고독사 예측 모델 구축을 위한 데이터마이닝 분석

32230811 김수인

목차

1. 목적 결정

고독사 문제는 점점 더 심각해지고 있지만, 고위험군을 사전에 예측하기 위한 데이터 기반 접근은 부족

따라서 본 프로젝트는 고독사 위험군을 일반인과 구분할 수 있는 예측 모델을 구축하고자 함
실제 통계 기반 확률 샘플링을 통해, 현실적 특성 분포를 반영한 비교 분석을 수행

고독사와 일반인 구분

확률 기반 반복 샘플링→ 현실 통계 반영 + 모델의 일반화 가능성 평가

민감도+불필요한 개입 방지에 초점

3가지 모델(KNN, Naive Bayes, Tree)을 동일한 조건으로 반복 비교
→가장 실용적이고 해석 가능한 예측 모델을 선정하는 것이 최종 목표

2. 데이터 수집

단순히 고독사자와 일반 사망자를 구분하는 데 그치지 않고, 현재 생존해 있는 일반인 중 '누가 고독사의 위험이 높은가'를 사전에 예측하는 것이 목적→ 따라서 비교 대상은 사망자가 아닌, 전체 인구 기반의 일반인 데이터를 설정

데이터 (모두 2021년 데이터)	출처
고독사 발생현황(성별, 거주지, 21대질환 보유자, 수술내역, 기초생활수급자 여부)	보건복지부
「장래인구추계」성 및 연령별 추계인구(1세별, 5세별) (2021)	통계청
「주거실태조사」(일반가구)지역별 소득계층별 주택유형	국토교통부
건강보험심사평가원,「건강보험통계」22대 분류별 진료현황(2021)	국민건강보험공단
특정 질병코드분류별 수술 진료환자수 및 진료건수(2021)	국민건강보험공단
「국민기초생활보장수급자현황」국민기초일반수급자수(2021)	보건복지부

<고독사 발생 현황: 성별·연령별>

연령별(1)	2021			
	계	남자	여자	2) 미상
계	3,378	2,817	529	32
19세이하	2	1	1	0
20대	53	37	16	0
30대	164	120	44	0
40대	526	436	88	2
50대	1,001	900	91	10
60대	981	860	114	7
70대	421	314	104	3
80대이상	203	135	67	1
3) 미상	27	14	4	9

<고독사 발생 현황: 최근 1년간 수술경험>

연령별(1)	2021
	수술경험
19세이하	1
20대	11
30대	25
40대	76
50대	192
60대	194
70대	80
80대이상	33
전체	612

<고독사 발생 현황: 고독사 사망 전 1년간 21대 질환 보유자>

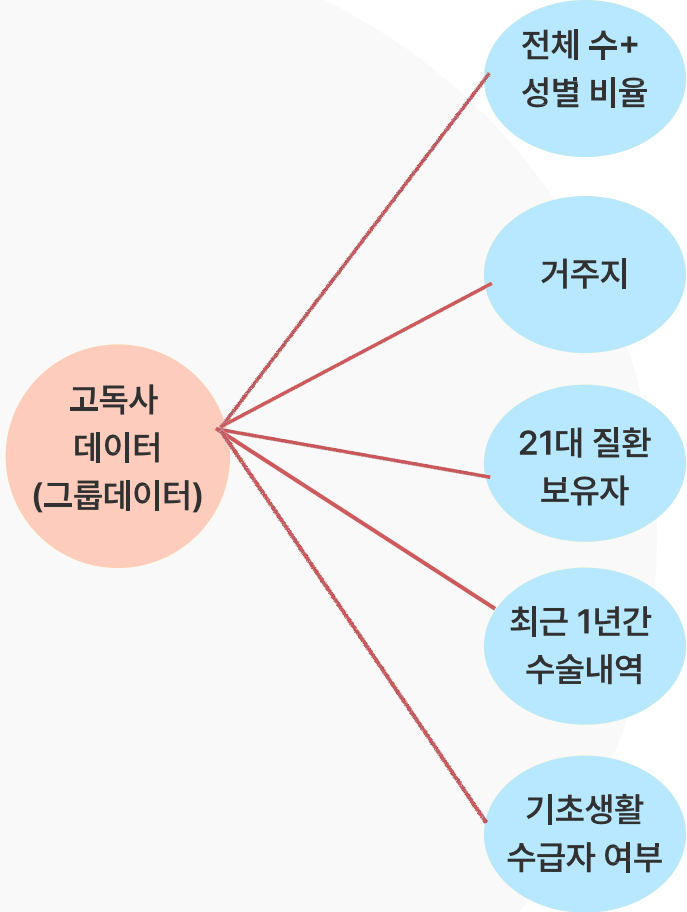
21대 질환	2019	2020	2021
특정감염성 및 기생충성 질환	214	65	211
신생물	156	178	183
혈액 및 조혈기관의 질환과 면역기전을 침범한 특정 장애	27	32	37
내분비, 영양 및 대사질환	478	550	533
정신 및 행동장애	557	617	593
신경계의 질환	255	252	249
눈 및 눈부속기의 질환	341	375	355
귀 및 유양돌기의 질환	124	112	96
순환기계의 질환	699	772	772
호흡기계의 질환	539	560	388
소화기계의 질환	683	714	671
피부 및 피하조직의 질환	371	395	353
근골격계 및 결합조직의 질환	711	810	717
비뇨생식기계의 질환	286	323	283
임신, 출산 및 산욕	0	0	0
주산기에 기원한 특정병태	0	0	0
선천성기형, 변형 및 염색체 이상	2	2	3
달리 분류되지 않은 증상, 징후와 임상 및 검사의 이상 소견	312	337	324
손상, 중독 및 외인에 의한 특정 기타 결과	499	563	518
건강상태 및 보건서비스 접촉에 영향을 주는 요인	70	86	148
전체(질병 2개 이상 포함)	6,324	6,743	6,434

<고독사 사망 전 1년간 기초생활수급 대상자 여부>

연령별(1)	2021		
	계	남자	여자
19세이하	0	0	0
20대	13	10	3
30대	28	21	7
40대	137	113	24
50대	417	377	40
60대	456	406	50
70대	179	136	43
80대이상	70	49	21
계	1,300	1,112	188

<고독사 발견 당시 거주지 현황>

거주지 유형(1)	거주지 유형(2)	2023	
		고독사 수 (명)	발생 비중 (%)
계	소계	3,661	100.0
주거	소계	3,316	90.6
	주택	1,762	48.1
	아파트	798	21.8
	원룸· 오피스텔	756	20.7
비주거	소계	345	9.4
	여관·모텔	137	3.7
	2) 고시원	143	3.9
	3) 기타	65	1.8



「장래인구추계」성 및 연령별 추계인구(1세별, 5세별)

「주거실태조사」(일반가구)지역별 소득계층별 주택유형

건강보험심사평가원,「건강보험통계」22대 분류별 진료현황

특정 질병코드분류별 수술 진료환자수 및 진료건수

「국민기초생활보장수급자현황」국민기초일반수급자수

<장래인구추계-성별 및 연령별 추계인구>

전체 수+
성별 비율

가정별	성별	연령별	2021
중위 추계(기본 추계: 출산율-중위 / 기대수명-중위 / 국제순이동-중위)	전체	계	51,769,539
		0 - 4세	1,618,830
		5 - 9세	2,231,797
		10 - 14세	2,297,714
		15 - 19세	2,345,886
		20 - 24세	3,278,327
		25 - 29세	3,675,202
		30 - 34세	3,306,451
		35 - 39세	3,683,026
		40 - 44세	3,944,406
		45 - 49세	4,213,912
		50 - 54세	4,425,748
		55 - 59세	4,113,626
		60 - 64세	4,064,749
		65 - 69세	2,894,911
		70 - 74세	2,090,525
		75 - 79세	1,567,444
		80세이상	2,016,985
		80 - 84세	1,157,812
		85 - 89세	605,532
		90 - 94세	203,889
		95 - 99세	43,381
		100세 이상	6,371
	남자	계	25,870,941
		0 - 4세	830,797
		5 - 9세	1,143,757
		10 - 14세	1,183,139
		15 - 19세	1,216,428
		20 - 24세	1,711,404
		25 - 29세	1,962,374
		30 - 34세	1,745,107

<지역별 소득계층별 주택유형>

거주지

시도구분(1)	소득구분(1)	2021						
		계	단독주택	아파트	연립주택	다세대주택	비거주용 건물내 주택	주택이외의 거처
전국	전체	100.00	30.40	51.50	2.10	9.30	1.50	5.20
	저소득층	100.00	45.10	33.50	2.10	9.30	2.00	8.00
	중소득층	100.00	23.50	58.20	2.30	10.60	1.30	4.10
	고소득층	100.00	13.70	76.00	1.80	6.20	0.90	1.50
수도권	전체	100.00	22.40	52.00	2.40	14.90	1.40	6.90
	저소득층	100.00	32.80	34.50	2.00	16.50	1.90	12.30
	중소득층	100.00	20.10	52.90	2.80	17.20	1.40	5.60
	고소득층	100.00	11.50	74.90	2.10	9.10	0.70	1.70
광역시 등	전체	100.00	28.40	59.10	1.40	5.90	1.40	3.80
	저소득층	100.00	42.00	40.50	1.80	8.00	1.60	6.00
	중소득층	100.00	21.00	68.80	1.40	5.30	1.10	2.40
	고소득층	100.00	10.00	85.30	0.60	1.50	1.50	1.10
도지역	전체	100.00	44.20	45.90	2.20	2.60	1.70	3.50
	저소득층	100.00	60.00	28.30	2.30	2.40	2.30	4.60
	중소득층	100.00	31.20	59.90	2.10	3.20	1.10	2.60
	고소득층	100.00	23.10	71.40	1.70	1.30	1.20	1.30

<22대 분류별 진료현황>

21대 질환 보유자

질병22대분류별(1)	진료형태별(1)	2021				
		진료실인원수 (명)	입내원일수 (일)	요양급여일수 (일)	진료비 (천원)	급여비 (천
계	계	47,539,838	874,351,947	10,025,189,077	92,308,501,084	69,210,819,525
	입원	6,968,102	136,354,318	208,897,686	34,902,328,938	28,365,198,647
	외래	47,321,571	737,997,629	9,816,291,391	57,406,172,147	40,845,620,878
특정감염성 및 기생충성 질환(A00-B99)	계	9,682,481	25,678,671	210,704,169	2,487,206,932	1,841,430,962
	입원	416,006	3,268,819	5,507,822	871,485,400	712,847,887
	외래	9,502,371	22,409,852	205,196,347	1,615,721,533	1,128,583,075
신생물(C00-D48)	계	4,363,598	34,153,964	381,845,099	10,951,938,668	9,558,699,014
	입원	685,943	13,887,933	25,256,015	5,934,552,880	5,180,878,676
	외래	4,304,976	20,266,031	356,589,084	5,017,385,788	4,377,820,338
혈액 및 조혈기관의 질환과 면역기전을 침범한 특정 장애(D50-D89)	계	715,292	1,876,799	34,246,645	502,249,512	414,183,083
	입원	26,049	229,140	479,029	118,644,351	100,454,201
	외래	704,329	1,647,659	33,767,616	383,605,161	313,728,883
내분비, 영양 및 대사질환(E00-E90)	계	8,196,479	43,709,702	1,580,096,768	5,066,452,618	3,489,624,833
	입원	157,911	2,028,982	3,958,537	433,340,051	336,061,815
	외래	8,156,868	41,680,720	1,576,138,231	4,633,112,567	3,153,563,018
정신 및 행동장애(F00-F99)	계	3,622,695	54,958,970	605,586,125	4,754,660,280	3,547,911,353
	입원	205,680	30,146,043	31,585,577	2,638,569,095	2,001,740,450
	외래	3,509,058	24,812,927	574,000,548	2,116,091,184	1,546,170,903
신경계의 질환(G00-G99)	계	3,638,264	27,892,989	327,823,290	3,486,731,771	2,620,110,276
	입원	291,463	15,451,467	18,020,841	2,224,262,265	1,751,452,458
	외래	3,527,022	12,441,522	309,802,449	1,262,469,506	868,657,818
눈 및 눈부속기의 질환(H00-H59)	계	14,606,479	38,864,164	226,243,732	3,936,002,442	2,820,663,607
	입원	532,561	1,011,442	1,824,940	991,982,735	793,682,894
	외래	14,586,240	37,852,722	224,418,792	2,944,019,708	2,026,980,712
귀 및 유양돌기의 질환(H60-H95)	계	5,178,029	14,594,471	88,953,947	899,895,601	612,699,014
	입원	98,210	397,868	1,095,557	164,815,648	118,417,681
	외래	5,153,120	14,196,603	87,858,390	735,079,953	494,281,333
순환기계의 질환(I00-I99)	계	10,346,369	79,080,181	2,689,869,543	12,145,714,224	9,144,539,225
	입원	731,259	16,111,653	24,669,723	5,337,380,442	4,452,847,514
	외래	10,234,906	62,968,528	2,665,199,820	6,808,333,782	4,691,691,711

<특정 질병코드분류별 수술 진료환자수>

최근 1년간
수술내역

2021년	1세	남자	3978	4854
2021년	1세	여자	2812	3564
2021년	2세	남자	5261	6313
2021년	2세	여자	3892	4765
2021년	3세	남자	5854	6999
2021년	3세	여자	4734	5751
2021년	4세	남자	6496	7567
2021년	4세	여자	5354	6311
2021년	5세	남자	7154	8133
2021년	5세	여자	6360	7221
2021년	6세	남자	7696	8772
2021년	6세	여자	6779	7772
2021년	7세	남자	7218	8190
2021년	7세	여자	6200	7116
2021년	8세	남자	7342	8375
2021년	8세	여자	6569	7623
2021년	9세	남자	8672	10078
2021년	9세	여자	8019	9355
2021년	10세	남자	9048	10530
2021년	10세	여자	8512	10019
2021년	11세	남자	9149	11055
2021년	11세	여자	8705	10765

<국민기초생활보장수급자현황>

기초생활
수급자 여부

행정구역별(시도)(1)	연령별(1)	연령별(2)	2021		
			계	남자	여자
전국	계	소계	2,268,852	1,011,833	1,257,019
	0 - 4세	소계	27,473	14,167	13,306
		0세	2,747	1,418	1,329
		1세	4,653	2,411	2,242
		2세	5,639	2,917	2,722
		3세	6,685	3,444	3,241
		4세	7,749	3,977	3,772
	5 - 9세	소계	79,313	40,683	38,630
		5세	9,653	4,927	4,726
		6세	11,117	5,645	5,472
		7세	17,305	8,844	8,461
		8세	18,966	9,743	9,223
		9세	22,272	11,524	10,748
	10 - 14세	소계	130,213	66,546	63,667
		10세	23,928	12,148	11,780
		11세	24,687	12,755	11,932
		12세	24,323	12,432	11,891
		13세	27,410	14,090	13,320
		14세	29,865	15,121	14,744
	15 - 19세	소계	147,840	74,359	73,481
		15세	28,386	14,580	13,806
		16세	29,278	14,847	14,431
		17세	32,467	16,540	15,927
		18세	34,064	17,258	16,806
		19세	23,645	11,134	12,511
	20 - 24세	소계	91,310	39,985	51,325
		20세	20,251	7,601	12,650
		21세	20,312	7,730	12,582

3. 데이터 탐색 및 정제

일반인데이터→ 고독사데이터 형식에 맞게 분류

<일반인 거주지 데이터>

<고독사 거주지 데이터>

(1) 일반인데이터→ 고독사데이터 형식에 맞게 분류

<일반인 기초생활수급자 데이터>

0세~5세(남: /여:)
5세~10세(남: /여:)
15세~20세(남: /여:)
20세~25세(남: /여:)
...
80세~85세(남: /여:)
85세~90세(남: /여:)
90세~95세(남: /여:)
90세~100세(남: /여:)
100세 이상(남: /여:)

<고독사 기초생활수급자 데이터>

19세 이하(남: /여:)
20대(남: /여:)
30대(남: /여:)
40대(남: /여:)
50대(남: /여:)
60대(남: /여:)
70대(남: /여:)
80대 이상(남: /여:)

(2)비율계산-일반인

나이 = c("19세이하"=8394227, "20대"=6953529, "30대"=6989477, "40대"=8158318,
"50대"=8539374, "60대"=6959660, "70대"=3657969, "80대이상"=2026985)
#성별은 5:5비율
거주지 ←c("주택"=0.29,"아파트"=0.524,"원룸오피스텔"=0.114,"기타"=0.072)
질병←c(감염성=9682102, 신생물=4363598, 혈액=715292, 내분비=8196479,정신=3622695, 신경=3638264,
눈=14606479, 귀=5178029, 순환=10346369, 호흡기=19076580, 소화기=30612231, 피부=14182746,
근골격=18105833, 비뇨생식=9826285,임신출산=390234, 주산기=133200, 선천성=360585, 증상징후
=10011417, 손상중독=14201513, 건강상태=9045163)
수술여부←c("19세이하"=449954,"20대"=715210,"30대"=892477,"40대"=1322581,"50대"=1891875,"60
대"=2124805,"70대"=1346092,"80대이상"=595237)/51769539
수급자여부←c("19세이하"=384839,"20대"=135124,"30대"=103025,"40대"=226099,"50대"=353396,"60
대"=425661,"70대"=344681,"80대이상"=296027)/51769539

(2)비율계산-고독사

나이 = c("19세이하", "20대", "30대", "40대", "50대", "60대", "70대", "80대이상"),
남성 = c(1, 37, 120, 436, 900, 860, 314, 135),
여성 = c(1, 16, 44, 88, 91, 114, 104, 67)
거주지 ← c("주택"=0.481,"아파트"=0.218,"원룸오피스텔"=0.207,"기타"=0.094)
질병 ← c(감염성=65, 신생물=178, 혈액=32, 내분비=550, 정신=617, 신경=252, 눈=375, 귀=112, 순환=772, 호흡기=560, 소화기=714, 피부=353, 근골격=717, 비뇨생식=283, 임신출산=0, 주산기=0, 선천성=3,증상징후=324, 손상중 독=518, 건강상태=148)/3378
수술여부←c("19세이하"=1,"20대"=11,"30대"=25,"40대"=76,"50대"=192,"60대"=194,"70대"=80,"80대이상"=33)/3378
수급자여부← c("20대"=13,"30대"=28,"40대"=137,"50대"=417,"60대"=456,"70대"=179,"80대이상"=70)/3378

(3)샘플링

고독사 1000명+일반인1000명

⇒ 실제 통계 기반 확률 샘플링

⇒ 질병중 임신출산은 여성에게만 부여

⇒ 각 질병은 통계청 및 복지부가 제공한 단일 질병 보유율에 기반해 개별 샘플링→복수 질환 동시 보유 패턴은 결합 확률 정보의 부재로 인해 구현이 어려워, 프로젝트에서는 질병 간 독립성을 전제로 시뮬레이션을 수행함.

seed 다르게 30번 반복

⇒ 단일 표본에 의존시 통계적으로 우연적 편향이나 대표성 문제가 발생

⇒ seed 값을 변경해 30회의 반복 샘플링과 모델링을 수행, 모델의 성능이 샘플에 따라 얼마나 안정적인지를 검증

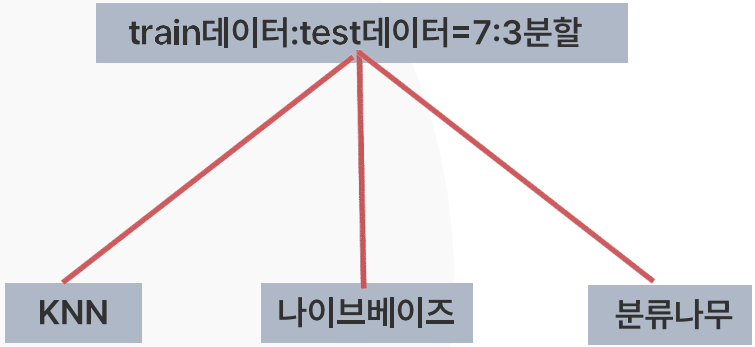
샘플링된 데이터셋 구성 변수

변수	설명
age	19세 이하,20대,...,70대, 80대 이상
gender	여성, 남성
질병	질병O=1, 질병X=0
recipient	수급O=1, 수급X=0
surgery	수술O=1, 수술X=0
residence	주택, 아파트, 원룸오피스, 기타
label	고독사=1, 일반인=0
iteration	반복 식별자 변수 → 1부터 30까지 각 반복 번호를 부여하여 모델 성능의 반복 비교 및 평균 산출에 활용

샘플링된 데이터셋

age	gender	질병_감염	질병_신생	질병_혈액	질병_내분	질병_정신	질병_신경	질병_눈	질병_귀	질병_순환	질병_호흡	질병_소화	질병_피부	질병_근골	질병_비뇨	질병_임신	질병_주산	질병_선천	질병_증상	질병_손상	질병_건강	surgery	recipient	residence	label	iteration
60대	남	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0	0	0	0	0	0	0	0	원룸오피스	1	1
50대	남	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	주택	1	1
40대	남	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	주택	1	1
40대	남	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	주택	1	1
50대	남	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	주택	1	1
60대	여	0	0	0	0	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	1	주택	1	1
60대	남	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	주택	1	1
60대	남	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	아파트	1	1
40대	남	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	원룸오피스	1	1
60대	남	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	주택	1	1
80대이상	남	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	아파트	1	1
40대	남	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	주택	1	1
40대	남	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	원룸오피스	1	1
80대이상	남	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	주택	1	1
60대	남	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	주택	1	1
60대	남	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	원룸오피스	1	1
70대	남	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	주택	1	1
50대	남	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	아파트	1	1
60대	남	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	기타	1	1
50대	남	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	주택	1	1
40대	여	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	아파트	1	1

4. 데이터마이닝 방법 결정



데이터는 iteration으로 구분하여 반복별 분석

1. KNN

- 사용 패키지: class::knn()

k = 16 (사전 실험을 통해 선정)

- 전처리

KNN은 거리 기반 모델→age, gender, residence 수치형 행렬로 변환 (model.matrix 사용)

- 예측및 성능

분류 결과(0, 1)→정오행렬로 Accuracy, Precision, Recall, F1-score 계산

2. 나이브베이즈

- 사용 패키지

e1071::naiveBayes()

- 전처리

별도 전처리 필요x

- 예측및 성능

type = "class": 예측 결과→정오행렬로 Accuracy, Precision, Recall, F1-score 계산

type = "raw"로 확률값 추출 가능 → auc계산

3. 분류나무

- 사용 패키지 rpart::rpart()

- 파라미터 설정

cp = 0.005

minsplit = 5

maxdepth = 8

- 전처리

별도 전처리 필요x

- 예측및 성능

type = "class": 예측 결과→정오행렬로 Accuracy, Precision, Recall, F1-score 계산

type = "prob"로 양성 클래스 확률 추출 가능

5. 모델 성능 비교 결과

모델	정확도 평균	정확도 평균 표준편차
KNN	0.7704	0.0224
나이브 베이즈	0.6763	0.0336
Tree	0.7966	0.0234

- KNN: 정확도는 Tree보다 낮지만, 가장 낮은 표준편차→ 반복 실험 간 성능의 일관성이 좋음
- 나이브베이즈: 정확도 평균 가장낮음 + 표준오차 가장 높음→ 불안정하고 부정확한 모델
- Tree: 정확도 평균이 가장 높고 표준편차도 KNN이랑 별로 차이 안남→가장 정확한 예측력

모델	양성예측도	민감도	특이도	F1-score	AUC
KNN	0.7235	0.9273	0.6493	0.8125	0.7883
나이브 베이즈	0.6093	0.9597	0.3870	0.7447	0.8741
Tree	0.8081	0.8066	0.8088	0.8065	0.8516

- 양성예측도: 예측된 고독사자 중 실제 고독사자의 비율
- 민감도: 실제 고독사자 중 맞게 예측된 비율
- 특이도: 실제 일반인을 일반인으로 맞게 분류한 비율
- F1-score: 양성예측도와 민감도 조화평균

- **KNN: 민감도는 매우 높지만, 특이도는 낮아 일반인 분류에서 오분류 위험**
- **나이브베이즈: AUC는 높지만, 특이도, 양성예측도가 낮아 과잉 탐지 경향**
- **Tree: 정밀도, 특이도, F1-score 모두 균형 잡힘(0.8이상)**

모델 선정 기준

고독사 예측 모델의 실용적 목적: 단순히 예측 정확도를 높이는 것이 아니라, 위험군(고독사자)을 최대한 놓치지 않으면서도, 불필요한 개입(실제 일반인을 고독사 위험군으로 잘못 분류)도 최소화!

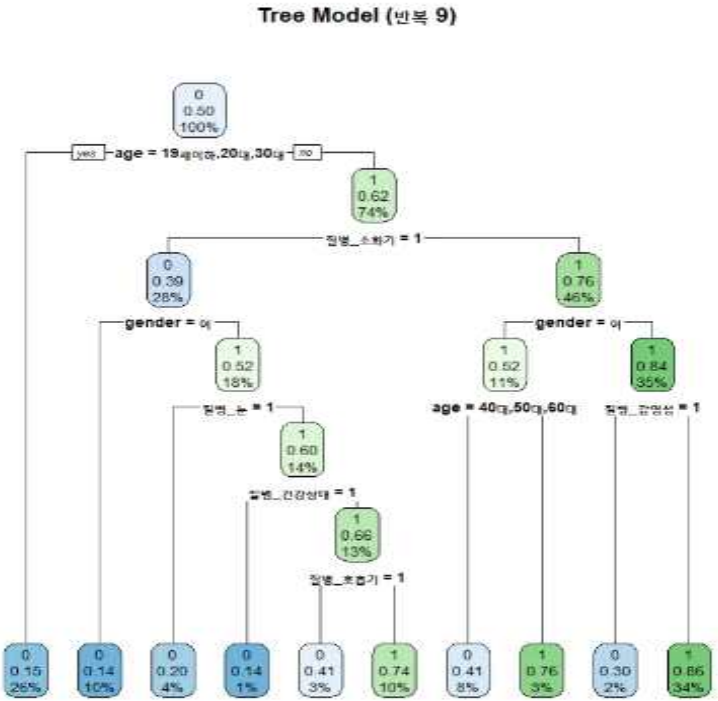
⇒ 고독사를 최대한 잡아내는 민감도와, 일반인을 고독사로 과도하게 분류하지 않는 양성예측도와 특이도 사이의 균형이 중요

⇒ 따라서 단일 지표보다는 양성예측도-민감도-특이도-F1-score를 종합적으로 고려한 판단이 필요

모델	양성예측도	민감도	특이도	F1-score	AUC
KNN	0.7235	0.9273	0.6493	0.8125	0.7883
나이브 베이즈	0.6093	0.9597	0.3870	0.7447	0.8741
Tree	0.8081	0.8066	0.8088	0.8065	0.8516

Tree모델이 이 균형에서 가장 우수!

6. 선정된 모델의 실제 예측 구조

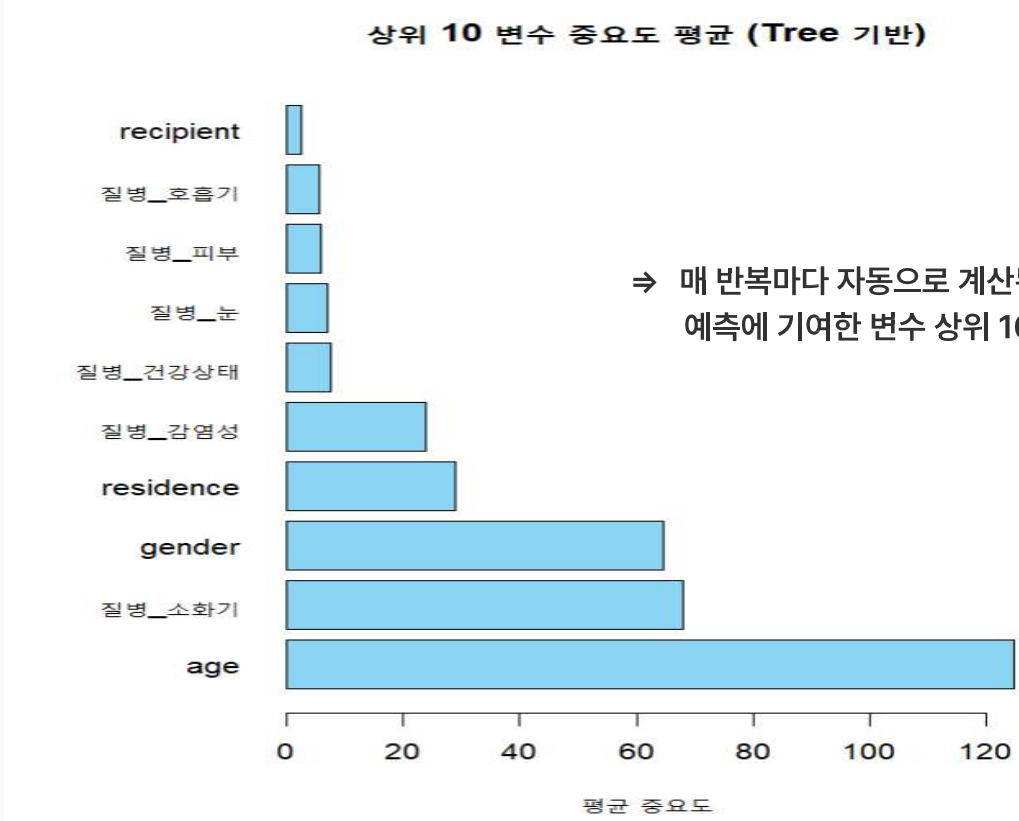


[분류나무 모델에서 30회 반복 실험 중, 전체 평균 정확도에 가장 근접한 반복의 결과를 시각화한 의사결정나무 구조]

- 주요 예측경로

연령 → 소화기 질환 → 성별 → 눈 질환, 건강상태접촉에 영향을 주는 요인, 호흡기 질환

본 시뮬레이션은 실제 통계 기반의 특성 분포를 반영한 확률 샘플링을 통해 구성되었으며, 변수 간 상관관계를 직접적으로 반영하지는 않았으나, 현실적인 비율 기반 데이터 내에서 반복적으로 도출된 의사결정 흐름이라는 점에서 모델의 해석 가능성과 실용성을 함께 고려한 결과로 볼 수 있음.



⇒ 매 반복마다 자동으로 계산된 변수 중요도를 평균 내서 구한 예측에 기여한 변수 상위 10개

6. 결론

고독사 위험군과 일반인을 구분할 수 있는 예측 모델을 구축하기 위해, **실제 통계 기반의 확률 분포**를 반영한 **시뮬레이션 데이터를 생성**

30회 반복 실험 결과, **Tree 모델**은 가장 높은 정확도 평균(0.7966)와 더불어 정밀도(0.8081), 특이도(0.8088), F1-score(0.8065)에서 **가장 균형 있는 성능**을 보임

Tree분기⇒ 연령 → 소화기 질환 → 성별 → 눈/건강상태접촉에 요인을 주는 질병/호흡기' 순으로 주요 결정경로가 형성
변수 중요도⇒ **연령, 소화기 질환, 성별순으로 영향력이 크게 나타남**

1. 향후 고독사 위험군을 조기에 발견할 수 있는 예측 시스템 구축을 위한 기초적 탐색자료로 활용

2. 현실통계 기반의 시뮬레이션 데이터를 바탕으로 트리모델 적용→ 연령, 질환등의 영향을 정량적으로 분석
+예측경로를 해석 ⇒ 우선개입이 필요한 고위험군을 사전에 파악하는데 참고자료가 될 수 있을 것이라 기대

6. 결론-한계점

1. 질병 간 관계를 고려 X

이 프로젝트는 각 질병을 독립적으로 샘플링했기 때문에, 실제처럼 여러 질병이 함께 나타나는 경우(다질환 패턴)나 질병 간 연관성은 반영되지 않았다.

2. 수급자, 수술 여부 등은 비율에 따른 추정값

일부 변수는 개인의 실제 이력 대신, 전체 통계 비율을 기준으로 샘플링된 값이기 때문에 정확한 개인 예측보다는 전체적인 경향 파악에 적합하다.

3. 시뮬레이션 데이터 기반의 예측 모델

실제 데이터를 그대로 사용한 것이 아니라 현실적인 비율을 반영한 가상의 데이터(시뮬레이션)로 분석했기 때문에 실제 적용 전에는 실제 데이터로 검증이 꼭 필요하다.

들어주셔서
감사합니다.

-최적의 k값 찾을때 (n)^(1/2)로 계산한 결과 44.7이 나와서 50까지 계산함

```r
#사전준비비
library(class)
library(caret)
library(dplyr)

godok <- read.csv("C:\\Users\\User\\OneDrive\\바탕 화면\\고득_weighted.csv")
general <- read.csv("C:\\Users\\User\\OneDrive\\바탕 화면\\일반_weighted.csv")
godok$label <- as.factor(godok$label)
general$label <- as.factor(general$label)
#병합
data <- rbind(godok, general)

#범주형 변수 factor 처리
data$gender <- as.factor(data$gender)
data$age <- as.factor(data$age)
data$residence <- as.factor(data$residence)
dummy_vars <- dummyVars(~ ., data = data[, -which(names(data) == "label")])
data_dummy <- predict(dummy_vars, newdata = data) %>% as.data.frame()
data_dummy$label <- data$label

#학습검증분할(7:3)
set.seed(123)
train_idx <- sample(1:nrow(data_dummy), 0.7 * nrow(data_dummy))
train_data <- data_dummy[train_idx, ]
test_data <- data_dummy[-train_idx, ]
x_train <- train_data[, -ncol(train_data)]
y_train <- train_data$label
x_test <- test_data[, -ncol(test_data)]
y_test <- test_data$label
```

```r
#최적의k
for (k in 1:50) {
  pred_k <- knn(train = x_train, test = x_test, cl = y_train, k = k)
  acc <- sum(pred_k == y_test) / length(y_test)
  cat("k =", k, "정확도 =", round(acc, 4), "\n")
}
#k=16채택
knn_pred <- knn(train = x_train, test = x_test, cl = y_train, k = 16)
confusionMatrix(knn_pred, y_test)
```

```
k = 1 정확도 = 0.7583
k = 2 정확도 = 0.7633
k = 3 정확도 = 0.7667
k = 4 정확도 = 0.7733
k = 5 정확도 = 0.7767
k = 6 정확도 = 0.7783
k = 7 정확도 = 0.77
k = 8 정확도 = 0.7733
k = 9 정확도 = 0.775
k = 10 정확도 = 0.7767
k = 11 정확도 = 0.78
k = 12 정확도 = 0.7833
k = 13 정확도 = 0.7817
k = 14 정확도 = 0.78
k = 15 정확도 = 0.78
k = 16 정확도 = 0.7883
k = 17 정확도 = 0.7867
k = 18 정확도 = 0.785
k = 19 정확도 = 0.7833
k = 20 정확도 = 0.7817
```