호흡계통의질환’이 사망원인으로서 비중이 커졌습니다.

[표 – 12] 시대별 남녀 전(全)연령 사망원인조사 순위

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 전(全)연령 | 1985년 | | 1995년 | | 2005년 | | 2015년 | |
| 남 | 여 | 남 | 여 | 남 | 여 | 남 | 여 |
| 특정 감염성 및 기생충성 질환 (A00-B99) | 6 | 7 | 8 | 9 | 8 | 10 | 9 | 9 |
| 신생물 (C00-D48) | 3 | 3 | 1 | 3 | 1 | 2 | 1 | 2 |
| 혈액 및 조혈기관질환과 면역메커니즘을 침범하는 특정장애 (D50-D89) | 14 | 14 | 14 | 14 | 15 | 16 | 14 | 13 |
| 내분비, 영양 및 대사 질환 (E00-E88) | 8 | 8 | 7 | 7 | 7 | 5 | 7 | 7 |
| 정신 및 행동장애 (F01-F99) | 11 | 11 | 9 | 8 | 10 | 8 | 11 | 11 |
| 신경계통의 질환 (G00-G98) | 9 | 9 | 10 | 12 | 9 | 9 | 8 | 6 |
| 눈 및 눈부속기의 질환 (H00-H57) | 18 | 18 | 17 | 18 | 18 | 19 | 18 | 19 |
| 귀 및 유돌의 질환 (H60-H93) | 17 | 19 | 18 | 19 | 17 | 18 | 17 | 18 |
| 순환계통 질환 (I00-I99) | 1 | 2 | 2 | 1 | 2 | 1 | 2 | 1 |
| 호흡계통의 질환 (J00-J98,U04) | 7 | 6 | 6 | 5 | 5 | 6 | 4 | 4 |
| 소화계통의 질환 (K00-K92) | 5 | 5 | 5 | 6 | 6 | 7 | 6 | 8 |
| 피부 및 피하조직의 질환 (L00-L98) | 16 | 17 | 16 | 15 | 16 | 15 | 16 | 14 |
| 근골격계통 및 결합 조직의 질환 (M00-M99) | 13 | 12 | 12 | 10 | 13 | 12 | 12 | 12 |
| 비뇨생식계통의 질환 (N00-N98) | 10 | 10 | 11 | 11 | 11 | 11 | 10 | 10 |
| 임신, 출산 및 산후기 (O00-O99) | 19 | 15 | 19 | 17 | 18 | 17 | 18 | 17 |
| 출생전후기에 기원한 특정병태 (P00-P96) | 15 | 16 | 15 | 16 | 12 | 13 | 13 | 15 |
| 선천 기형, 변형 및 염색체 이상 (Q00-Q99) | 12 | 13 | 13 | 13 | 14 | 14 | 15 | 16 |
| 달리 분류되지 않은 증상, 징후 (R00-R99) | 2 | 1 | 4 | 2 | 4 | 3 | 5 | 3 |
| 질병이환 및 사망의 외인 (V01-Y89) | 4 | 4 | 3 | 4 | 3 | 4 | 3 | 5 |

1983년 ~ 2018년도까지의 데이터를 사용하면 두 가지 문제점이 발생합니다.

첫째로, [표 – 10, 11, 12]를 보면, 시대별로 사망원인비중이 달라지고 있습니다. 가장 눈에 띄는 ‘달리 분류되지 않는 증상’ 및 ‘신생물’은 과거 1985년과 2015년의 사망원인 순위와 비율에 차이가 큼을 알 수 있습니다. 우리가 앞으로 사용할 모델은 현재의 사망원인을 분석하여 그 분석 값을 토대로 건강나이를 산출하는 것이므로, 현재 모델 설정과 맞지 않는 오래된 데이터는 제거하고자 합니다.

둘째, 앞으로 사용할 통계청 마이크로 데이터의 연령대 조사 방식이, 이는 2000년 이전과 2000년 이후가 달라집니다. 2000년 이전에는 90세 이상 인구가 80세 이상 인구로 통합되어 있지만, 2000년 이후에는 80대와 90세 이상 인구가 분류되어 있습니다. 이는 2000년 이전과 2000년 이후 데이터를 합쳐서 사망원인 나이를 분석하게 되면, 분류되는 나이에 차이가 생김을 의미합니다.

따라서, 조사범위 변경 전 데이터인 2000년 이전의 데이터를 삭제합니다.

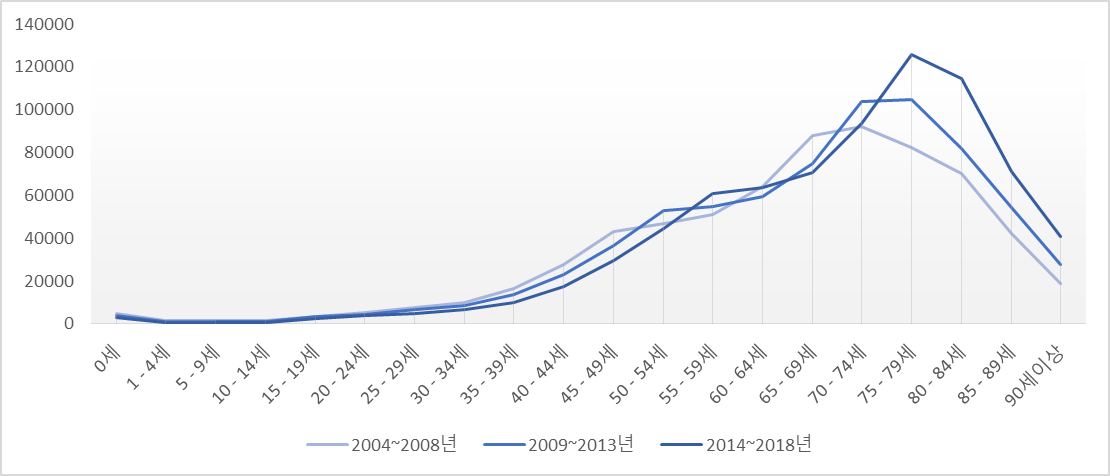
② 2004년 ~ 2018년[[11]](#footnote-11)

[표 - 13] 5년 단위 사망 통계 요약(summary)

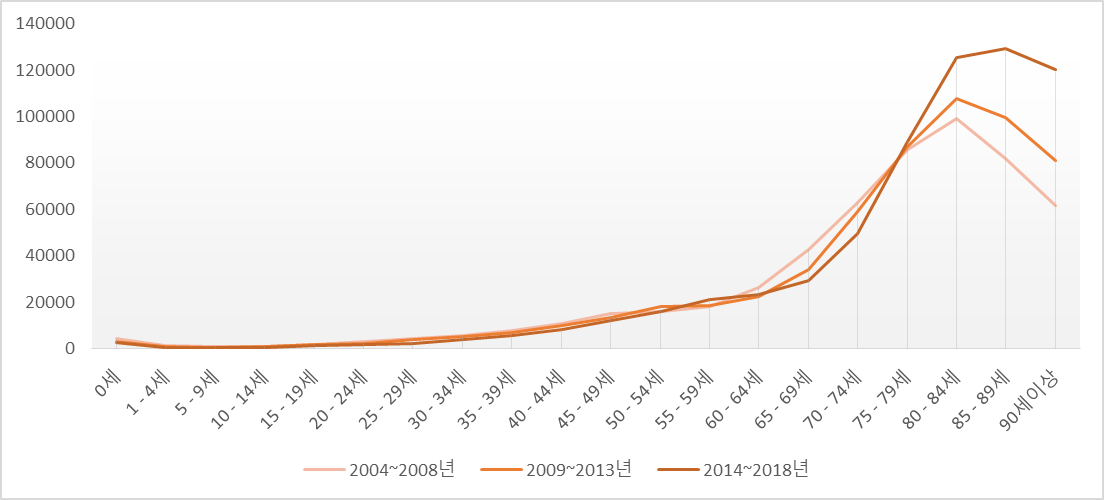
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 2004~2008 | | 2009~2013 | | 2014~2018 | |
| 성별 | 남 | 녀 | 남 | 녀 | 남 | 녀 |
| Mean | 67.0 | 75.0 | 69.1 | 76.8 | 71.4 | 79.0 |
| Std | 15.8 | 15.9 | 15.2 | 15.1 | 14.4 | 14.2 |
| 25% | 58.0 | 69.0 | 60.0 | 71.0 | 63.0 | 74.0 |
| 50% | 69.0 | 78.0 | 72.0 | 80.0 | 74.0 | 82.0 |
| 75% | 78.0 | 85.0 | 80.0 | 87.0 | 82.0 | 88.0 |

\*수치는 소수점 둘째 자리에서 반올림 한 값을 사용

[그림 - 02] 5년 단위 남성의 연령별 사망자수



[그림 - 03] 5년 단위 여성의 연령별 사망자수



[표 – 13], [그림 – 02, 03]를 통해 도출 가능한 특징은 다음과 같습니다.

◼ [그림 – 02]를 보면 전체적으로 시간이 흐름에 따라 그래프가 우 편향되는 경향을 보이며, 세 그래프의 양상이 동일합니다. 일례로, 사망자 수가 최대로 나타나는 구간을 자세히 살펴보면, 그 연령대 범위가 뒤로 1단위씩 밀리고 있습니다.

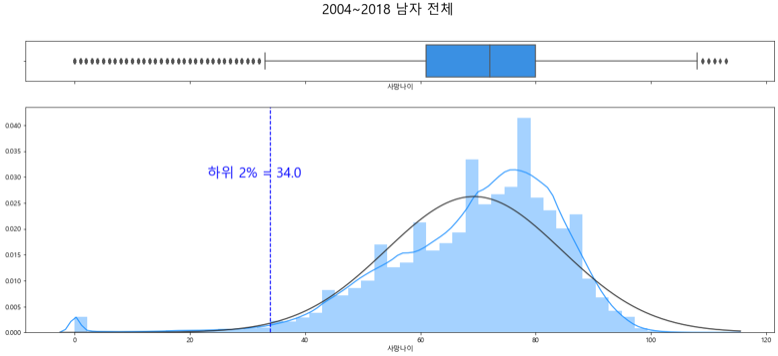
◼ [표 - 13]에서, 남성과 여성의 평균 나이 값이 약 8살 차이로 그 값이 시대가 지나도 일정합니다.

◼ [표 – 13]에서, 남성과 여성 모두 5년 단위 데이터의 평균이 약 2살씩 증가함을 알 수 있습니다. 또한, 남성과 여성의 표준편차가 시대가 지나면서 점점 줄어들고 있습니다. [그림 - 02]와 [그림 – 03]에서도 볼 수 있듯이, 시간이 지남에 따라 사망 연령대가 평균에 밀집되고, 중년층 연령의 사망자 수가 감소하면서 고령자 사망이 증가했습니다.

**4) 이상치(outlier) 제거**

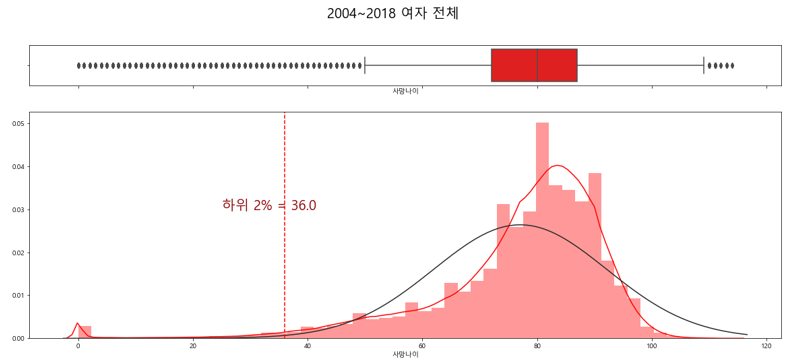
우리는 2004년 ~ 2018년 전체 남녀별 사망 나이 분포를 알아보려 합니다. 다음은 데이터를 정규화[[12]](#footnote-12)시킨 모습과, 이상치 비율은 어느 정도 되는지 나타낸 그래프 입니다.

[그림 - 04] 2004년 ~ 2018년 남성 전체의 boxplot과 분포표



남성 그래프 boxplot을 보면, 약 30세 이하 및 110세 이상의 사망자 수는 그 값이 너무 적어 이상치(outlier)로 나타납니다. 따라서, 정확한 분포를 보기 위해 하위 2%에 해당하는 34세 미만의 데이터는 제거합니다.

[그림 - 05] 2004년 ~ 2018년 여성 전체의 boxplot과 분포표



마찬가지로, 여성 그래프 boxplot에서도 초저(低)연령 및 초고령자의 데이터는 이상치(outlier)로 나타나기 때문에, 하위 2%에 해당하는 36세 미만의 데이터는 제거합니다.

결과적으로, 우리는 사망연령 하위 2%를 제거하고자 합니다. 또한, 보험업계에서 기초요율을 산정 시 최근 5개년 자료를 바탕으로 산출하므로 최근 5개년 데이터만 사용하고자 합니다.

**5) 자료의 신뢰성 검토**

우리가 가지고 있는 데이터는 아무리 많아도 표본집단이기 때문에 모집단을 대표할 만큼 정확한지 확인해야 합니다. 위에서 사용한 데이터의 표본 크기가 적당한지 분석하겠습니다. 분석 방법으로는 가설 검정을 채택하였습니다.

귀무가설 H0 = 정규분포이다.

대립가설 H1 = 정규분포가 아니다.

신뢰도 95%에서 검정 시 기각역 α가 0.05인데, p-value가 0.05 이상이면 귀무가설(H0)을 기각 할 수 없습니다. 남성의 경우 p-value=0.4608, 여성의 경우 p-value=0.2050의 값이 나왔고 따라서 귀무가설을 기각할 수 없습니다. 그러므로 데이터가 정규분포를 따르고, 신뢰할 수 있습니다.

**(2) 세부 사망원인 순위**

앞서 언급한 전처리 과정을 통한 데이터를 사용하여 사망원인을 소분류(103항목)로 재확인 해보았습니다.

[표 - 14] 2014년 ~ 2018년 30세이상 남성의 사망원인 별 사망자수 [[13]](#footnote-13)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 사망원인별(103항목) | 사망자 수(명) | 사망자 비율 | 누적사망자 비율 |
| 1 | 폐암 | 65,194 | 9.878% | 9.878% |
| 2 | 뇌혈관 질환 | 56,449 | 8.553% | 18.430% |
| 3 | 노쇠 | 54,047 | 8.189% | 26.619% |
| 4 | 폐렴 | 44,314 | 6.714% | 33.333% |
| 5 | 간암 | 40,727 | 6.171% | 39.504% |
| 6 | 허혈성 심장질환 | 38,849 | 5.886% | 45.390% |
| 7 | 기타 심장질환 | 32,788 | 4.968% | 50.358% |
| 8 | 위암 | 26,729 | 4.050% | 54.408% |
| 9 | 간질환 | 25,494 | 3.863% | 58.270% |
| 10 | 다발성 골수종 및 악성형질세포 악성신생물 | 25,294 | 3.832% | 62.103% |
| 11 | 당뇨병 | 24,275 | 3.678% | 65.781% |
| 12 | 결장,직장 및 항문의 악성신생물 | 24,048 | 3.644% | 69.424% |
| 13 | 만성 하기도질환 | 22,036 | 3.339% | 72.763% |
| 14 | 나머지 호흡계통 질환 | 16,912 | 2.562% | 75.325% |
| 15 | 나머지 비뇨생식계통 질환 | 15,164 | 2.298% | 77.623% |
| 16 | 췌장암 | 14,703 | 2.228% | 79.850% |
| 17 | 나머지 신경계통 질환 | 14,251 | 2.159% | 82.010% |
| 18 | 나머지 소화계통 질환 | 10,464 | 1.585% | 83.595% |
| 19 | 전립선암 | 8,922 | 1.352% | 84.947% |
| 20 | 고혈압성 질환 | 8,291 | 1.256% | 86.203% |
| 21 | 패혈증 | 7,406 | 1.122% | 87.325% |
| 22 | 알츠하이머병 | 7,212 | 1.093% | 88.418% |
| 23 | 나머지 정신 및 행동장애 | 7,078 | 1.072% | 89.490% |
| 24 | 식도암 | 6,787 | 1.028% | 90.519% |
| 25 | 호흡기 결핵 | 6,041 | 0.915% | 91.434% |
| 26 | 방광암 | 5,149 | 0.780% | 92.214% |
| 27 | 비호지킨 림프종 | 5,081 | 0.770% | 92.984% |
| 28 | 백혈병 | 4,565 | 0.692% | 93.675% |
| 29 | 구강 및 인두의 악성신생물 | 4,354 | 0.660% | 94.335% |
| 30 | 정신활성물질 사용에의한 정신 및 행동장애 | 3,867 | 0.586% | 94.921% |

[표 - 15] 2014년 ~ 2018년 30세이상 여성의 사망원인 별 사망자수[[14]](#footnote-14)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 사망원인별(103항목) | 사망자수 | 사망자 비율 | 누적사망자 비율 |
| 1 | 노쇠 | 70950 | 12.177% | 12.177% |
| 2 | 뇌혈관 질환 | 60199 | 10.332% | 22.510% |
| 3 | 폐렴 | 41201 | 7.072% | 29.581% |
| 4 | 기타 심장질환 | 41155 | 7.064% | 36.645% |
| 5 | 허혈성 심장질환 | 33092 | 5.680% | 42.325% |
| 6 | 당뇨병 | 24185 | 4.151% | 46.476% |
| 7 | 폐암 | 22885 | 3.928% | 50.403% |
| 8 | 다발성 골수종 및 악성형질세포 악성신생물 | 21338 | 3.662% | 54.066% |
| 9 | 고혈압성 질환 | 19042 | 3.268% | 57.334% |
| 10 | 알츠하이머병 | 18209 | 3.125% | 60.459% |
| 11 | 결장, 직장 및 항문의 악성신생물 | 18010 | 3.091% | 63.550% |
| 12 | 나머지 비뇨생식계통 질환 | 17241 | 2.959% | 66.510% |
| 13 | 나머지 정신 및 행동장애 | 14292 | 2.453% | 68.963% |
| 14 | 나머지 신경계통 질환 | 14037 | 2.409% | 71.372% |
| 15 | 간암 | 14037 | 2.409% | 73.781% |
| 16 | 위암 | 13916 | 2.388% | 76.170% |
| 17 | 췌장암 | 12971 | 2.226% | 78.396% |
| 18 | 만성 하기도 질환 | 12946 | 2.222% | 80.618% |
| 19 | 나머지 소화계통 질환 | 12418 | 2.131% | 82.749% |
| 20 | 나머지 호흡계통 질환 | 11861 | 2.036% | 84.785% |
| 21 | 유방암 | 10669 | 1.831% | 86.616% |
| 22 | 패혈증 | 10210 | 1.752% | 88.368% |
| 23 | 간질환 | 7782 | 1.336% | 89.704% |
| 24 | 난소암 | 5206 | 0.894% | 90.598% |
| 25 | 골격계통 및 결합조직의 질환 | 4760 | 0.817% | 91.415% |
| 26 | 자궁경부암 | 4097 | 0.703% | 92.118% |
| 27 | 비호지킨 림프종 | 3615 | 0.620% | 92.738% |
| 28 | 호흡기 결핵 | 3388 | 0.581% | 93.320% |
| 29 | 나머지 순환계통질환 | 3294 | 0.565% | 93.885% |
| 30 | 백혈병 | 3291 | 0.565% | 94.450% |

위 [표 – 14]와 [표 – 15]는 최근 사망자들의 소분류 사망원인 순위입니다.

이를 통하여 확인할 수 있는 특징은

◼ 소분류 사망원인 103항목 중 95%의 해당하는 사람들은 상위 30개 항목으로 사망합니다[[15]](#footnote-15). 이는 사망원인이 일정 질병에 치우쳐 있다는 것으로 생각할 수 있습니다.

◼ 상위 30개의 항목을 분석해보면 절반이 암 및 나머지 질환[[16]](#footnote-16)으로 사망하고 있습니다.

◼ 암 및 나머지 질환을 제외한 질병목록 중 장기 별로 분류해보면 폐, 간, 심장, 혈관 크게 4가지 정도로 분류됩니다.

결과적으로, 현재 살아있는 사람들 또한 상위 30가지(장기로 보면 5가지정도)에 속하는 질병으로 사망할 확률이 아주 높습니다. 따라서, 이에 해당하는 질병을 보험나이 산출 시 적용하여 새롭게 보험나이를 재산출 할 필요가 있습니다.

모든 사람이 모든 질병에 대해 진단을 받는다면 가장 정확한 지표 산출이 가능하지만, 보험계약자는 단지 보험료 할인을 받기 위해 큰 비용이 드는 검진을 받지 않습니다. 따라서, 보험나이 산출하기 위해 국가건강검진 진단 시 나타나는 기본검사 데이터를 사용하려 합니다.

사망자들의 사망 원인 통계와 국가 건강 검진 데이터를 종합해보면,

◼ 폐 – 흡연상태

◼ 간 – 감마지티피, ALT, AST, 음주여부

◼ 심장 – 수축기 혈압, 이완기 혈압, 혈색소

◼ 혈관 – 총 콜레스테롤, 트리글리세라이드, HDL, LDL, 혈청 크레아니틴, 공복혈당, 요단백

◼ 식◦운동 습관 – BMI, 허리둘레

등의 지표로 장기 별 건강을 체크할 수 있습니다.

**2. 본격적인 건강나이 산출**

**(1) 실제 데이터 소개 (건강검진 데이터)[[17]](#footnote-17)**

이 데이터는 앞으로 건강 나이를 산출하는데 있어 사용할 데이터입니다.

◼ '신장(5Cm단위)'과 '체중(5Kg단위)'의 각각의 값은 의미가 없기 때문에, BMI 지수[[18]](#footnote-18)로 합산하여 나타냈습니다.

◼ '기준년도','가입자일련번호','시도코드','시력(좌)','시력(우)','청력(좌)','청력(우)','구강검진 수검여부','치아우식증유무','치석유무','데이터 기준일자' 항목의 값은‘(2) 세부 사망원인 순위’의 결과에 의해 건강 나이와 거리가 멀기 때문에 제거했습니다.

◼ 결측치(NaN)가 존재하는 행을 제거하였습니다.

◼ 마찬가지로 남성과 여성 데이터를 분리하였습니다.

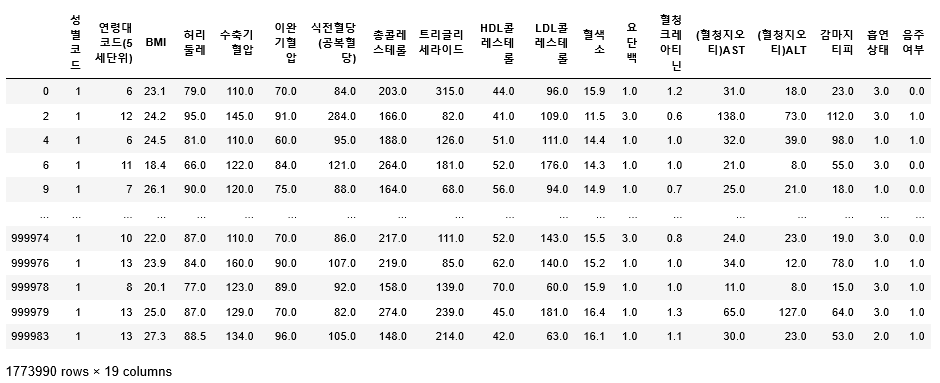
데이터의 연령대 코드는 5세 단위로 나타나 있습니다. 이는 ‘국민건강정보데이터 건강검진정보 사용자 매뉴얼\_20171027.hwp’에서 제공하는 지표를 따릅니다. 이를 참고하여,

◼ 이상치(outlier)로 판단했던 4그룹 이하에 해당하는 데이터를 제거하였습니다.

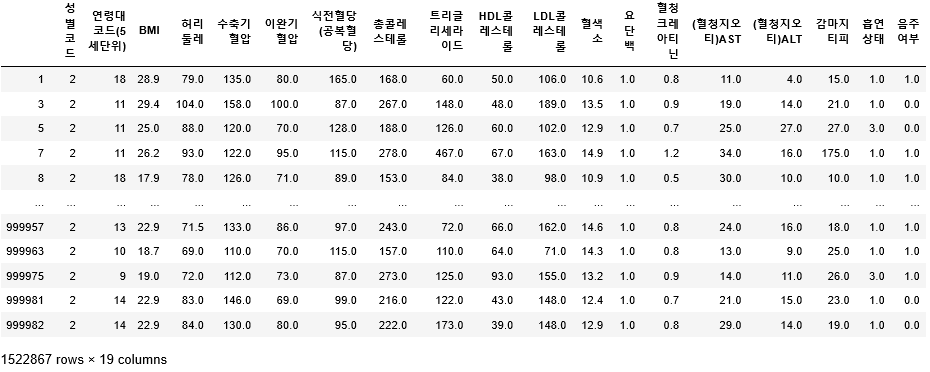
[표 - 16] 연령대 그룹 구분

|  |  |  |  |
| --- | --- | --- | --- |
| **그룹** | **연령대** | **그룹** | **연령대** |
| 1 | 20~24세 | 8 | 55~59세 |
| 2 | 25~29세 | 9 | 60~64세 |
| 3 | 30~34세 | 10 | 65~69세 |
| 4 | 35~39세 | 11 | 70~74세 |
| 5 | 40~44세 | 12 | 75~79세 |
| 6 | 45~49세 | 13 | 80~84세 |
| 7 | 50~54세 | 14 | 85세+ |

[표 - 17] 예시 – 2014~2018년도 남성 데이터



[표 - 18] 예시 – 2014~2018년도 여성 데이터



**(2) 분석 방법론 소개**

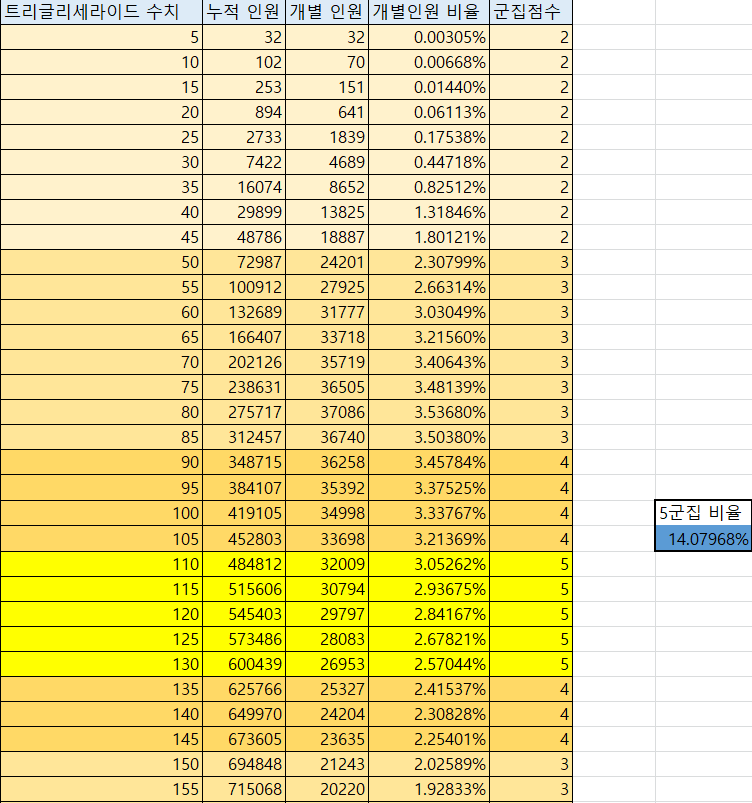
건강나이를 산출하기 위한 분석 방법론으로 ‘군집화(clustering)방법’과 ‘회귀분석(Regression) 방법’을 사용하여 예측합니다.

**1) 군집 분석(clustering analysis)**

군집분석(clustering analysis)이란, 주어진 데이터들의 특성을 고려해 데이터 집단을 정의하고, 데이터 집단의 대표할 수 있는 대표점을 찾는 방법입니다.

우선, 각각의 독립변수 17개를 개별적으로 군집화 시킵니다.

[그림 - 06] 예시-독립변수(트리글리세라이드)의 군집점수



이때, 군집 개수(K) = 5개로 설정하였으며

1점 = 건강상태가 매우 좋지 않음 or 데이터 값이 이상함.

2점 = 건강상태가 좋지 않음

3점 = 정상범위를 조금 초과하는 상태를 보임.

4점 = 정상범위 속하지만 상위15%의 건강상태를 보이지 못함.

5점 = 정상범위 속하면서 상위15%[[19]](#footnote-19)의 건강상태를 보임

으로 군집을 분류하였습니다.

**2) 회귀 분석(regression analysis)**

통계학에서 회귀 분석(regression analysis)이란, 관찰된 연속형 독립변수들 X에 대한 종속변수 Y 모형을 구한 뒤 적합도를 측정해 내는 분석 방법입니다. 우리는 종속 변수 Y를 ‘연령대코드(5세단위)’로 두고 나머지 17항목들을 독립변수 X로 둔 다음, X 값이 변할 때 Y 값의 변화를 알아보고자 합니다. 머신러닝을 이용해 다양한 회귀 분석 방법을 시도해 보고, 가장 정확한 분석 방법을 최종 모델로 선택합니다.

대표적인 회귀 모형 평가 지표인 RMSE(Root Mean Square Error)[[20]](#footnote-20)를 각각의 모델에서 구하고, 그 값을 비교하여 어떤 모델이 가장 정확한지 살펴봅니다. 같은 크기의 데이터를 사용했을 때, RMSE 값이 작을수록 정확한 모델입니다.

**① 선형 회귀(linear regression)**

선형 회귀는 종속 변수 y와 한 개 이상의 독립변수 x의 선형 상관 관계를 모델링 하는 회귀 분석 기법입니다. 건강검진 데이터는 총 17개의 독립변수를 가지고 있으므로 다중 선형 회귀에 속합니다.

[그림 - 07] 선형 회귀 모델링 코드

|  |  |
| --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17 | from sklearn.linear\_model import LinearRegression    lin\_reg = LinearRegression()  lin\_reg.fit(manAlldata\_prepared, manAlldata\_labels) #남성 데이터 fitting  lin\_reg\_2 = LinearRegression()  lin\_reg\_2.fit(womanAlldata\_prepared, womanAlldata\_labels) #여성 데이터 fitting    lin\_man\_predictions = lin\_reg.predict(manAlldata\_prepared)  lin\_mse = mean\_squared\_error(manAlldata\_labels, lin\_man\_predictions)  lin\_rmse = np.sqrt(lin\_mse)  lin\_woman\_predictions = lin\_reg.predict(womanAlldata\_prepared)  lin\_mse\_2 = mean\_squared\_error(womanAlldata\_labels, lin\_woman\_predictions)  lin\_rmse\_2 = np.sqrt(lin\_mse\_2)    #모델 성능  lin\_rmse -> 2.4313537153127744  lin\_rmse\_2 -> 2.8194991345369447 |

**② 다항 회귀(Polynomial Regression)**

우리가 가지고 있는 데이터가 비선형 형태일 수도 있습니다. 이 때 비선형 데이터를 선형 모델을 사용하여 훈련시키는 방법이 다항 회귀 분석입니다. 각 특성의 거듭제곱을 새로운 특성으로 추가하고, 이 확장된 특성을 포함하여 선형모델을 훈련시킵니다.

[그림 - 08] 다중 회귀 모델링 코드

|  |  |  |
| --- | --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22  23  24 | from sklearn.preprocessing import PolynomialFeatures  from sklearn.linear\_model import LinearRegression  from sklearn.metrics import mean\_squared\_error  poly\_features = PolynomialFeatures(degree=2, include\_bias = False)  man\_poly = poly\_features.fit\_transform(manAlldata\_prepared) #남성 데이터 fitting  poly\_features\_2 = PolynomialFeatures(degree=2, include\_bias = False)  woman\_poly = poly\_features\_2.fit\_transform(womanAlldata\_prepared) #여성 데이터 fitting  poly\_reg = LinearRegression()  poly\_reg.fit(man\_poly, manAlldata\_labels)  poly\_reg\_2 = LinearRegression()  poly\_reg\_2.fit(woman\_poly, womanAlldata\_labels)    poly\_man\_predictions = poly\_reg.predict(man\_poly)  poly\_mse = mean\_squared\_error(manAlldata\_labels, poly\_man\_predictions)  poly\_rmse = np.sqrt(poly\_mse)  poly\_woman\_predictions = poly\_reg\_2.predict(woman\_poly)  poly\_mse\_2 = mean\_squared\_error(womanAlldata\_labels, poly\_woman\_predictions)  poly\_rmse\_2 = np.sqrt(poly\_mse\_2)    # 모델 성능  poly\_rmse -> 2.2995827447552215  poly\_rmse\_2  -> 2.2500690345891305 |  |

**③ 그래디언트 부스팅 회귀(Gradient Boosting)**

그래디언트 부스팅 회귀는 부스팅 방법을 적용한 결정 트리(Decision Tree)의 앙상블 입니다. 이전까지의 오차를 보정하도록 예측기를 순차적으로 추가하되, 이전 예측기가 만든 잔여 오차(residual error)에 새로운 예측기를 학습시킵니다. 여기서 부스팅 방법이란, 약한 학습기를 여러 개 연결하여 강한 학습기를 만드는 앙상블 방법입니다. 결정 트리 회귀 모델이란, 설정한 깊이(depth)만큼 루트 노드(깊이가 0인 맨 꼭대기 노드)부터 리프 노드(자식 노드를 가지지 않는 노드)까지의 트리를 만들어 값을 예측하는 모델입니다.

[그림 - 09] 그래디언트 부스팅 모델 코드

|  |  |  |
| --- | --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19 | from sklearn.ensemble import GradientBoostingRegressor  gbrt = GradientBoostingRegressor(max\_depth=4, n\_estimators=200, learning\_rate=0.1)  gbrt.fit(manAlldata\_prepared, manAlldata\_labels)    #남성 데이터 fitting  gbrt\_2 = GradientBoostingRegressor(max\_depth=4, n\_estimators=200, learning\_rate=0.1)  gbrt\_2.fit(womanAlldata\_prepared, womanAlldata\_labels) # 여성 데이터 fitting  gbrt\_predictions = gbrt.predict(manAlldata\_prepared)  gbrt\_mse = mean\_squared\_error(manAlldata\_labels, gbrt\_predictions)  gbrt\_rmse = np.sqrt(gbrt\_mse)  gbrt\_woman\_predictions = gbrt\_2.predict(womanAlldata\_prepared)  gbrt\_mse\_2 = mean\_squared\_error(womanAlldata\_labels, gbrt\_woman\_predictions)  gbrt\_rmse\_2 = np.sqrt(gbrt\_mse\_2)    #모델 성능  gbrt\_rmse -> 2.035533120408944  gbrt\_rmse\_2 -> 1.9389905450750227 |  |

**④ 랜덤 포레스트(Random Forest) 회귀**

랜덤 포레스트는 쉽게 말해, 배깅 방법을 적용한 결정 트리(Decision Tree)의 앙상블 입니다. 배깅 방법이란, 한 예측기에 적용할 훈련세트의 서브 셋을 무작위로 구성하고, 중복을 허용하여 샘플링 하는 방법입니다. 즉, 랜덤 포레스트는 결정 트리를 예측기로 사용하여 훈련시킨 다음, 모든 예측값들을 모아서 예측 평균을 계산해주는 모델입니다.

[그림 - 10] 랜덤 포레스트 모델 코드

|  |  |  |
| --- | --- | --- |
| 1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20 | from sklearn.ensemble import RandomForestRegressor  from sklearn.metrics import mean\_squared\_error    forest\_reg = RandomForestRegressor()  forest\_reg.fit(manAlldata\_prepared, manAlldata\_labels) #남성 데이터 fitting  forest\_reg\_2 = RandomForestRegressor()  forest\_reg\_2.fit(womanAlldata\_prepared, womanAlldata\_labels) #여성 데이터 fitting  man\_predictions = forest\_reg.predict(manAlldata\_prepared)  forest\_mse = mean\_squared\_error(manAlldata\_labels, man\_predictions)  forest\_rmse = np.sqrt(forest\_mse)    woman\_predictions = forest\_reg\_2.predict(womanAlldata\_prepared)  forest\_mse\_2 = mean\_squared\_error(womanAlldata\_labels, woman\_predictions)  forest\_rmse\_2 = np.sqrt(forest\_mse\_2)    #모델 성능  forest\_rmse -> 0.9333785329612521  forest\_rmse\_2 -> 0.9076896072510086 |  |

결과적으로 모든 모델의 RMSE 값을 비교해보면, 랜덤 포레스트 모델의 값이 가장 작게 나왔습니다. 따라서 다양한 회귀 모델 중 상대적으로 가장 정밀도가 높은 모델은 ‘랜덤 포레스트 모델’이라 할 수 있습니다. 즉, 랜덤 포레스트 회귀 모델로 분석했을 때, 실제 값(실제 나이)과 예측 값(건강 나이) 사이의 오차범위가 터무니없이 크게 벌어질 확률이 다른 모델보다 작다고 할 수 있습니다.

또한, 랜덤 포레스트 방법은 각 특성의 중요도를 구할 수 있습니다. 이를 살펴보면 다음과 같이 나타납니다.

[표 - 19] 각 특성의 중요도[[21]](#footnote-21)

|  |  |  |
| --- | --- | --- |
|  | 남성 | 여성 |
| BMI | 0.10914351685070237 | 0.07920338108331867 |
| 허리둘레 | 0.06528447013628211 | 0.1591478578892926 |
| 수축기 혈압 | 0.06595647709129956 | 0.12782126889370643 |
| 이완기 혈압 | 0.047660765733508255 | 0.0456744906380294 |
| 식전혈당(공복혈당) | 0.09896576308195908 | 0.0638502601344397 |
| 총콜레스테롤 | 0.055755849407824166 | 0.04340889815787169 |
| 트리글리세라이드 | 0.06006225755627357 | 0.06951973982318624 |
| HDL 콜레스테롤 | 0.049932576560960534 | 0.04858773973306012 |
| LDL 콜레스테롤 | 0.05064875159042984 | 0.06044892007305266 |
| 혈색소 | 0.14201169929959753 | 0.05321276261262968 |
| 요단백 | 0.003877564318490201 | 0.003990437525974998 |
| 혈청 크레아티닌 | 0.031847254533663695 | 0.031312917285435025 |
| (혈청지오티)AST | 0.054523199942784276 | 0.1101433242583151 |
| (혈청지오티)ALT | 0.06288126998348063 | 0.04735605279753724 |
| 감마지티피 | 0.05328109335412249 | 0.04111768076801544 |
| 흡연 상태 | 0.05355820394705415 | 0.007731582595259777 |
| 음주 여부 | 0.00747812401193811 | 0.007472685730875337 |

이 결과를 보면 건강 나이를 산출 시 가장 중요하게 작용하는 5 항목이 남성과 여성이 차이가 있음을 알 수 있습니다. 우선, 남성의 건강에 중요하게 작용하는 항목은 혈색소, BMI, 식전혈당, 수축기 혈압, (혈청지오티)ALT 입니다. 반면, 여성의 건강에 중요하게 작용하는 항목은 허리 둘레, 수축기 혈압, BMI, (혈청지오티)AST, 트리글리세라이드 입니다.

**⑤ 보험료 산출 방법**

현재 보험 나이는 만 나이 기준으로 사용합니다. 위의 결과를 바탕으로 보험 나이를 재산출하는 방법은 다음과 같습니다.

보험나이 = 만 나이 – {(만 나이 연령대(5세단위) + (만 나이 – 만 나이 연령대 중 최저 나이) \* 0.2) - 건강나이 연령대(5세 단위)} \* 5

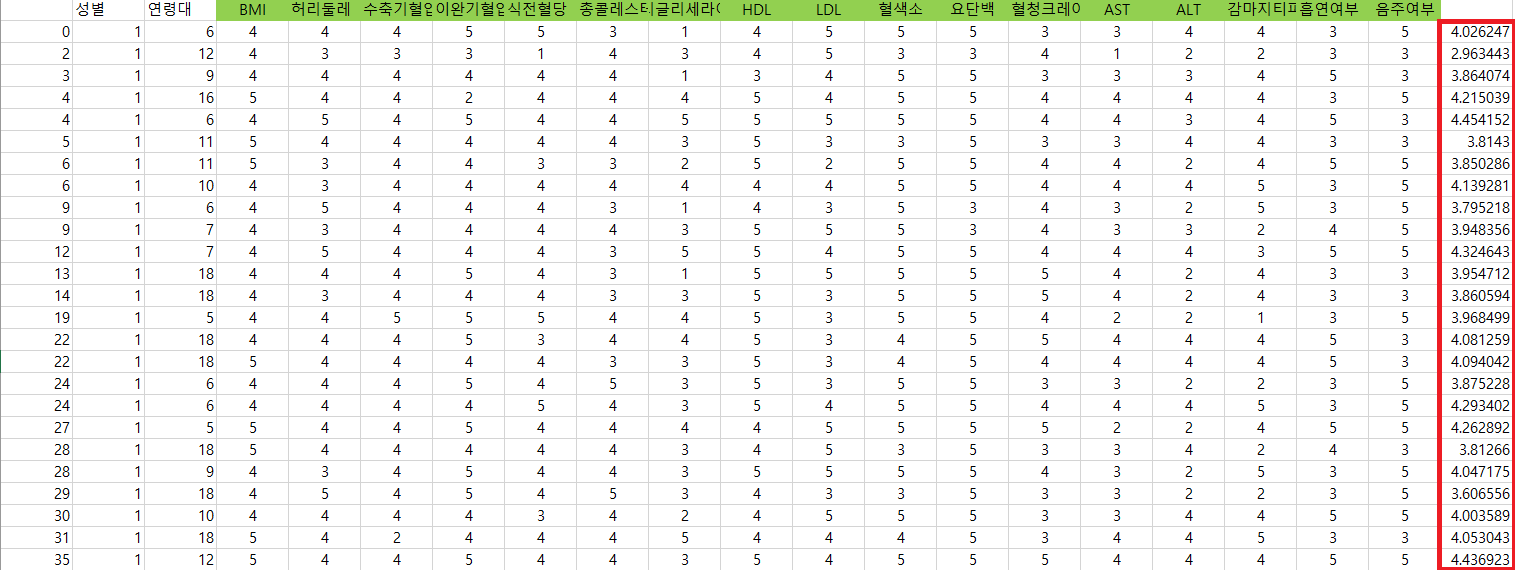
이 식을 통해 보험 나이를 재산출 합니다.

**(3) 실제 예시를 통해 예측하기**

**1) 군집화 모델로 예측**

사망보험가입자가 보험료 할인을 받기 위한 목적으로 보험회사에 자신이 받은 건강검진표를 제출하게 되면, 이를 개별 군집점수로 환산하여 총 군집점수를 산출합니다.

[그림 - 11] 남성 17항목을 통한 군집화 점수산출



이때, 총 군집점수 = ∑ (각 항목점수 x 각 항목의 가중치)를 통하여 산출하였으며 가중치는 랜덤 포레스트 방식으로 산출한 [표 - 19]를 적용하여 구했습니다. 이를 군집점수 크기에 따라 내림차순으로 정리한 다음 상위 10%[[22]](#footnote-22)를 자릅니다

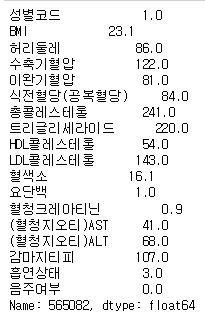
[그림 - 12] 남성 17항목을 통한 군집화 점수 상위자 나열



이를 통하여 상위 5%에 해당하는 초고(高)군집점수 계약자는 보험료 10%를 1년간 할인해주는 제도로써, 차상위 5%에 해당하는 고(高)군집점수 계약자는 보험료 5%를 1년간 할인해주는 제도로써 보험계약자 할인정책을 시행합니다. 또는 직전 1년의 보험료할인액을 환급해주는 시스템으로서 정책을 실시합니다.

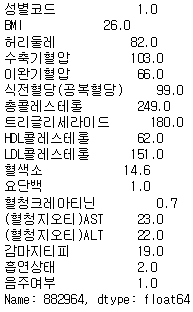
**2) 회귀 모델로 예측**

가장 정밀도가 높았던 랜덤 포레스트 회귀 모델을 이용하여 실제로 건강나이를 예측해보겠습니다. 다음은 그 결과입니다.



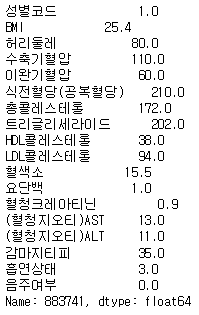
이 남성은 실제 만 나이가 63세입니다. 그러므로 연령대 분류를 하면 9 그룹에 들어갑니다. 9 그룹의 최저 나이는 60세 입니다. 랜덤 포레스트 모델 분석 결과, 예측값(건강 나이 연령대 분류)도 9.0이 나왔습니다. 이때 보험 나이를 계산하면,

63 – {9.0 + (63 - 60)\* 0.2 – 9.0}\* 5 = 60세 입니다.



이 남성은 실제 만 나이가 76세입니다. 그러므로 연령대 분류를 하면 12 그룹에 들어갑니다. 12 그룹의 최저 나이는 75세 입니다. 랜덤 포레스트 모델 분석 결과, 예측값(건강 나이 연령대 분류)는 11.7이 나왔습니다. 이때 보험 나이를 계산하면,

76 – {12.0 + (76 - 75)\* 0.2 – 11.7}\* 5 = 73.5세 입니다.



이 남성은 실제 만 나이가 66세입니다. 그러므로 연령대 분류를 하면 10 그룹에 들어갑니다. 10 그룹의 최저 나이는 65세 입니다. 랜덤 포레스트 모델 분석 결과, 예측값(건강 나이 연령대 분류)는 10.4이 나왔습니다. 이때 보험 나이를 계산하면

69 – {10.0 + (66 - 65)\* 0.2 – 10.4}\* 5 = 70세 입니다

**Ⅳ. 결론**

**1. 보험 계약자가 얻을 수 있는 이점**

**(1) 할인된 가격의 보험료**

[표 – 06] 에서 볼 수 있듯이, 생명 보험의 가입 의향이 없는 가장 큰 요인은 경제적 여유가 없기 때문입니다. 즉, 보험료가 비싸다고 느끼는 사람이 많다는 것입니다. 건강 상태에 따른 보험료 할인 정책을 적용하게 되면, 자신의 건강 상태에 따라 할인 혜택을 받을 수 있는 기회가 주어집니다.

이 논문에서 제안하는 군집 분류 방법을 사용하면, 동일한 나이라도 국민건강검진을 통해 자신의 건강이 보험회사에서 정한 상위 x%에 속하게 되면 보험료를 할인 받을 수 있습니다. 또는 회귀 분석 방법을 사용하면, 보험 나이가 실제 만 나이보다 어리게 나올 시 보험료를 할인 받을 수 있습니다.

[표 - 20] 60세 만기 전기납, 주 계약 1억원 기준

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 삼성생명 (다이렉트 2.0) | 36(세) | 37(세) | 38(세) | 39(세) | 40(세) | 41(세) | 42(세) | 43(세) | 44(세) | 45(세) |
| 28,000 | 29,000 | 30,000 | 32,000 | 33,000 | 35,000 | 36,000 | 38,000 | 39,000 | 40,00 |

예를 들어, 위 [표 – 20] 기준 현재 44세 남성이 가입할 경우, 월 보험료는 약 39,000원 [[23]](#footnote-23)입니다. 만약 이 사람의 건강 연령대가 4.5로 예측될 경우

재 산출된 보험나이가 44 – {5 + (44 – 40) \* 0.2 – 4.5} \* 5 = 37.5세가 되어, 월 보험료가 29,500원으로 감소하게 됩니다.

**(2) 건강 케어 서비스 혜택**

현재 사망 보험을 가입하면 사망이라는 보험사고 발생 시에만 보험금을 지급받고 보험 계약이 종료됩니다. 이 논문에서 제안하는 보험료 할인 정책을 도입하면, 보험회사에서는 보험계약자 개개인의 건강 지표 데이터를 가지게 되고, 이를 통해 건강 나이를 산출합니다. 계약자 입장에서는 보험 회사에게 건강에 대한 정보를 제공받고 이를 통해 자신의 건강 수준이 어느 정도인지 객관적인 지표로 알 수 있습니다. 더 나아가 보험 회사가 헬스 케어 서비스를 제공한다면, 지속적으로 건강 관리를 받을 수 있습니다.

**2. 보험 회사가 얻을 수 있는 이점**

**(1) 유동성 문제 해결 및 투자 수익 극대화**

현행 사망보험상품의 경우 사망을 보험사고로 인식하여 보험계약자에게 보험금을 지급하지만, 실질적으로 보험계약자의 사망확률에 대해서 예측할 수 있는 방법이 없었습니다. 하지만, 보험료할인정책을 통하여 보험계약자에게 개인별 건강 검진 표를 지속적으로 받게 된다면, 논문에서 제안하는 분석 방법론들을 사용하여 보험 계약자의 사망확률을 높은 정확도로 예측할 수 있게 됩니다.

수치적으로 정확하게 나타낸 사망 확률을 이용하면 현금 유동성 시기를 비교적 정확히 예측할 수 있습니다. 그러므로 보험회사의 자본포트폴리오를 설계할 때 유동성 리스크를 감소시키고 투자 수익을 극대화시킬 수 있습니다.

**(2) 자본 듀레이션 연장**

건강하면 오래 산다[[24]](#footnote-24)는 것은 이미 많은 논문을 통해 과학적으로 밝혀졌습니다. 개별 계약자에게 일정조건 만족 시 보험료 할인 정책을 내걸게 되면, 보험계약자는 일정조건을 만족시키기 위해 건강상태를 유지하려고 합니다. 따라서, 실제로 보험계약자의 수명은 일정부분 증가하게 되어 보험회사 입장에서 보험료를 굴릴 수 잇는 자본 듀레이션의 기간이 길어지게 됩니다.

**(3) 다양한 서비스 제공 및 고객 확보**

예를 들어 ‘DB 건강 케어 챗봇’이라는 서비스를 도입한다고 했을 때, 보험 계약자가 자신의 객관적 건강 지표를 알아보기 위해 이 서비스를 사용하게 되고, 그러면 보험회사로의 접근성이 높아집니다. 접근성이 높아지면, 꼭 사망 보험이 필요하지 않은 저 연령층도 잠재적 고객으로의 확보가 가능해집니다. 결과적으로 높은 기업 인지도, 신뢰도와 광고효과의 증가를 얻을 수 있습니다.

**3. 국가적 차원에서 얻을 수 있는 이점**

**(1) 국가 사회 비용 절감**

국가 차원에서 국민 건강 증진을 위해 국가가 부담해야 하는 노력의 일부를 보험회사가 함께 담당하게 되므로 이에 수반되는 사회적 비용이 절감됩니다.

**(2) 국민의 건강 인식 제고**

보험회사가 건강나이 서비스를 제공함으로써, 국민들의 건강에 대한 관심을 높일 수 있고 마찬가지로 이에 부담되는 사회적 비용이 절감됩니다.

**4. 논문의 한계점**

**(1) 데이터 상의 한계점**

[표 – 15]에서 여성의 경우 부인과 질환이 소분류 103항목 중 20위권 안에 있습니다. [표 – 14, 15]에서 암의 경우, 최상위권에 분포되어 있지만 유전적 요인이 크게 작용합니다. 그렇지만 국가 건강 검진 데이터로는 유전적 요소, 부인과 질환 검진 데이터를 확인할 수 없습니다. 따라서 건강나이를 산출 시 적용하지 못하였습니다. 만약 건강 검진 항목에 필수항목으로 부인과 질환을 검사한다면, 여성의 경우 좀 더 정확한 건강 나이를 예측할 수 있을 것입니다. 유전적 데이터는 사기업이 이익 창출의 목적으로 사용이 불가능하기 때문에 법률적으로 허가가 나지 않는 이상 보험회사에서 데이터를 사용하지 못한다는 한계가 존재합니다. 현재로서는 보험 설계사가 대면 판매를 할 시 보험계약자에게 직접 유전질환 여부를 확인 방법 밖에 없습니다.

**(2) 통계적 한계점**

[표 – 10, 11]을 참고하면, 사고로 인한 죽음의 비율이 상위 5항목 안에 있음을 알 수 있습니다. 이 논문에서는 건강나이를 산출 시 사고로 인한 죽음은 건강 나이와 관련이 없기 때문에 전처리 과정에서 데이터를 제거했습니다. 하지만 건강나이가 높은 확률로 예측 가능하더라도, 사고로 인한 죽음의 비율이 20%으로 매우 큽니다. 따라서 건강 나이로 예측한 사망 시기와 예상치 못한 사고로 인한 사망 시기의 차이가 크게 벌어질 수 있습니다. 이를 해결할 수 있는 방안으로는, 건강나이에 대한 기초요울을 산정할 때, 안전할증률(safety margin)을 부과하여 산출하면 됩니다.

**Ⅴ. 참고 문헌**

[1] 핸즈온 머신러닝 - 한빛미디어 오렐리앙 제롱 지음

[2] 공공데이터포털 https://www.data.go.kr/

[3] 국가통계포털 <http://kosis.kr/index/index.do>

[4] 국가통계 마이크로데이터 <https://mdis.kostat.go.kr/index.do>

[5] 위키백과- 클러스터 분석 기법<https://ko.wikipedia.org/wiki/%ED%81%B4%EB%9F%AC%EC%8A%A4%ED%84%B0_%EB%B6%84%EC%84%9D>

[6] 위키백과 - 회귀 분석 기법

[7] 한국보건사회연구원2014-22-9-3 연구보고서

[8] 국민건강정보데이터 건강검진정보 사용자 매뉴얼\_20171027.hwp

[9] kiri 연구보고서 2009-05-02 자동차보험 요율 산출 방법

[10] kiri 연구보고서 2019-11-05 2019년 보험소비자 설문조사

[11] 생명보험 협회 13차, 14차, 15차 생명보험 성향조사

[12] 질병분류 정보센터 http://www.koicd.kr

1. 2019.6월 ~ 11월의 평균계약건수 [↑](#footnote-ref-1)
2. 총 보험 가입 금액이란, 각 보험 계약에서 사망 발생 시, 전액 보상할 경우 지급해야 하는 보험금이라 정의됩니다. [↑](#footnote-ref-2)
3. KB생명의 경우, 사망보험에 치중되어있지 않은 모습을 보여주고 있는데, 이는 기업전략 및 상품의 특성으로 인한 예외외인 기업입니다. [↑](#footnote-ref-3)
4. 보험 설계사가 보험 계약 당시 전반적인 보험 내용을 설명했을 때 확실하게 이해한 정도에 대한 설문조사 [↑](#footnote-ref-4)
5. 일반적인 보험료이며, 기업, 성별, 환경적 요소 및 특약 선택에 따라 보험료는 변동할 수 있습니다. [↑](#footnote-ref-5)
6. 자동차보험료=기본보험료×차등화요소×할인·할증요율(심도)×사고건수요율(빈도) [↑](#footnote-ref-6)
7. 금융감독원 2017.07.11 브리핑자료에서 자동차보험할인할증제도 설명 시 든 예시로 실제 보험료 할인 폭은 회사별 상이할 수 있습니다. [↑](#footnote-ref-7)
8. 앞으로 사용하는 모든 데이터는 남녀 각각 분리하여 분석합니다. [↑](#footnote-ref-8)
9. 1983년이 가장 연한 색이며 시간이 지날수록 색이 진해지도록 설정했습니다. [↑](#footnote-ref-9)
10. 각 시대 별 상위 5 항목을 색칠하였고 순위마다 밝기를 다르게 하였습니다. 색이 진해질수록 사망률 비율이 높습니다. [↑](#footnote-ref-10)
11. 5년 단위의 비교가 용이하도록 가장 최근 데이터(2018년)부터 2004년의 데이터까지를 사용합니다. [↑](#footnote-ref-11)
12. 정규화된 값. [↑](#footnote-ref-12)
13. [표 -14]는 소분류 103항목 중 상위 95%의 사망원인에 해당하는 항목만을 나열하였습니다. [↑](#footnote-ref-13)
14. [표 -15]는 소분류 103항목 중 상위 95%의 사망원인에 해당하는 항목만을 나열하였습니다. [↑](#footnote-ref-14)
15. 남성의 경우 50% = 7항목, 60% = 9항목, 70% = 12항목, 80% = 16항목이며, 여성의 경우 50% = 7항목, 60% = 10항목, 70% = 13항목, 80% = 18항목으로 나타납니다. [↑](#footnote-ref-15)
16. 질병관리본부에서 정확한 질병 코드를 예측할 수 없는 질병을 나머지 질환으로 분류하고 있습니다. [↑](#footnote-ref-16)
17. 데이터에 관한 자세한 설명은 공공 데이터 포털에‘국민건강정보데이터 건강검진정보 사용자 매뉴얼\_20171027.hwp’ 에서 참고할 수 있습니다. [↑](#footnote-ref-17)
18. BMI = [↑](#footnote-ref-18)
19. 상위 15%는 사망보험가입자가 보험료할인을 위하여 제출한 경우 임의의 일정 계약자비율 만큼 보험료를 할인 적용시켜주기 위하여 대입한 임의의 값으로 변경될 수도 있는 수치입니다, [↑](#footnote-ref-19)
20. 추정값 또는 모델이 예측한 값과 실제 관찰되는 값의 차이를 예측하는 척도. RMSE() = = [↑](#footnote-ref-20)
21. 각 특성의 중요도 전체 합은 1 입니다. [↑](#footnote-ref-21)
22. 우리가 보험료 할인 대상을 지정한 가상의 비율이며, 이는 바뀔 수 있습니다. [↑](#footnote-ref-22)
23. 이 계약은 2012년 삼성생명 다이렉트 2.0의 보험료입니다. [↑](#footnote-ref-23)
24. 한국보건사회연구원2014-22-9-3연구보고서 참고 [↑](#footnote-ref-24)