심장병에 대한 다변량 자료 분석

정보통계학과 20160854 김민경

1. 서론 : 연구동기
2. 본론
    2.1 데이터 출처 및 변수 설명
    2.2 기초통계량
    2.3 상관분석
3. 통계분석
    3.1 주성분 분석
    3.2 인자분석
    3.3 로지스틱 회귀분석
    3.4 신경망
    3.5 의사결정나무를 이용한 앙상블 기법
    3.6 ROC Curve & AUC
    3.7 군집분석
4. 결론 : 연구 요약 및 한계점과 의의

초록

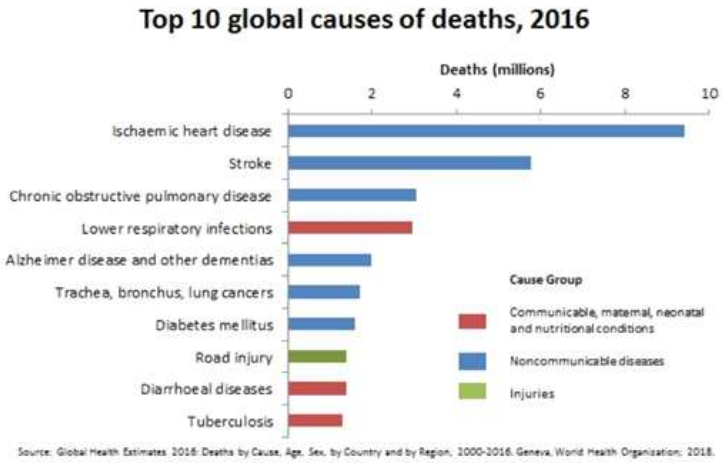
본 연구는 국내외적으로 주요 3대 사인으로 꼽히는 심장병 데이터에 대해 논의하고자 수행되었다. 이 연구는 심장병 데이터의 다변량 자료 분석과 예측 분석을 통해 심장병에 대한 인사이트를 얻고자하였다. 학부에서 배운 R 프로그래밍을 사용하여 데이터 분석을 진행하였다. 본 연구의 결과 로지스틱 회귀분석이 심장병 예측에 가장 적합하다는 결론과 탈세미아 유전 혈액 장애, 가슴 통증 유형, 휴식과 관련된 운동으로 인한 ST 우울증이 심장병 예측에 주요하다는 결론이 도출되었다. 4차 산업 혁명의 IOT와 AI의 기술을 접목하여 심장병의 조기진단에 활용할 수 있다는 의미에서 이 연구는 의의를 가진다.

1. 서론

통계청의 사망원인통계 자료(2019)를 살펴보면 2018년 1월부터 2019년 4월까지의 사망 사건에 대하여 3대 사인은 암, 심장 질환, 폐렴으로 전체 사망의 45%를 차지하였다. 특히, 심장질환은 2008년부터 현재까지 증가 추세에 있다. 또한 WHO의 세계10대 사망원인 발표(2018)에 따르면 2016년 전 세계 사망자 5690만 명 중 절반 이상(54%)이 10대 원인으로 사망했고 그 중 총 1,520만 명의 사망자를 낸 세계 최대 사망원인은 심장병과 뇌졸중이었다. 이러한 질병들은 지난 15년 동안 전 세계적으로 주요 사망원인으로 남아있다.

위와 같은 통계자료를 접하고 심장 질환에 대한 연구의 필요성을 느껴 심장병 데이터를 분석하게 되었다.

그림 1.1 WHO Top 10 global causes of deaths, 2016



2. 본론

2.1 데이터 출처 및 변수 설명

Kaggle 사이트를 통해 UCI에서 제공한 Heart disease 데이터를 다운받아 사용하였다. Heart disease 데이터는 본래 76개의 변수를 가진 데이터이지만 Kaggle에서는 14개의 변수만을 공개하였고 5개의 연속형 변수, 9개의 범주형 변수로 이루어진 303명 환자의 데이터이다.

표 2.1 변수 설명

변수명	변수설명	변수명	변수설명
age	나이	thalach	최대 심박수(bpm)
sex	성별 (0=female,1=male)	exang	운동 유발 협심증 (0=no, 1=yes)
cp	가슴 통증 유형 (0: 전형적인 협심증, 1: 비정형 협심증, 2: 비혈관 통증, 3: 무증상적 협심증)	oldpeak	휴식과 관련된 운동으로 인한 ST 우울증
trestbps	휴식기 혈압(mmhg)	slope	피크 운동 ST 세그먼트의 기울기 (1: 업슬로핑, 2: 플랫, 3: 다운슬로핑)
chol	혈청 콜레스테롤(mg/dl)	ca	플러로소피에 의해 색칠된 주요 혈관 수
fbs	단식 혈당 120mg/dl 이상 (0 = false,1 = true)	thal	탈라세미아라고 불리는 혈액 장애(유전) (thal: 3 = 정상, 6 = 고정 결함, 7 = 되돌릴 수 있는 결함)
restecg	휴면 심전도 결과 (0 = 정상, 1 = ST-T파 이상이 있는 경우,	target	심장병 여부 (0 = no, 1 = yes)

	2 = 에스테스의 기준에 의해 발생 가능하거나 확실한 좌심실 비대증을 보이는 경우)		

age변수는 환자들의 나이를 나타낸 연속형 변수이고 sex변수는 환자들의 성별을 0인 경우 여성으로 1인 경우 남성으로 나타낸 범주형 변수, cp변수는 가슴 통증 유형을 0: 전형적인 협심증, 1인 경우 비정형 협심증, 2인 경우 비혈관 통증, 3인 경우 무증상적 협심증으로 나타낸 범주형 변수이다. 협심증(Angina Pectoris)란 “관상 동맥에 콜레스테롤과 같은 이물질이 쌓여 혈관이 좁아지면서 생기는 병”(서울특별시 서울의료원 심혈관센터)을 말한다. 빠른 속도로 걷거나 운동 등으로 심장이 일이 많이 할 때 흉통이 생기는 증상이 가장 전형적인 협심증, 일시적인 관상동맥의 경련에 의한 흉통이 비정형 협심증, 혈관에 의한 흉통이 아닌 경우 비혈관 통증, 아무리 심한 활동을 하여도 통증이 없는 경우가 무증상적 협심증이다. trestbps의 변수는 휴식기 혈압을 나타내는 연속형 변수로 일반적으로 혈압이 최대 140mmhg이상, 최저 90mmhg이상일 때 고혈압이라 부르며 심장비대, 심부전, 협심증으로 이어질 수 있는 것으로 알려져있다. chol변수는 혈청 콜레스테롤을 나타내는 연속형 변수로 관상동맥에 콜레스테롤이 쌓일 경우 혈관이 좁아지며 협심증을 유발할 수 있다. fbs변수는 단식 혈당이 120mg/dl이상인 경우 1, 120mg/dl미만인 경우 0으로 나타내는 범주형 변수이다. 일반적으로 체내 혈당이 180mg/dl이상인 경우 소변에서 당이 배출되며 당뇨병의 증상이 나타나는 것으로 알려져있다. restecg변수는 휴면 심전도 결과를 0인 경우 정상, 1인 경우 ST-T파 이상이 있는 경우, 2인 경우 에스테스의 기준에 의해 발생 가능하거나 확실한 좌심실 비대증을 보이는 경우로 나타내는 범주형 변수이다. exang변수는 운동 유발 협심증이 아닌 경우 0으로 운동 유발 협심증인 경우는 1로 나타내는 범주형 변수이다. oldpeak변수는 휴식과 관련된 운동으로 인한 ST 우울증을 나타내는 연속형 변수이다. slope변수는 피크 운동 ST 세그먼트의 기울기가 1인 경우 업슬로핑, 2인 경우 플랫, 3인 경우 다운슬로핑을 나타내는 범주형 변수이다. 일반적으로 업슬로핑인 경우 심근손상, 플랫인 경우 허혈성 심장질환, 다운슬로핑인 경우 좌심실 비대증을 보이는 것으로 알려져있다. ca변수는 플러로소피에 의해 색칠된 주요 혈관 수를 나타내는 연속형 변수로 혈관 수가 많을수록 심근의 에너지원인 산소와 영양소가 풍부해져 심장의 펌프 작용에 도움을 준다. thal변수는 탈라세미아라고 불리는 유전 혈액장애가 3인 경우 정상, 6인 경우 고정 결함, 7인 경우 되돌릴 수 있는 결함을 나타내는 범주형 변수이다. 마지막으로 target변수는 0인 경우 심장병이 없고 1인 경우 심장병이 있음을 나타내는 범주형 변수로 주어진 데이터의 출력변수이다.

2.2 기초통계량

분석에 앞서 데이터에 대한 전반적인 이해를 돕기 위해 데이터의 기초통계량을 구하였다. 우선, 심장병 데이터에 있는 연속형 변수 5개에 대해 최소값, 최대값, 중앙값, 평균, 표준편차를 구하고 boxplot을 구하였다. 범주형 변수 9개에 대해서는 target변수 값에 따른 빈도표를 작성하였다. 표 2.2를 보면 age(나이)변수의 경우 최소 29세에서 77세까지 고르게 분포해있는 것을 확인할 수 있고, tresbps(휴식기 혈압)과 chol(혈청 콜레스테롤)의 경우 5개 정도의 이상치를 발견할 수 있다. 표준편차를 살펴보면 chol(혈청 콜레스테롤)과 thalach(최대 심박수)의 편차가 큰 것을 확인할 수 있다.

표 2.2 심장병 데이터의 연속형 변수 기초통계량

변수명	최소값	최대값	중앙값	평균	표준편차
age	29	77	55	54.37	9.08
tresbps	94	200	130	131.6	17.53
chol	126	564	240	246.3	51.83
thalach	71	202	153	149.6	22.90
oldpeak	0	6.20	0.80	1.04	1.16

그림 2.2 연속형 변수에 대한 boxplot

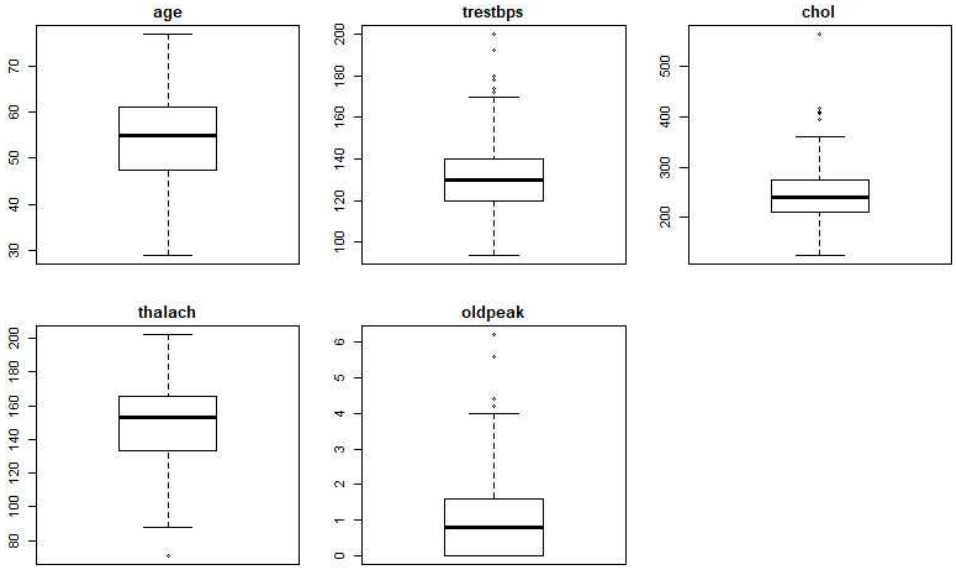


그림 2.2 심장병 데이터 범주형 변수 빈도표

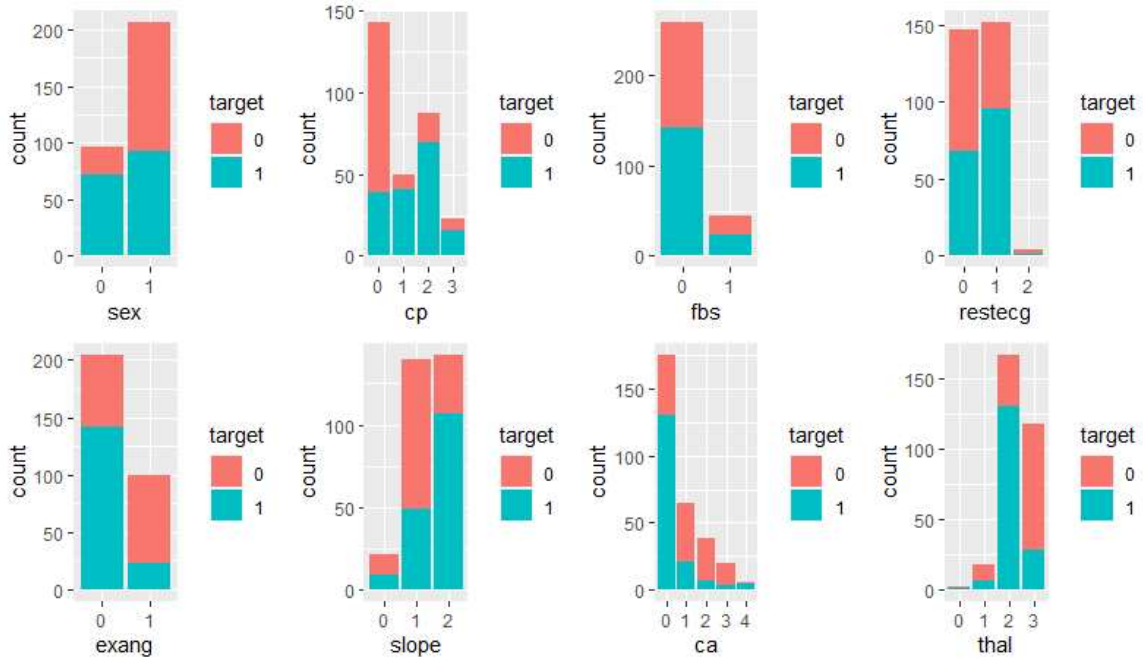


그림 2.2의 target변수의 값에 따른 범주형 변수의 빈도표를 살펴보면 sex(성별)은 여성의 경우 심장병인 경우가, 남성의 경우 심장병이 아닌 경우가 더 많음을 확인할 수 있다. cp(가슴 통증 유형)는 전형적인 협심증의 경우의 빈도가 가장 높고, 많은 경우 심장병이 없으나 나머지 가슴 통증 유형의 대부분은 심장병이 있음을 확인할 수 있다. fbs(단식 혈당)의 경우 120mg/dl이하인 사람의 빈도가 더 많으나 각각의 범주의 경우 심장병 여부는 비슷해 보인다. restecg(휴면 심전도 결과)가 정상인 경우와 ST-T파 이상이 있는 경우의 빈도가 많고 ST-T파 이상이 있는 경우 심장병인 사람의 수가 더 많은 것으로 보인다. exang(운동 유발 협심증)은 안정성 협심증이라고 불리며 심장이 일을 많이 할 때 협심증이 심해진다. 운동 유발 협심증이 아닌 경우 심장병인 사람이 많고 운동 유발 협심증인 경우 심장병이 없는 사람이 더 많은 것으로 보인다. slope(피크 운동 ST 세그먼트의 기울기)를 살펴보면 플랫인 경우 심장병인 사람이 적고, 다운 슬로핑인 경우 심장병인 사람이 많음을 확인할 수 있다. ca(플러로소피)에 의해 색칠된 주요 혈관 수를 살펴보면 그 수가 많을수록 심장병인 사람의 수가 줄어드는 것을 확인할 수 있다.

2.3 상관분석

각 변수간의 선형 관계를 파악하기 위해 상관분석을 진행하였다. 상관계수는 1에 가까울수록 강한 양의 선형관계를 갖고, -1에 가까우면 강한 음의 선형관계, 0에 가까우면 선형관계가 아님을 나타낸다.

그림 2.3 상관행렬

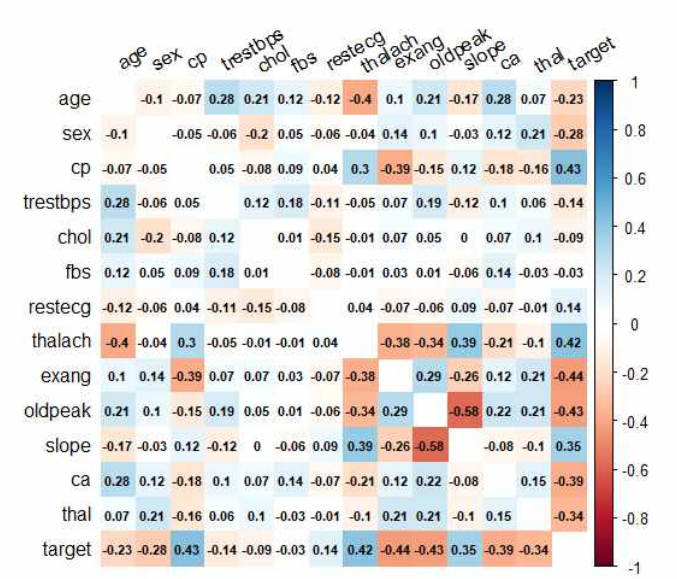


그림 2.3을 살펴보면 비교적 높은 양과 음의 상관계수를 갖는 변수관계는 보이지 않지만, slope변수와 oldpeak변수가 -0.58로 다소 높은 음의 상관계수를 갖는 것을 확인할 수 있다. 또한 많은 변수들의 상관계수 값이 0에 가까운 것으로 보아 선형관계가 없는 것으로 볼 수 있다.

3. 통계분석

3.1 주성분 분석

차원축소는 분석대상이 되는 변수의 수를 줄이는 일련의 탐색적 자료분석의 과정으로 변수를 직접적으로 변환하는 변수변환(feature transformation)과 변수를 선택하는 변수선택(feature selection)에 기초한 방법으로 구분된다. 주성분 분석은 변수변환에 기초한 방법으로 원래 변수들에 대한 선형 변환을 통하여 주성분이라 불리는 새로이 생성된 변수들이 서로 직교하도록 하는 다변량 자료분석기법이다. 원 변수의 분산을 바탕으로 한 직교변환을 이용하기 때문에 새로 생성된 변수들 간에는 서로 상관관계가 없다. 주성분이 설명하는 분산의 크기는 제1주성분에서 제p주성분까지 순차적으로 감소한다. 변수의 단위가 다르거나 또는 분산의 차이가 큰 경우, 표본상관행렬을 이용하여 주성분을 구하면 해석하는데 편리하다.

표 3.1 표본상관행렬의 분산 설명력

주성분	표준편차	분산비율	누적분산비율
PC1	1.66	0.21	0.21
PC2	1.24	0.12	0.33
PC3	1.10	0.09	0.42
PC4	1.09	0.09	0.52
PC5	1.01	0.08	0.59
PC6	0.98	0.07	0.67
PC7	0.93	0.07	0.74
PC8	0.88	0.06	0.79
PC9	0.85	0.06	0.85
PC10	0.79	0.05	0.90
PC11	0.73	0.04	0.94
PC12	0.65	0.03	0.97
PC13	0.61	0.03	1.00

그림 3.1 Scree-plot

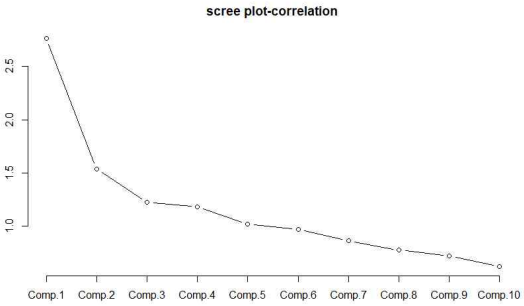


표3.1는 상관행렬을 사용하여 주성분 분석을 실시한 결과를 표로 정리한 것이다. 네 번째 주성분 PC4까지의 누적분산비율이 0.52이므로 4개의 주성분이 전체 분산의 약 52%를 설명하고 있고, 다섯 번째 주성분 PC5까지 누적분산비율이 0.59로 5개의 주성분이 전체 분산의 59%정도를 설명하고 있다. 스크리 그래프의 가파른 정도를 보고 큰 고유값과 작은 고유값을 구분하여 자연스럽게 적절한 개수를 정할 수 있다. 그림3.1을 참고하면 네 번 째, 다섯 번 째 주성분 이후의 기울기가 완만해지는 것으로 보아 주성분 개수는 4, 5개로 충분함을 확인할 수 있다.

전체를 합한 데이터만 이용하여 분석할 경우, 집단간 차이가 크면 세분된 결과가 상쇄되어 집단의 특성을 나타내지 못할 수도 있다. 따라서 target=0(심장병이 없다.), target=1(심장병이 있다.)의 두 개의 집단으로 나누어 주성분 분석을 진행한 결과와 전체 자료에 대한 주성분분석 결과를 진행한 결과를 비교하였다.

표 3.2 target에 따른 주성분 계수

	target=0(심장병 “no”)				target=1(심장병 “yes”)			
	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
age	0.248	0.385	0.174	0.233	0.501		0.375	0.160
sex	-0.265	-0.255	0.146	0.456	-0.171	-0.472	-0.191	0.245
cp	-0.268	0.213	0.568	-0.166	0.176	-0.468		-0.236
trestbps	0.268	0.356		-0.324	0.266	-0.270	0.195	-0.283
chol	0.140	0.266	-0.361	-0.249	0.213	0.207	0.301	-0.332
fbs	0.169	0.304			0.139	-0.416	0.214	
restecg		-0.269	0.205	-0.344	-0.226		0.139	0.384
thalach	-0.349	0.296		-0.321	-0.424	-0.256		-0.413
exang	0.320	-0.321	-0.390			0.149	-0.168	0.391
oldpeak	0.433	-0.158	0.321		0.419	-0.159	-0.386	
slope	-0.472	0.186	-0.339	0.119	-0.372		0.494	
ca	0.207	0.335		0.455		-0.242	0.426	0.238
thal		-0.126	-0.253	-0.292		-0.287	0.136	0.367
표준편차	1.488	1.310	1.140	1.13	1.445	1.262	1.195	1.145
누적분산 비율	0.170	0.302	0.402	0.499	0.160	0.283	0.393	0.494

표 3.3 전체 자료의 주성분 계수

	PC1	PC2	PC3	PC4	PC5
age	0.314	0.406			0.307
sex		-0.378	-0.555	0.255	
cp	-0.275	0.297	-0.357	-0.288	-0.163
trestbps	0.184	0.438	-0.204		-0.188
chol	0.117	0.365	0.408	0.343	-0.320
fbs		0.317	-0.482		0.233
restecg	-0.128	-0.221		-0.266	0.394
thalach	-0.416		-0.158	0.184	-0.323
exang	0.361	-0.263	0.126	0.115	
oldpeak	0.420		-0.110	-0.326	-0.251
slope	-0.380			0.495	0.247
ca	0.273		-0.184	0.328	0.435
thal	0.222	-0.201	-0.125	0.389	0.509
표준편차	1.66	1.24	1.10	1.09	1.01
누적분산 비율	0.21	0.33	0.42	0.52	0.59

target=0(심장병이 없다.)의 경우 첫 번째 주성분에 대하여 피크 운동 ST 세그먼트의 기울기(slope)에 비해 휴식과 관련된 운동으로 인한 ST 우울증(oldpeak)이 높을수록 첫 번째 주성분 값이 커지는 것을 확인 할 수 있다. 즉, ST 우울증이 높을수록 피크 운동 ST 세그먼트의 기울기가 업슬로핑이라는 해석을 할 수 있다. 두 번째 주성분에 대하여 나이(age)가 많을수록, 휴식기 혈압(trestbps)이 높을수록 운동유발 협심증(exang)이 없다는 것을 확인 할 수 있다. 세 번째 주성분에 대해서 무증상적 협심증(cp)일수록 운동유발 협심증(exang)이 아니라는 것을 확인 할 수 있다. 네 번째 주성분에 대해서 남성(sex)일수록, 플러로소피에 의해 색칠

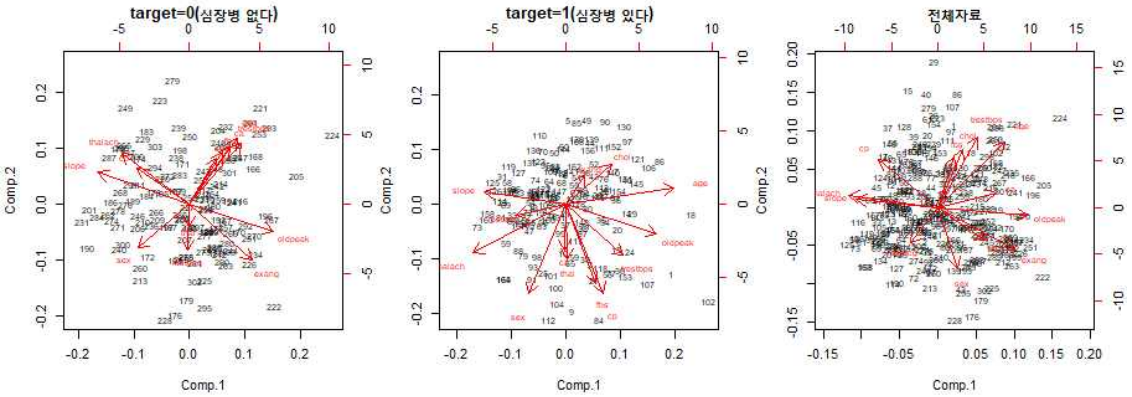
된 주요 혈관 수(ca)가 많을수록 휴면 심전도 결과(restecg)가 정상인 것을 확인할 수 있다.

target=1(심장병이 있다.)의 경우 첫 번째 주성분에 대하여 나이(age)가 많을수록 최대 심박수(thalach)가 낮다는 것을 확인 할 수 있다. 두 번째 주성분에 대하여 혈청 콜레스테롤(chol)이 높을수록 여자(sex)이고 가슴 통증 유형이 전형적인 협심증(cp)이라는 것을 확인할 수 있다. 세 번째 주성분에 대하여 피크 운동 ST 세그먼트의 기울기가 다운슬로핑(slope)일수록 휴식과 관련된 운동으로 인한 ST 우울증(oldpeak)이 낮다는 것을 확인할 수 있다. 네 번째 주성분에 대하여 휴면 심전도 결과가 좌심실 비대증(restecg)이고 운동유발 협심증(exang)이 있으면 최대 심박수(thalach)가 낮다는 것을 확인 할 수 있다.

전체자료에 대한 주성분 분석의 경우 첫 번째 주성분은 최대 심박수(thalach)에 비해 휴식과 관련된 운동으로 인한 ST 우울증(oldpeak)이 높을수록 첫 번째 주성분 값이 커지므로 최대 심박수와 ST 우을증의 대비성분이라고 할 수 있다. 두 번째 주성분은 휴식기 혈압(trestbps)이 높을수록 여성(sex)이라고 할 수 있고, 세 번째 주성분은 혈청 콜레스테롤(chol)이 높을수록 여성(sex)이라고 할 수 있다. 네 번째 주성분은 피크 운동 ST 세그먼트의 기울기가 다운 슬로핑일수록 휴식과 관련된 운동으로 인한 ST 우울증(oldpeak)이 낮다는 것을 확인할 수 있다. 다섯 번째 주성분은 탈라세미아라고 불리는 혈액 장애(thal)가 높을수록 최대 심박수(thalach)와 휴식기 혈압(trestbps)가 낮다는 것을 확인 할 수 있다.

target=1(심장병이 있다.)와 전체 자료에 대한 주성분 분석의 결과가 일부 비슷한데 이는 target=0(심장병이 없다.)의 전체 데이터 개수가 138개, target=1(심장병이 있다.)의 전체 데이터 개수가 165개로 target=1의 데이터가 target=0의 데이터보다 많아 전체자료에 대해 더 많은 영향을 주었을 것으로 짐작할 수 있다. 세 집단의 주성분 분석에서 공통적으로 나타난 주성분은 휴식과 관련된 운동으로 인한 ST 우울증(oldpeak)과 피크 운동 ST 세그먼트의 기울기(slope)의 대비이다. 이는 앞서 진행한 상관분석 결과에서 oldpeak변수와 slope변수가 다소 높은 음의 상관계수를 갖는 결과와 연관 지을 수 있다.

그림 3.2 주성분 그래프



주성분 그래프는 2차원 좌표축에 첫 번째 주성분과 두 번째 주성분을 그려 두 주성분 간의 관계와 패턴을 도출할 수 있으며 또한 전체 데이터가 주성분을 통해 변화되어 나타내는 관계도 알 수 있다. 주성분 그래프에서 가까운 거리와 방향일수록 변수들의 상관성이 높아진다. target=0(심장병이 없다.)의 주성분 그래프에서 (thalach, cp, slope), (sex), (oldpeak, exang), (thal, restecg), (chol, thal, ca, age, trestbps)의 그룹별로 상관성이 높은 것으로 보인다. target=1(심장병이 있다.)의 주성분 그래프에서는 (age), (oldpeak, trestbps), (fbs,

cp), (ca, thal), (sex), (thalach, slope, restecg), (exang, chol)의 그룹별로 상관성이 높은 것으로 보인다. 전체자료의 주성분 그래프에서는 (cp), (fbs, chol, trestbps, age), (ca), (oldpeak), (thal, exang), (sex), (restecg), (slope, thalach)의 그룹별로 상관성이 높은 것으로 보인다. 세 집단의 주성분 그래프에서 공통적으로 (thalach, slope)이 같은 그룹으로 묶이며 상관성이 높은 것으로 나왔는데 이는 2.3절에서 진행한 상관분석의 결과 slope변수와 thalach변수 0.39의 상관계수를 가지며 양의 상관관계를 보이는 것과 연관지어 이해할 수 있다. 또한 slope변수와 oldpeak변수는 정반대의 방향으로 향하는 것을 확인할 수 있는데 이는 역시 2.3절에서 진행한 상관분석의 결과 slope변수와 oldpeak변수가 -0.58의 다소 높은 음의 상관계수를 가지므로 강한 음의 선형관계를 가지는 결과와 일치한다.

3.2 인자분석

인자분석은 변수들의 묶음마다 어떤 공통요인들이 선형적으로 결합되어 관찰되었다는 가정 하에서, 공통요인을 식별하는 것에 주안점을 두는 다변량 분석법이다. 인자분석에서 같은 그룹에 속하는 변수들의 상관성은 높고, 다른 그룹에 속하는 변수들끼리의 상관성은 낮다. 따라서 한 그룹에 속하는 변수들은 하나의 인자에 의해 관찰된 속성들로서, 이들의 높은 상관성은 잠재된 인자에 의해 설명된다고 할 수 있다. 따라서 인자분석을 통하여 변수들의 군집적 특성을 파악하는 것이 가능하다. 인자분석은 공분산행렬을 근사 또는 차원 축소한다는 의미에서는 주성분 분석과 유사하지만, 주성분은 관측된 변수들의 선형결합식으로 정의되는 반면, 인자분석에서는 변수들이 공통인자들의 선형결합식으로 정의된다는 측면에서 구별된다. 인자분석에서는 적절한 변환을 통해 인자에 대한 해석이 용이한 단순한 구조를 만들기 위해 회전변환을 시도한다. 회전변환의 종류로 직교회전인 Varimax 회전과 Quartimax 회전이 있으며 비직교회전을 이용한 Promax 회전 등이 있다.

본 분석에서는 주성분을 이용한 인자분석과 최대우도법을 이용한 인자분석을 사용하고, 인자의 해석을 위해 Varimax회전 방법과 Promax 회전 방법을 모색하도록 한다. 적절한 인자 개수를 구하기 위해 'H_0: 인자 개수는 4개이다.'에 대한 varimax 회전 인자분석의 카이제곱 검정 결과 p-value=0.00101 < 0.05이므로 유의수준 5%에서 주어진 귀무가설을 기각한다. 'H_0: 인자 개수는 5개이다.'에 대한 varimax 회전 인자분석의 카이제곱 검정 결과 p-value=0.143 > 0.05이므로 유의수준 5%에서 주어진 귀무가설을 기각할 수 없어 인자 5개인 모형이 적합하다. 그림 3.3의 스크리 그래프에서도 인자 개수 5개 이후 그래프의 기울기에 거의 변화가 없는 것을 보아 인자 개수는 5개가 적합하다고 볼 수 있다.

그림 3.3 인자 개수에 대한 스크리 그래프

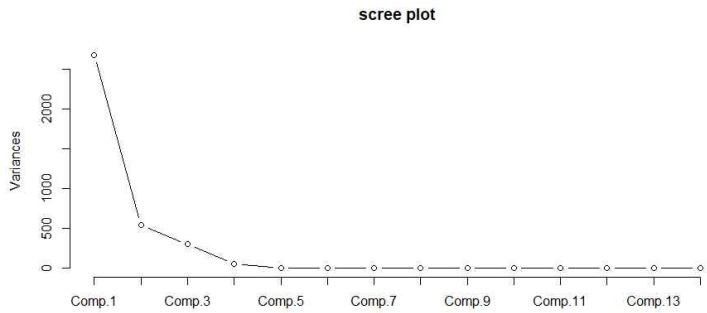


표 3.4 최대우도법을 이용한 인자적재 값

	Factor1	Factor2	Factor3	Factor4	Factor5	공통성	특수성
age	0.225	-0.353	0.127	0.496	0.252	0.501	0.499
sex	0.690	0.720				0.995	0.005
cp	-0.249	0.169		-0.192	0.635	0.538	0.462
trestbps			0.243	0.328	0.213	0.218	0.782
chol	-0.122	-0.158	0.109	0.345		0.172	0.828
fbs				0.203	0.197	0.083	0.917
restecg				-0.220		0.06	0.940
thalach	-0.750	0.657				0.995	0.005
exang	0.371	-0.158	0.196	0.107	-0.365	0.346	0.654
oldpeak	0.322	-0.172	0.835			0.837	0.163
slope	-0.306	0.249	-0.511	0.115		0.431	0.569
ca	0.235		0.175	0.327		0.197	0.803
thal	0.211		0.204	0.153	-0.162	0.144	0.856
누적 분산 설명량	0.124	0.221	0.311	0.369	0.424		

최대우도법을 이용한 인자분석 결과를 살펴보도록 한다. 추출된 요인에 의해 각 변수의 설명력을 의미하는 공통성을 살펴보면 나이(age), 성별(sex), 가슴 통증 유형(cp), 최대 심박수(thalach), 휴식과 관련된 운동으로 인한 ST 우울증(oldpeak)변수가 0.5이상의 값을 가지어 설명력이 있다고 할 수 있다. Factor1에서는 성별(sex)의 요인적재 값이 크게 나타나고 Factor2에서는 성별(sex)과 최대 심박수(thalach)의 요인 적재 값이 크게 나타난다. Factor3에서는 휴식과 관련된 운동으로 인한 ST 우울증(oldpeak)의 요인 적재값이 크게 나타나며 Factor4에서는 나이(age)의 요인 적재 값이 크게 나타나며 Factor5에서는 가슴 통증 유형(cp)의 요인 적재 값이 크게 나타난다. Factor1은 “성별관련 인자”, Factor2는 “성별과 최대 심박수 관련 인자”, Factor3는 “휴식과 관련된 운동으로 인한 ST우울증관련 인자”, Fator4는 “나이관련 인자”, Factor5는 “가슴 통증 유형과 관련된 인자”라고 설명할 수 있다.

표 3.5 varimax 회전과 promax회전변환 후 인자적재 값

	varimax 회전					promax회전				
	Factor1	Factor2	Factor3	Factor4	Factor5	Factor1	Factor2	Factor3	Factor4	Factor5
age		-0.120		-0.328	0.605			0.675	-0.254	-0.152
sex		0.988	0.135			-0.137		-0.108		0.665
cp			-0.723	0.105		0.762		0.158		0.164
trestbps	0.154				0.432		-0.102	0.445	0.112	
chol		-0.228	0.140		0.312	-0.195		0.305	0.141	-0.342
fbs					0.266	0.149		0.283		0.101
restecg					-0.229			-0.188		
thalach	-0.269		-0.271	0.917		-0.507	-0.109	-0.201	0.907	
exang	0.212		0.506	-0.197		-0.135	-0.636		-0.114	
oldpeak	0.879		0.178		0.169		0.856			
slope	-0.611		-0.113	0.203		-0.266	0.138	0.385		0.139
ca	0.117		0.214		0.348	-0.348			0.108	0.167
thal	0.149	0.168	0.274		0.132	0.663		-0.155		-0.196

varimax회전변환 후 인자분석 결과를 살펴보도록 한다. Factor1은 휴식과 관련된 운동으로 인한 ST 우울증(oldpeak)의 요인 적재 값이 크게 나타나며 Factor2는 성별(sex)과 관련된 요인 적재 값이 크게 나타난다. Factor3는 운동 유발 협심증(exang)과 관련된 요인 적재 값이 크게 나타나고 Factor4는 최대 심박수(thalach)의 요인 적재 값이 크게 나타나고 Factor5에서는 나이(age)의 요인 적재 값이 크게 나타난다. Factor1은 “휴식과 관련된 운동으로 인한 ST 우울증과 관련된 인자”, Factor2는 “성별과 관련된 인자”, Factor3는 “운동 유발 협심증과 관련된 인자”, Factor4는 “최대 심박수와 관련된 인자”, Factor5는 “나이와 관련된 인자”라고 설명할 수 있다.

promax회전변환 후 인자분석 결과를 살펴보도록 한다. Factor1은 가슴 통증 유형(cp)와 탈라세미아라고 불리는 혈액 장애(thal)의 요인 적재 값이 크게 나타나며 Factor2는 휴식과 관련된 운동으로 인한 ST 우울증(oldpeak)의 요인 적재 값이 크게 나타난다. Factor3는 나이(age)의 요인 적재 값이 크게 나타나고 Factor4는 최대 심박수(thalach)의 요인 적재 값이 크게 나타나고 Factor5는 성별(sex)의 요인 적재 값이 크게 나타난다. Factor1은 “혈액 장애에 의한 가슴 통증 유형과 관련된 인자”, Factor2는 “휴식과 관련된 운동으로 인한 ST 우울증과 관련된 인자”, Factor3는 “나이와 관련된 인자”, Factor4는 “최대 심박수와 관련된 인자”, Factor5는 “성별과 관련된 인자”라고 설명할 수 있다.

최대우도법, varimax회전변환, promax회전변환에서 공통적으로 나타난 인자는 “성별관련 인자”, “나이관련 인자”, “휴식과 관련된 운동으로 인한 ST우울증과 관련된 인자”임을 확인할 수 있다.

3.3 로지스틱 회귀분석

판별분석은 정규분포를 가정하고 $P(Y=1|X=x)$를 추정하여 분류하는 방법으로서 자료의 생성모형, 즉 결합분포에 기반한 방법이다. 이에 비해 로지스틱 회귀모형은 $P(Y=1|X=x)$에 대하여 특정 분포를 가정하는 것이 아니라 특정한 연결함수를 이용하는 기술모형기반의 방법이다. 즉, (Y,X)의 결합분포를 정의가 없이도 사용이 가능하다. 로지스틱 회귀의 경우 분포에 대한 가정을 하지 않기 때문에 적용범위가 더 넓어 판별분석보다 널리 사용된다. 따라서 주어진 데이터의 출력변수인 target변수의 범주를 예측하기 위해 로지스틱 회

귀분석을 진행하도록 한다. 로지스틱 회귀에서는 대체손실함수로 로지스틱 손실함수 $\log(1+\exp(-Yf(X)))$를 사용한다. 로지스틱 회귀분석 과정에서 과대적합 문제를 피하기 위해 변수선택을 실시한다. 변수선택이 필요한 이유는 설명력이 없는 변수들이 모형에 포함되면 올바른 모형에 비해 예측오차가 증가할 수 있기 때문이다. 변수선택 방법은 전모형탐색법, 전진선택법, 후진소거법 중, AIC를 이용한 전진선택법(forward selection)을 사용하도록 한다. AIC는 모형의 복잡도에 벌점을 주는 방법으로서 모든 후보 모형들에 대해 AIC가 최소가 되는 모형을 선택한다. BIC는 AIC에 비해 더 단순한 모형을 선택하지만 AIC의 예측오차가 더 좋기에 본 분석에서는 AIC를 사용하도록 한다.

분석에 앞서 주어진 데이터는 변수의 단위가 다르기 때문에 표준화과정을 거치도록 한다. 또한 데이터의 예측력을 검정하기 위해서 데이터를 학습데이터와 검증데이터를 같은 비율로 나누어 분석을 진행하였다.

그림 3.4 로지스틱 회귀분석 결과

```
Call:
glm(formula = target ~ oldpeak + ca + thal + thalach + sex +
    cp + chol, family = binomial, data = tr.h)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-2.53567  -0.21594   0.08917   0.33364   2.99983

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.9722     1.8868   3.165 0.001550 **
oldpeak      -6.9811     1.9479  -3.584 0.000339 ***
ca           -5.4233     1.2525  -4.330 1.49e-05 ***
thal         -6.2435     1.5398  -4.055 5.02e-05 ***
thalach       6.9225     2.0352   3.401 0.000670 ***
sex          -2.5487     0.7952  -3.205 0.001351 **
cp            1.8619     0.8374   2.223 0.026189 *
chol         -6.0231     3.0711  -1.961 0.049856 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 220.45  on 160  degrees of freedom
Residual deviance:  80.74  on 153  degrees of freedom
AIC: 96.74

Number of Fisher Scoring iterations: 7
```

AIC를 이용한 전진선택법에 의한 로지스틱 회귀분석의 결과를 살펴보면 출력변수인 target변수를 제외한 13개의 입력변수 중에서 oldpeak, ca, thal 등의 7개의 변수만을 선택하였다. 선택된 7개의 변수들은 유의수준 5%에서 모두 유의한 결과를 보이고 있다.

표 3.6 로지스틱 회귀 오분류표

실제\예측	0	1
0	46	22
1	12	62

표 3.6은 주어진 학습자료에 대해서 추정된 사후확률 $\hat{P}(y=1|x)$과 절단값 0.5를 이용하여 구축된 분류기 $\widehat{C_c}(x)$에 대해서 작성된 오분류표이다. 정분류율은 0.67, 오분류율은 0.329, 민감도와 특이도는 각각 0.837, 0.676이다. 민감도는 $y=1$인 클래스에 속한 자료 중 정분류된 자료의 비율이며, 특이도는 $y=0$인 클래스에 속한 자료 중 정분류된 자료이다. 즉, 심장병 여부를 분류하는 문제에서 심장병이 있는 사람 중 약 84%를 심장병이 있다고 예측할 것이며, 심장병이 없는 사람 중 약 32%를 심장병이 있다고 예측할 것이다.

민감도와 특이도는 절단값에 따라 달라지지만 동시에 크게 할 수 없다. ROC(Receiver Operating Characteristic)곡선과 ROC 곡선아래의 면적인 AUC(Area Under the Curve)를 통해 여러 절단값에서의 민감도와 특이도의 관계를 살펴보면 예측력이 가장 좋은 모형을 선택할 수 있다. 따라서 본 분석에서는 앞으로 진행할 신경망 모형, 의사결정나무를 사용한 앙상블 모형의 ROC 곡선과 AUC를 모두 비교하여 예측력이 가장 우수한 모형을 선택하도록 한다.

3.4 신경망

신경망은 인간의 두뇌구조를 모방한 입력값과 출력값 간의 매우 복잡한 형태의 비선형 지도학습법이다. 신경망 모형은 예측력이 좋으나 복잡성으로 인하여 해석이 어렵고 모형 구축시 고려해야할 사항이 많다. 입력자료의 선택의 민감성을 피하기 위해 표준화 과정을 진행하여 입력변수 값들의 범위가 변수 간에 큰 차이가 없도록 하고 범주형 변수는 같은 범위를 갖도록 가변수화하였다. 신경망은 입력층(input layer)과 출력층(output layout)으로만 이루어진 단층신경망과 입력층, 은닉층(hidden layer), 출력층으로 이루어진 다층신경망이 있다. 입력층은 각 입력변수에 대응되는 노드로 구성되고 은닉층은 입력층으로부터 전달받은 변수 값들의 선형결합을 비선형함수로 처리하여 출력층 또는 다른 은닉층으로 전달한다. 출력층은 출력변수에 대응되는 노드로 클래수의 수만큼 출력노드가 생성된다. 본 분석에서는 은닉층을 포함한 다층신경망을 적합하였다. 은닉층은 대체로 하나일 때 모든 매끄러운 함수를 근사적으로 표현할 수 있다고 알려있으므로 은닉층은 하나로 설정하였고, 적절한 은닉노드의 수를 결정하기 위해 은닉노드의 수가 2개부터 5개인 경우에 대하여 오분류율을 비교하였다.

그림 3.5 은닉노드 개수에 따른 오분류율

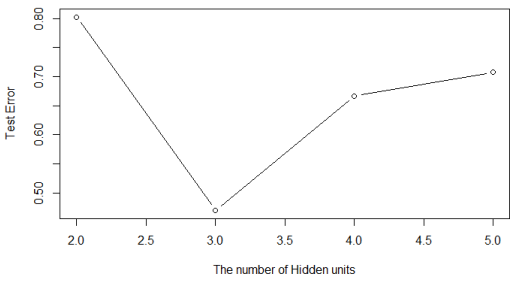


그림 3.6 은닉노드 3개인 다층신경망 적합 결과

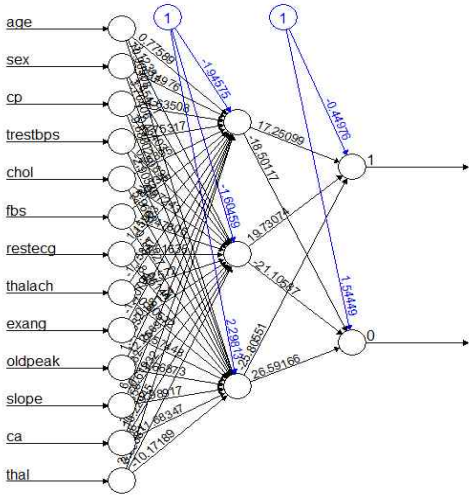


그림 3.5를 보면 오분류율 값은 은닉노드가 2개일 때 0.8027, 3개일 때 0.4694, 4개일 때 0.6666, 5개일 때 0.7074으로 오분류율이 최소가 되는 은닉노드 수는 3개임을 확인하였다. 따라서 은닉층이 1개, 은닉노드가 3개인 다층신경망 모형에 학습데이터를 적합한 결과가 그림 3.5와 같다.

표 3.6 신경망 오분류표

실제\예측	0	1
0	52	9
1	18	70

은닉노드 수가 2개인 신경망 모형의 오분류표는 표 3.6과 같다. 정분류율은 0.818, 오분류율은 0.181, 민감도와 특이도는 각각 0.795, 0.852이다. 즉, 심장병 여부를 분류하는 문제에서 심장병이 있는 사람 중 약 80%를 심장병이 있다고 예측할 것이며, 심장병이 없는 사람 중 약 15%를 심장병이 있다고 예측할 것이다. 이 결과는 3.5절의 로지스틱 회귀분석의 결과와 비교하면 심장병이 없는 사람을 심장병이 있다고 오분류하는 False positive의 비율은 감소했으나, 심장병이 있는 사람을 없다고 오분류하는 False negative의 값이 로지스틱 회귀의 경우 0.162였으나 신경망의 경우 0.204로 증가하였음을 확인할 수 있다. 분류에서 False positive보다 False negative가 더 위험하므로 정분류율과 오분류율 외에도 False negative의 비율도 함께 비교해야함에 주의한다.

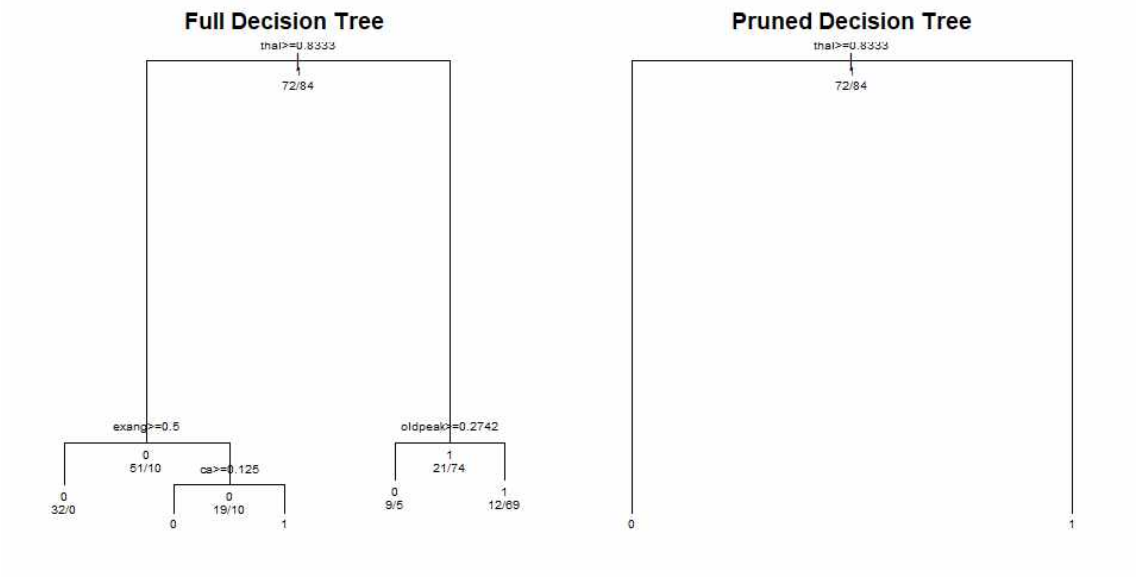
3.5 의사결정나무를 이용한 앙상블 기법

3.5.1 의사결정나무

의사결정나무는 주어진 입력값에 대하여 출력값을 예측하는 모형으로서 각 변수의 영역을 반복적으로 분할함으로써 전체 영역에서의 규칙을 생성하는 지도학습기법이다. 의사결정

나무는 결과를 나무형태의 그래프로 표현하여 다른 지도학습법에 비해 예측력은 떨어지나 해석력이 좋다. 의사결정나무는 모형에 대한 가정이 필요 없는 비모수적인 방법으로 연속형 변수와 범주형 변수를 모두 취급할 수 있다. 의사결정나무의 형성과정은 크게 성장(growing), 가지치기(prunning), 타당성 평가, 해석 및 예측의 순서로 이루어진다.

그림 3.6 의사결정나무 나무모형



의사결정나무 역시 3.3절의 로지스틱 회귀분석과 3.4절의 신경망 모형과 같이 입력변수를 표준화하고 데이터를 학습데이터와 검증데이터를 같은 비율로 나누어 진행하였다. 의사결정나무는 학습데이터를 가지고 복잡도를 0으로하여 최대 모형을 생성하고 10묶음 교차확인오차를 진행하여 적합하였다. 가지치기는 최대 모형에서 교차확인오차가 최솟값을 갖는 복잡도 값을 가지고 가지치기를 진행하였다.

그림 3.6은 가지치기 이전의 최대 나무모형과 가지치기 이후의 나무모형의 결과이다. 가지치기 이전의 나무모형을 살펴보면 의사결정나무의 분류에서 가장 중요한 변수로 thal변수를 선택하여 1차 분류를 하였고, exang변수와 oldpeak변수로 2차 분류를, ca변수로 3차 분류를 통해 최종적으로 5개의 그룹으로 분류하였다. 가지치기 이후의 나무모형은 thal변수를 기준으로 그 값이 0.8333보다 크거나 같으면 0으로, 작으면 1로 분류하여 최종적으로 2개의 그룹으로 분류하였다.

표 3.7 의사결정나무 오분류표

실제\예측	0	1
0	38	28
1	18	63

가지치기를 진행한 의사결정나무 모형의 오분류표는 표 3.7과 같다. 정분류율은 0.687, 오분류율은 0.3129, 민감도와 특이도는 각각 0.777, 0.576이다. 즉, 심장병 여부를 분

류하는 문제에서 심장병이 있는 사람 중 약 78%를 심장병이 있다고 예측할 것이며, 심장병이 없는 사람 중 약 42%를 심장병이 있다고 예측할 것이다. 또한 False negative의 값이 0.222로 앞서 진행한 로지스틱 회귀분석과 신경망 모형에 비해 값이 가장 큰 것을 확인할 수 있다.

3.5.2 앙상블 기법

의사결정나무는 복잡한 나무모형에 대해서는 분산이 매우 큰 불안정한 방법이다. 의사결정나무의 분산을 줄이기 위해 배깅, 랜덤포레스트, 부스팅과 같은 앙상블 기법을 사용한다. 앙상블이란 주어진 자료로부터 여러 개의 예측모형들을 만든 후 이러한 예측모형들을 결합하여 하나의 최종 예측모형을 만드는 방법이다. 본 분석에서는 3.6.1절에서 진행한 의사결정나무 모형으로 배깅, 랜덤포레스트, 부스팅의 앙상블 기법을 활용하도록 한다.

(1) 배깅(Bootstrap Aggreating)

배깅(Bootstrap Aggreating)은 불안정한 예측모형에서 불안정성을 제거함으로써 예측력을 향상시키는 모형으로 주어진 자료에 대하여 여러 개의 붓스트랩(bootstrap)자료를 생성하고 각 붓스트랩 자료에 예측모형을 만든 후 결합하여 최종 예측모형을 만드는 방법이다. 배깅 예측모형은 k묶음 교차확인오차를 진행하지 않고 복잡도가 0, 각 노드에 최소 5개의 관측값 포함, 최종 모형까지 최대 10개의 노드를 가진 최대 나무모형을 복원추출로 50개 생성하여 예측 모형을 생성하였다.

그림 3.7 배깅 모형의 주요 변수

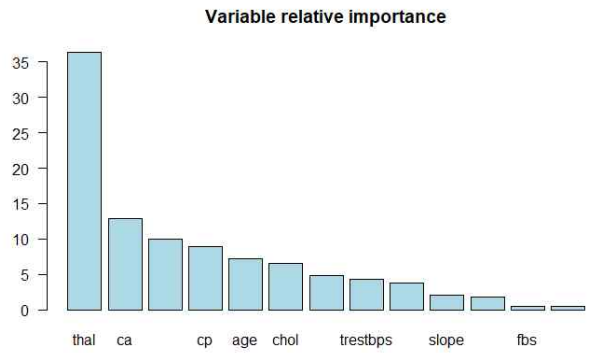


그림 3.7은 배깅 모형 생성 후 주요 변수들을 출력한 결과이다. thal변수(탈세미아라고 불리는 유전 혈액장애)의 중요도가 36.307로 가장 높았고, ca변수(플러로소피에 의해 색칠된 주요 혈관 수)가 12.876, oldpeak변수(휴식과 관련된 운동으로 인한 ST 우울증)가 10.047, cp변수(가슴 통증 유형)가 8.915로 중요한 변수로 출력되었다.

표 3.7 배깅 오분류표

실제\예측	0	1
0	52	14
1	19	62

배깅 모형의 오분류표는 표 3.7과 같다. 정분류율은 0.776, 오분류율은 0.224, 민감도와 특이도는 각각 0.765, 0.788이다. 즉, 심장병 여부를 분류하는 문제에서 심장병이 있는 사람 중 약 77%를 심장병이 있다고 예측할 것이며, 심장병이 없는 사람 중 약 21%를 심장병이 있다고 예측할 것이다. 또한 False negative의 값이 0.234로 앞서 진행한 로지스틱 회귀분석, 신경망 모형, 의사결정나무 모형에 비해 값이 가장 큰 것을 확인할 수 있다.

(2) 부스팅(boosting)

부스팅(boosting)은 예측력이 약한 모형들을 결합하여 강한 예측모형을 만드는 방법으로 일반적으로 앙상블 방법중 예측력이 가장 우수하다. 부스팅 예측모형은 k묶음 교차확인 오차를 진행하지않고 복잡도 0, 최종 모형까지 최대 5개의 노드를 가진 최대 나무모형을 복원추출로 50번 반복하며 가중치를 배분하였다.

그림 3.8 부스팅 모형의 주요 변수

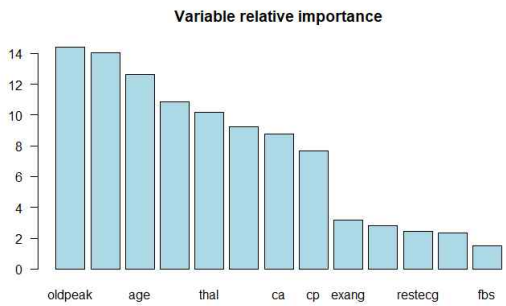


그림 3.8은 부스팅 모형 생성 후 주요 변수들을 출력한 결과이다. oldpeak변수(휴식과 관련된 운동으로 인한 ST 우울증)가 14.391로 가장 중요한 변수로 출력되었고, chol변수(혈청 콜레스테롤) 14.044, age변수(나이)가 12.621, thal변수(탈세미아라고 불리는 유전 혈액장애)의 중요도가 10.171로 중요한 변수로 출력되었다.

표 3.8 부스팅 오분류표

실제\예측	0	1
0	51	15
1	19	62

부스팅 모형의 오분류표는 표 3.8과 같다. 정분류율은 0.769, 오분류율은 0.231, 민감도와 특이도는 각각 0.765, 0.773이다. 즉, 심장병 여부를 분류하는 문제에서 심장병이 있는 사람 중 약 77%를 심장병이 있다고 예측할 것이며, 심장병이 없는 사람 중 약 23%를 심장병이 있다고 예측할 것이다. 또한 False negative의 값이 0.235로 앞서 진행한 로지스틱 회귀분석, 신경망 모형, 의사결정나무 모형, 배깅 모형에 비해 값이 가장 큰 것을 확인할 수 있다.

(3) 랜덤포레스트(random forest)

랜덤포레스트(random forest)는 배깅보다 더 많은 무작위성을 주어 최종학습기를 생성하는 방법이다. 랜덤포레스트는 무작위성을 최대로 주기 위해 붓스트랩과 더불어 입력변수들에 대한 무작위 추출을 진행한다. 랜덤포레스트 모형은 13개의 변수중 5개의 변수를 선택하여 100개의 의사결정나무를 만들어 적합하였다.

정확도에 따른 변수 중요도를 살펴보면 thal변수(탈세미아라고 불리는 유전 혈액장애)의 중요도가 10.223으로 가장 중요한 변수로 출력되었고 ca변수(플러로소피에 의해 색칠된 주요 혈관 수)가 9.84, cp변수(가슴 통증 유형)가 6.455로 중요한 변수로 출력되었다. 지니지수에 따른 변수 중요도는 thal변수(탈세미아라고 불리는 유전 혈액장애)의 중요도가 15.026으로 가장 중요한 변수로 출력되었고 oldpeak변수(휴식과 관련된 운동으로 인한 ST 우울증)가 10.405, cp변수(가슴 통증 유형)가 10.105로 중요한 변수로 출력되었다.

표 3.9 랜덤포레스트 오분류표

실제\예측	0	1
0	54	12
1	18	63

랜덤포레스트 모형의 오분류표는 표 3.8과 같다. 정분류율은 0.796, 오분류율은 0.204, 민감도와 특이도는 각각 0.777, 0.818이다. 즉, 심장병 여부를 분류하는 문제에서 심장병이 있는 사람 중 약 78%를 심장병이 있다고 예측할 것이며, 심장병이 없는 사람 중 약 18%를 심장병이 있다고 예측할 것이다. 또한 False negative의 값이 0.222로 앞서 진행한 의사결정나무모형의 False negative의 값이 비슷하다.

3.6 ROC Curve & AUC

적합한 모형들의 예측력을 비교하기 위해 앞서 3.4절에서 소개한 ROC 곡선과 AUC를 비교하여 로지스틱 회귀, 신경망, 앙상블 기법들 중 가장 예측력이 좋은 모형을 선택하도록 한다.

그림 3.9 ROC Curve

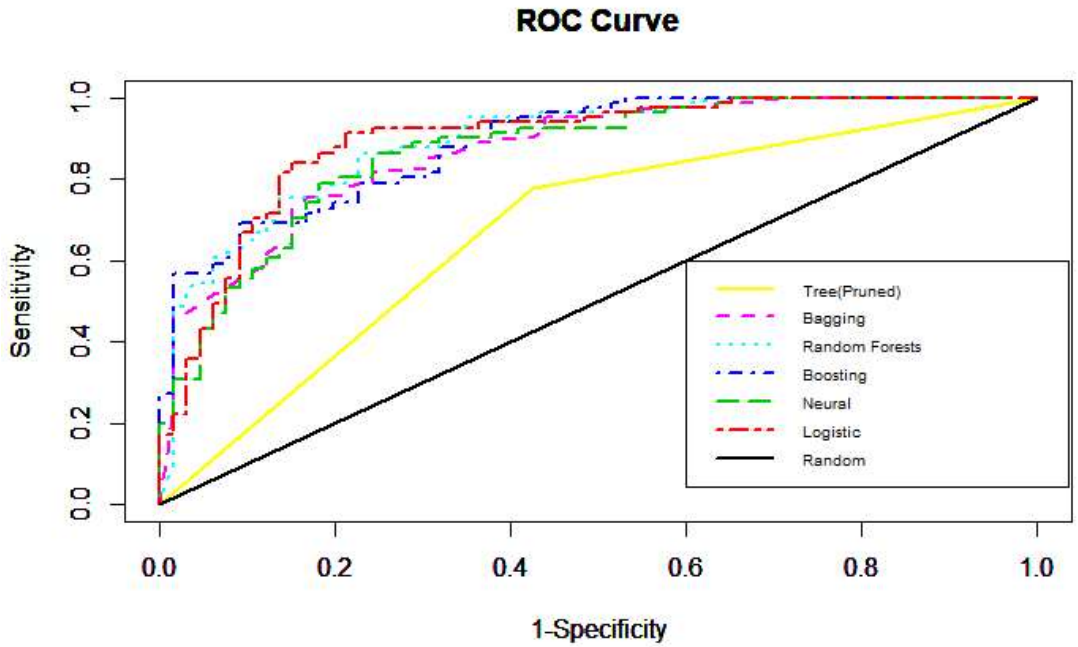


그림 3.7은 3.4절, 3.5절, 3.6절에서 진행한 로지스틱 회귀, 신경망, 의사결정나무를 이용한 앙상블 모형의 예측력을 ROC 곡선으로 나타낸 것이다. ROC 곡선을 보면 의사결정나무의 예측력이 가장 떨어지는 것을 확인할 수 있는데 이는 앙상블 기법을 통해 의사결정나무의 예측력이 향상되었다고 해석할 수 있다. 그러나 의사결정나무를 제외한 다른 모형들은 ROC 곡선으로는 비교가 어려우므로 AUC를 구해 비교하도록 한다.

표 3.10 AUC 비교

모형	로지스틱 회귀	신경망	의사결정나무	배깅	랜덤포레스트	부스팅
AUC	0.893	0.867	0.676	0.869	0.891	0.886

의사결정나무를 제외한 모형들의 AUC 값은 비슷하나 그 중 가장 큰 값을 갖는 모형은 로지스틱 회귀이다. 즉, 로지스틱 회귀, 랜덤포레스트, 부스팅, 배깅, 신경망, 의사결정나무 순으로 주어진 데이터의 출력변수인 target변수의 범주를 가장 잘 예측했다고 할 수 있다.

표 3.11 False negative 비교

모형	로지스틱 회귀	신경망	의사결정나무	배깅	랜덤포레스트	부스팅
FN	0.162	0.204	0.222	0.234	0.222	0.235

그러나 앞에서도 언급했듯이 오분류에서 False negative가 False positive보다 위험하므로 False negative의 값도 함께 참고하여 적절한 예측모형을 선택할 수 있어야한다. False negative의 값은 부스팅, 배깅, 의사결정나무-랜덤포레스트, 신경망, 로지스틱 회귀 순으로 높은 값을 가지는 것을 확인할 수 있다. 따라서 예측력은 높고 False negative는 낮은 로지스틱 회귀모형이 주어진 데이터의 출력변수 범주 예측에 가장 적합하다고 할 수 있다.

3.7 군집분석

군집분석은 적절한 기준에 대하여 동일한 군집에 속하는 관측값들을 서로 유사하도록 여러 개의 부분집합들로 할당하는 비지도학습법이다. 군집분석은 모집단 또는 범주에 대한 사전정보가 없는 자료에 대한 탐색적인 방법이므로 분석자의 주관에 따라 결과나 해석이 달라진다. 군집분석은 확률 모형에 기초한 모형인 가우스 혼합 모형과 확률 모형을 가정하지 않은 계층적 군집법과 비계층적 군집법(k-means)이 있다. 군집분석은 비슷한 자료끼리 군집을 형성하는 것 외에도 특이값을 갖는 개체의 발견이나 결측값의 보정 등에 활용될 수 있다. 본 군집분석에서는 확률 모형을 가정하지 않은 계층적 군집법과 비계층적 군집법을 실행하도록 한다. 확률 모형을 가정하지 않은 군집방법에서는 관측값들의 유사성을 측정하는 측도로 거리를 사용하는데 본 분석에서는 유클리드 거리를 사용하도록 한다. 또한 군집분석은 자료들 간의 거리를 이용하여 수행되기 때문에 자료의 단위가 결과에 큰 영향을 미친다. 따라서 각각의 변수에 대해 표준화를 실시한 후에 군집분석을 적용하도록 한다.

3.7.1 비계층적 군집분석(k-means)

비계층적 군집법은 주어진 군집수 k에 대해서 군집내 거리제곱합의 합을 최소화하는 것을 목적으로 한다. 비계층적 군집법은 계층적 군집법에 비하여 계산량이 적어 대용량 데이터를 빠르게 처리할 수 있다는 장점이 있지만 모든 변수가 연속형이어야 한다는 단점이 있다. 주어진 데이터의 변수는 연속형 변수와 범주형 변수로 되어있으므로 비계층적 군집법은 진행하지 않도록 한다.

3.7.2 계층적 군집분석

계층적 군집분석은 순차적으로 가까운 관측값들끼리 묶어주는 병합방법과 먼 관측값들을 나누어 가는 분할 방법이 있다. 병합방법에는 군집들 간의 거리를 측정하는 방법에 따라 최단연결(single linkage), 최장연결(complete linkage), 평균연결(average linkage), Ward 방법(ward method) 등이 있다. 최단연결법은 두 군집 사이의 거리를 각 군집에서 하나씩 관측값을 뽑았을 때 나타날 수 있는 모든 조합의 거리의 최솟값을 측정하는 방법이고 최장연결법은 두 군집 사이의 거리를 각 군집에서 하나씩 관측값을 뽑았을 때 나타날 수 있는 거리의 최대값으로 측정한다. 본 분석에서는 최단연결, 최장연결, 평균연결, Ward 방법을 모두 시행하여 비교하도록 한다.

그림 3.9 Average Silhouette methods

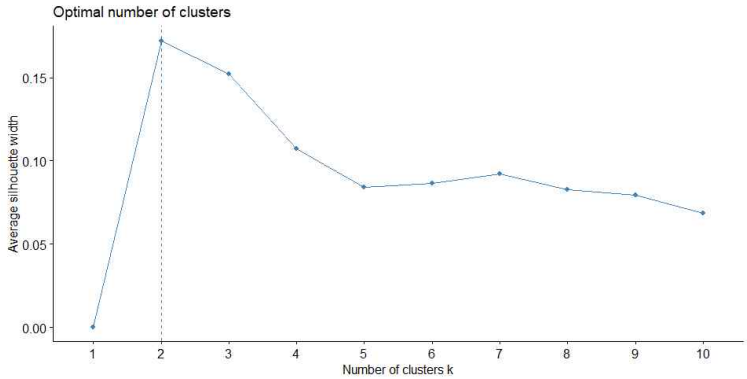


그림 3.9는 최적 군집수를 결정하기 위해 Average Silhouette 방법을 사용한 결과로 2개의 군집 개수가 적당함을 확인할 수 있다. 결과에 따라 계층적 군집분석의 군집수를 2개로 설정하여 분석을 진행하도록 한다.

그림 3.10 계층적 군집분석(최단연결, 최장연결, 평균연결, Ward)

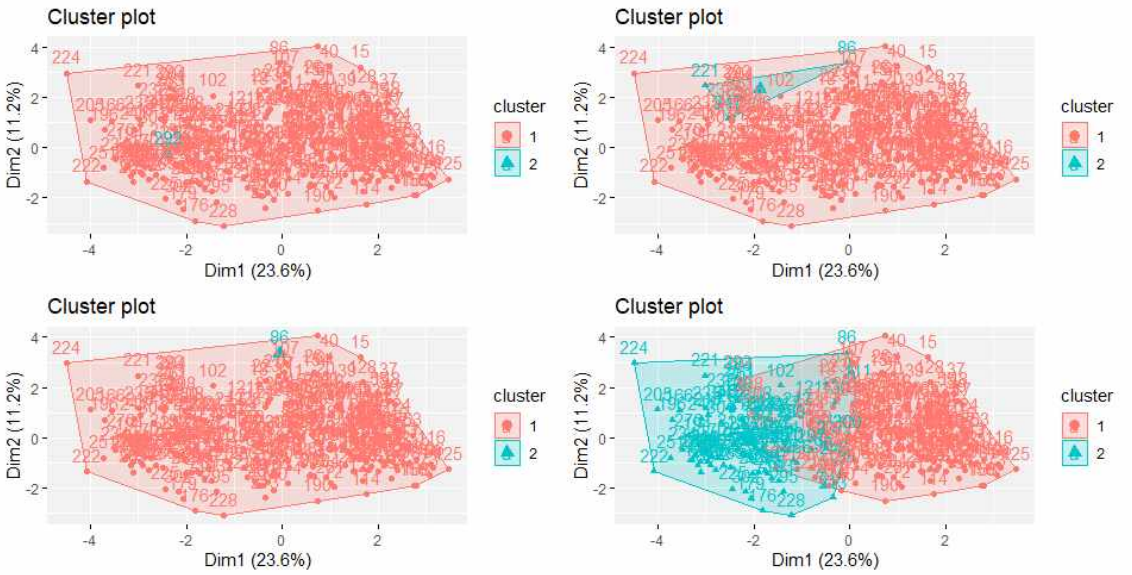


그림 3.10은 계층적 군집분석의 결과로 왼쪽 상단부터 차례대로 최단연결, 최장연결, 평균연결, ward 방법을 군집 수 2개로 분석한 결과이다. 표준화한 변수를 사용하였음에도 군집분석 결과가 좋지 않고 최단연결, 최장연결, 평균연결법은 다수의 경우를 군집1(target변수의 범주에서 '심장병이 없다')로 분류하였음을 확인할 수 있다. 비교적 ward방법이 다른 방법들에 비해 군집이 잘 나누어 진 것처럼 보이나 더 자세한 군집분석 결과의 예측력을 확인하기 위한 오분류율은 다음과 같다.

표 3.14 평균연결 오분류표

실제\예측	0	1
0	138	0
1	164	1

표 3.15 ward방법 오분류표

실제\예측	0	1
0	57	81
1	146	19

표 3.12 최단연결 오분류표

실제\예측	0	1
0	137	1
1	165	0

표 3.13 최장연결 오분류표

실제\예측	0	1
0	138	0
1	164	1

표 3.16 계층적 군집분석 결과 오분류율 비교

모형	최단연결	최장연결	평균연결	ward
오분류율	0.455	0.459	0.459	0.518

표 3.16을 보면 최단연결, 최장연결, 평균연결, ward방법 모두 오분류율이 0.4를 넘으며 예측력이 매우 좋지 않음을 확인할 수 있다. 표 3.15를 보면 ward방법은 두 그룹의 비율을 적절하게 분류하였으나 분류 결과의 예측력이 좋지 않아 오분류율이 높게 나옴을 확인할 수 있었다. 표 3.13, 3.14를 보면 최장연결, 평균연결의 경우 그룹1(target변수의 범주에서 '심장병이 없다')에 대해서는 완벽하게 분류했지만 그룹2(target변수의 범주에서 '심장병이 있다')에 대해서는 대부분의 값이 오분류되어 False negative가 높음을 확인할 수 있다.

전반적으로 계층적 군집분석의 예측력은 좋지 않으며 군집분류 결과도 좋지 않음을 확인할 수 있었다. 출력변수의 분류에 적합한 모형은 3.6절에서 진행한 모형들을 고려하는 것이 더 적절하게 보인다.

4. 결론

심장질환은 국내외적으로 주요 3대 사인으로 꼽히며 현재까지도 꾸준히 증가추세에 있으므로 분석의 필요성을 느끼고 상관분석, 주성분 분석, 인자분석, 로지스틱 회귀분석, 신경망 분석, 의사결정나무를 기반으로한 앙상블 기법과 군집분석을 진행하였다.

상관분석 결과 많은 변수들의 상관계수 값이 0에 가까운 것으로 보아 선형관계가 없는 것으로 보였고, slope변수와 oldpeak변수가 -0.58로 다소 높은 음의 상관계수를 갖는 것을 확인할 수 있었다. 이 결과는 주성분 분석의 결과에서도 확인할 수 있었다.

주성분 분석의 결과 심장병이 있는 그룹과 심장병이 없는 그룹, 전체의 그룹에서 공통적으로 slope변수와 oldpeak변수의 대비를 확인할 수 있었다. 즉, 피크 운동 ST 세그먼트의 기울기가 다운슬로핑일수록 휴식과 관련된 운동으로 인한 ST 우울증은 작은 값을 갖는다는 의미로 해석할 수 있다. 실제 심혈관 질환의 진단에서 피크 운동 ST 세그먼트의 기울기가 다운슬로핑인 경우 좌심실 비대증이 의심되고 플랫인 경우 허혈성 심장질환이 의심된다.

인자분석의 결과 최대우도법, varimax회전변환, promax회전변환에서 공통적으로 나타난 인자는 "성별관련 인자", "나이관련 인자", "휴식과 관련된 운동으로 인한 ST우울증과 관련된 인자"임을 확인할 수 있었으나 이는 차원축소에 큰 영향을 미치지 않는 결과로 해석되

어 향후의 분석에서 적용하지 않았다.

이후 3.4절부터 3.6.2절까지는 주어진 데이터의 출력변수인 target변수의 범주를 예측하기 위한 분석을 진행하였다. 분석 모형들이 예측의 주요변수로 뽑은 변수들은 각각 차이가 있었지만 공통적으로 모든 분석 모형들이 thal변수를 가장 많이 주요 변수로 뽑았고, 다음으론 oldpeak, cp, ca변수를 마지막으로 age, sex, chol변수를 주요 변수로 뽑았다. 즉, 심장병의 여부를 예측할 때 탈라세미아라고 불리는 유전 혈액장애의 여부와 휴식과 관련된 운동으로 인한 ST 우울증, 가슴 통증 유형, 플러로소피에 의해 색칠된 주요 혈관 수, 나이, 성별, 혈청 콜레스테롤 등이 중요함을 알 수 있다.

3.7절의 ROC 곡선과 AUC 결과를 확인한 결과, 로지스틱 회귀, 랜덤포레스트, 부스팅, 배깅, 신경망, 의사결정나무 순으로 주어진 데이터의 출력변수인 target변수의 범주를 가장 잘 예측하였다. 그러나 병의 진단에서 병이 있음에도 없다고 예측되는 경우(False negative)가 가장 위험하므로 예측력과 함께 False negative도 함께 비교해야 한다. 위 분석의 경우 로지스틱 회귀가 예측력도 가장 좋고 False negative도 가장 낮아 심장별 질환 예측에 가장 적합한 예측모형이라는 결론을 내릴 수 있었다. 그러나 병의 진단 예측은 정확도가 중요하므로 로지스틱 회귀의 예측성능을 더 높일 수 있는 방법의 연구가 필요해 보인다.

비계층적 군집분석의 결과 최단연결, 최장연결, 평균연결, ward방법 모두 예측성능이 좋지 않았다. 따라서 출력변수의 범주를 예측하는 모형으로는 3.4절에서 진행한 로지스틱 회귀분석을 진행하는 것이 더 바람직하게 보인다.

전반적인 분석에서 주목할 만한 결과를 살펴보면 일반적으로 남성의 경우가 여성보다 심장질환이 자주 발생한다는 것으로 알려져 있는 것과는 상반되게 여성의 심장병 비율이 더 높은 것을 확인할 수 있었다. 이는 주어진 데이터의 심장질환이 협심증, 심근경색, 동맥경화증, 심장판막증, 원발성 폐고혈압증 등 다양한 심장질환의 구분이 없기 때문으로 보인다. 협심증과 심근경색의 경우 중년이후의 남성에서 많이 볼 수 있으며 동맥경화증의 경우는 노화현상으로 모든 노인들에게서 볼 수 있고 원발성 폐고혈압증은 젊은 성인 여성에게 자주 발생하는 것으로 알려져있다. 본 연구에서 사용한 데이터는 모든 연령과 성별을 대상으로 초기 심장질환의 진단에 활용하기에 문제가 없으나 이후 다양한 심장질환의 데이터를 추가적으로 공급받아 구체적인 심장질환 병명의 진단에 활용하는 방안을 제안한다.

심장질환은 발병연령이 점차 감소하는 추세이며 뇌졸중, 돌연사로도 이어질 수 있으므로 조기진단이 매우 중요하다. 4차 산업혁명 시대에 빠른 속도로 발전하고 있는 IOT 기술과 AI를 활용하여 빠르고 간단하게 심장질환을 진단할 수 있는 기술을 개발한다면 심장병의 조기진단과 치료에 중요한 역할을 할 수 있을 것으로 기대하며 본 연구의 활용방향을 제시한다.

<참고문헌 및 사이트>

-김재희. (2011). R 다변량 통계분석. 교우사.
-덕성여자대학교 정보통계학과. (2019). 다변량 통계분석 모음집.
-박창이. (2015). R을 이용한 데이터 마이닝. 교우사.
-박평우. (2018). 허혈성 심장질환 진단을 위한 기계 학습 알고리즘 비교 연구. (국내석사학위논문).
-삼성서울병원 심장뇌혈관병원. 내분비/유방 질환 당뇨병.
http://www.samsunghospital.com/home/hbv/disease/info/view.do?CONT_ID=2690&CONT_SRC_ID=09a4727a8000f200&CONT_SRC=CMS&CONT_CLS_CD=001020001
-서울특별시 서울의료원. 심혈관 센터 질환안내. 심장질환. seoulmc.or.kr
-이금숙. (2016.10.25). [심장건강 길라잡이④-심장검진]심장병 조기 진단 가능한 '심전도 검사'아세요?. 헬스조선.
http://health.chosun.com/site/data/html_dir/2016/10/24/2016102401620.html
-통계청. 2019년 6월 24일. [2018년 사망원인 통계]
-WHO. (2018). The top 10 causes of death.
https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death
-ZHEXUE HUANG. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values.