The Journal of the Convergence on Culture Technology (JCCT)
Vol. 8, No. 4, pp.339-345, July 31, 2022. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2022.8.4.339

JCCT 2022-7-42

데이터마이닝을 이용한 심혈관질환 판별 모델 방법론 연구

A study of methodology for identification models of cardiovascular diseases based on data mining

이범주*

Bum Ju Lee*

요 약 심혈관 질환은 전 세계적으로 주요 사망원인들 중 하나이다. 본 연구는 보다 우수한 심혈관질환 판별 모델을 생성하기 위한 방법에 대한 연구로써, 3가지 변수 선택법과 7가지 머신러닝 알고리즘을 바탕으로 사회인구학적 변수들을 이용하여 고혈압과 이상지질혈증 판별모델들을 생성하고, 생성된 모델들의 성능을 비교 평가한다. 본 연구의 결과에서는 두 가지 질병 모두에서, 전체변수 및 correlation-based feature subset selection 메소드 기반 모델들에서는 naive Bayes 모델이 다른 머신러닝을 이용한 모델들보다 다소 우수한 판별 성능이 있는 것으로 나타났고, wrapper 메소드 기반 변수 선택법에서는 logistic regression 모델이 다른 모든 모델보다 성능이 다소 우수한 것으로 나타났다. 본 연구의 결과는 원격의료 및 대중보건 분야에서 향후 한국인의 심혈관질환 판별 및 예측 모델 생성을 위한 참고자료로 활용될 수 있을 것으로 기대된다.

주요어 : 심혈관질환, 판별모델, 데이터마이닝, 방법론, 머신러닝

Abstract Cardiovascular diseases is one of the leading causes of death in the world. The objectives of this study were to build various models using sociodemographic variables based on three variable selection methods and seven machine learning algorithms for the identification of hypertension and dyslipidemia and to evaluate predictive powers of the models. In experiments based on full variables and correlation-based feature subset selection methods, our results showed that performance of models using naive Bayes was better than those of models using other machine learning algorithms in both two diseases. In wrapper-based feature subset selection method, performance of models using logistic regression was higher than those of models using other algorithms. Our finding may provide basic data for public health and machine learning fields.

Key words : Cardiovascular Diseases, Identification Model, Data Mining, Methodology, Machine Learning

Ⅰ. 서 론

심혈관질환 (cardiovascular diseases)은 전 세계적으로 주요 사망 원인중 하나이다 [1]. 심혈관질환으로 인한 사망과 입원환자는 고령화에 따라 한국에서도 지난 십년간 지속적으로 증가하는 추세이며, 이러한 심혈관질환의 위험요인으로 고혈압 (hypertension), 이상지질혈증(dyslipidemia), 당뇨 (diabetes), 비만 (obesity), 대사증후군 (metabolic syndrome)등이 고려되고 있으며 이러한 질병들에 대한 지속적인 관리를 제안하고 있다 [1, 2].

*정회원, 한국한의학연구원 책임연구원 (단독저자)
접수일: 2022년 5월 24일, 수정완료일: 2022년 6월 21일
게재확정일: 2022년 7월 2일

Received: May 24, 2022 / Revised: June 21, 2022
Accepted: July 2, 2022
*Corresponding Author: bjlee@kiom.re.kr
Digital Health Research Division, Korea Institute of Oriental Medicine, Korea

한국에서도 고혈압과 이상지질혈증은 비만의 증가에 따라 그 유병율이 증가하는 추세이며 주요 심혈관 질환으로 고려되고 있다 [3, 4]. 고혈압과 다르게 이상지질혈증은 높은 triglyceride, LDL-cholesterol, triglyceride 레벨 그리고 낮은 HDL-cholesterol 레벨 중에서 한 가지 이상을 만족할 경우 이상지질혈증으로 판단한다 [5]. 고혈압과 이상지질혈증의 위험요인으로는 남자, 고령화, 가족력, 흡연, 비만 등의 매우 다양한 요인이 존재하고 있다 [3, 4, 5].

그동안 여러 위험요인들을 고려하여 머신러닝, 통계 또는 데이터마이닝 등을 기반으로 고혈압 및 이상지질혈증 판별 모델을 개발하는 연구들이 진행되어져 왔다 [6-9]. 고혈압 식별에 대한 선행연구에서 Lee et al [6]은 wrapper 기반 변수 선택법과 naive Bayes를 기반으로 사회인구학적 변수들만을 이용하여 남성 식별 모델을 생성한 후 0.850의 AUC 값과 0.495의 kappa값을 획득하였다. Kim et al. [7]은 나이브 베이시안 (naive Bayesian)을 이용하여 이상지질혈증의 위험요인들을 선별함과 더불어 발병을 예측하는 노모그램을 제시하였다. Lee et al. [8]의 연구에서는 한국성인의 심혈관질환과 흡연의 연관성을 Logistic regression을 이용하여 분석을 수행한 후, 중노년층의 고혈압 및 이상지질혈증의 위험요인에 대하여 흡연의 위험도를 평가하였다. Tang et al. [9]은 중국인을 대상으로 이상지질혈증 예측모델을 개발하기 위하여 random survival forest 알고리즘을 이용하였고, 남성에서는 0.731, 여성에서는 0.801의 C-statistics 값을 획득하였다.

지금까지, 이러한 심혈관질환 판별 또는 예측 모델의 선행연구들이 수행되었음에도 한국인을 대상으로 고혈압 및 이상지질혈증에 대한 다양한 모델 생성방법론에 대한 연구는 아직까지 미미한 실정이다. 따라서, 본 연구에서는 한국인 중노년을 대상으로 고혈압과 이상지질혈증에 대한 많은 판별 모델들을 생성하고, 생성된 모델들에 대한 성능 비교평가를 통하여 보다 우수한 질병 판별 모델 구현을 위한 방법론을 제시하고자 한다. 본 연구의 결과는 향후 한국인의 심혈관질환에 대한 대중보건의 발전에 기여할 것으로 판단된다.

II. 메소드

1. 데이터 셋

본 연구에서는 국민건강영양조사 2018년도 데이터를 기반으로 실험을 수행하였으며, 본 데이터는 총 7992 샘플들로 구성되어 있고, 그림 1과 같이 샘플 추출 과정을 통하여 최종 2941개의 샘플들을 도출하였다. 각 심혈관 질환에 사용된 샘플 수 및 변수들에 대한 기본 정보와 통계적 유의성은 표 1과 2에 기술하였다.

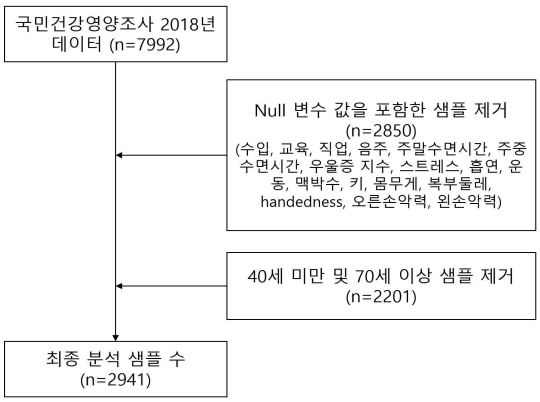

그림 1. 본 실험에 적용된 샘플 추출 프로세스
Figure 1. Process of sample extraction used in this study

2. 측정변수

모델생성을 위하여 본 연구에서는 사회인구학적 정보들만을 바탕으로 총 25가지 비침습적 변수들을 추출하였다. 25가지 변수는 두 가지 데이터 타입으로 구성되었다. 명목형 또는 범수형 변수로는 성별, 수입, 교육정도, 직업, 결혼유무, 음주, 스트레스, 흡연, 운동, 오른손잡이 또는 왼손잡이가 해당하고, 숫자형 변수로는 나이, 가족구성원수, 주중수면시간, 주말수면시간, 우울증지수, 맥박수, 수축기혈압, 이완기혈압, 키, 몸무게, 복부둘레, 복부와 키의 비율, 체질량지수, 오른손 악력, 왼손 악력이 해당한다.

3. 통계분석 및 모델링 기법

명목형 또는 범주형 변수 각각에 대한 통계적 유의성 분석을 위해서는 IBM SPSS (ver. 23) 소프트웨어를 이용한 교차분석이 수행되었고, 숫자형 변수의 유의성 도출을 위해서는 독립표본 t-test가 수행되었다.

변수들의 최적화를 바탕으로 다양한 모델 생성을 위해서 3가지 변수 선택법을 이용하였다. 3가지 변수선택법은 25개 변수 모두를 모델에 적용하는 full variables 메소드, 다중공선성 (multicollinearity)문제를 보완하기 위한 correlation-based feature subset selection (CFS) 메소드,

표 1. 고혈압 판별 모델에 사용된 샘플 및 변수들에 대한 통계분석 결과
Table 1. The results of statistical analysis for samples and variables used in hypertension model

Variable	Category	Normal		Hypertension		p-value
		Number or mean	% or SD	Number or mean	% or SD	
Subjects		2213		728		
Sex	Men	898	30.5%	351	11.9%	<0.001
	Women	1315	44.7%	377	12.8%	
Income	1st quadrant	508	17.3%	189	6.4%	0.077
	2nd quadrant	561	19.1%	186	6.3%	
	3rd quadrant	558	19.0%	193	6.6%	
	4th quadrant	586	19.9%	160	5.4%	
Education	Elementary or below	239	8.1%	209	7.1%	<0.001
	Middle school	243	8.3%	123	4.2%	
	High school	850	28.9%	255	8.7%	
	College or above	881	30.0%	141	4.8%	
Occupation	Expert	370	12.6%	74	2.5%	<0.001
	Office worker	263	8.9%	45	1.5%	
	Service	379	12.9%	114	3.9%	
	Farmer or fisher	81	2.8%	55	1.9%	
	Blue-collar worker	297	10.1%	104	3.5%	
	Elementary	198	6.7%	74	2.5%	
	Unemployed	625	21.3%	262	8.9%	
Marriage	Married	2119	72.1%	701	23.8%	0.525
	Single	94	3.2%	27	.9%	
Drinking	No	984	33.5%	323	11.0%	0.964
	Yes	1229	41.8%	405	13.8%	
Stress	Low	1673	56.9%	559	19.0%	0.516
	High	540	18.4%	169	5.7%	
Smoking	No	1789	60.8%	580	19.7%	0.489
	Yes	424	14.4%	148	5.0%	
Activity	No	1300	44.2%	474	16.1%	0.002
	Yes	913	31.0%	254	8.6%	
Handedness	Right	1976	67.2%	640	21.8%	0.589
	Left	84	2.9%	31	1.1%	
	Both	153	5.2%	57	1.9%	
Age		52.98	8.432	59.45	7.617	<0.001
Family number		3.055	1.198	2.654	1.161	<0.001
SDWD		415.0	71.63	415.6	75.64	0.841
SDWE		451.5	82.23	437.9	86.17	<0.001
Depressive score		2.065	3.278	2.140	3.340	0.594
Pulse		17.50	2.207	17.67	2.250	0.068
SBP		116.6	15.82	127.6	15.37	<0.001
DBP		77.08	10.08	79.57	9.696	<0.001
Height		163.0	8.338	162.5	8.889	0.150
Weight		63.22	11.10	67.42	11.29	<0.001
WC		81.46	9.101	87.32	8.823	<0.001
WHtR		0.500	0.052	0.538	0.054	<0.001
BMI		23.70	3.141	25.47	3.295	<0.001
Right grip strength		26.88	9.520	26.92	9.735	0.908
Left grip strength		26.00	9.159	26.17	9.515	0.667

(SD: standard deviation, WHtR: waist-to-height ratio, BMI: body mass index, WC: waist circumference, DBP: diastolic blood pressure, SBP: systolic blood pressure, SDWD: sleep duration during weekday, SDWE: sleep duration during weekend)

그리고 변수 선택과정이 black box형태로 진행되는 wrapper-based feature subset selection (wrapper) 메소드로 구성된다. CFS 방식은 단 한 번의 변수 선택을 통하여 추출된 변수리스트를 다양한 머신러닝 알고리즘에 동일하게 적용하였다. Wrapper 방식은 변수 선택시 각각의 머신러닝 알고리즘을 적용하여 변수 리스트를

표 2. 이상지질혈증 판별 모델에 사용된 샘플 및 변수들에 대한 통계분석 결과
Table 2. The results of statistical analysis for samples and variables used in dyslipidemia model

Variable	Category	Normal		Dyslipidemia		p-value
		Number or mean	% or SD	Number or mean	% or SD	
Subjects		2259		682		
Sex	Men	969	32.9%	280	9.5%	0.394
	Women	1290	43.9%	402	13.7%	
Income	1st quadrant	540	18.4%	157	5.3%	0.684
	2nd quadrant	568	19.3%	179	6.1%	
	3rd quadrant	569	19.3%	182	6.2%	
	4th quadrant	582	19.8%	164	5.6%	
Education	Elementary or below	281	9.6%	167	5.7%	<0.001
	Middle school	252	8.6%	114	3.9%	
	High school	869	29.5%	236	8.0%	
	College or above	857	29.1%	165	5.6%	
Occupation	Expert	371	12.6%	73	2.5%	<0.001
	Office worker	258	8.8%	50	1.7%	
	Service	377	12.8%	116	3.9%	
	Farmer or fisher	101	3.4%	35	1.2%	
	Blue-collar worker	318	10.8%	83	2.8%	
	Elementary	191	6.5%	81	2.8%	
	Unemployed	643	21.9%	244	8.3%	
Marriage	Married	2156	73.3%	664	22.6%	0.027
	Single	103	3.5%	18	.6%	
Drinking	No	958	32.6%	349	11.9%	<0.001
	Yes	1301	44.2%	333	11.3%	
Stress	Low	1736	59.0%	496	16.9%	0.027
	High	523	17.8%	186	6.3%	
Smoking	No	1809	61.5%	560	19.0%	0.240
	Yes	450	15.3%	122	4.1%	
Activity	No	1362	46.3%	412	14.0%	0.956
	Yes	897	30.5%	270	9.2%	
Handedness	Right	2005	68.2%	611	20.8%	0.471
	Left	86	2.9%	29	1.0%	
	Both	168	5.7%	42	1.4%	
Age		53.34	8.648	58.69	7.520	<0.001
Family number		3.017	1.202	2.752	1.178	<0.001
SDWD		415.0	72.02	415.7	74.66	0.838
SDWE		450.0	82.82	442.0	85.11	0.028
Depressive score		1.978	3.181	2.433	3.621	0.003
Pulse		17.50	2.274	17.68	2.019	0.060
SBP		118.5	16.62	122.0	15.41	<0.001
DBP		77.71	10.29	77.64	9.204	0.858
Height		163.3	8.419	161.6	8.559	<0.001
Weight		63.79	11.26	65.84	11.26	<0.001
WC		81.96	9.367	86.05	8.715	<0.001
WHtR		0.502	0.054	0.533	0.052	<0.001
BMI		23.84	3.214	25.13	3.262	<0.001
Right grip strength		27.08	9.532	26.25	9.685	0.046
Left grip strength		26.25	9.199	25.36	9.377	0.027

(SD: standard deviation, WHtR: waist-to-height ratio, BMI: body mass index, WC: waist circumference, DBP: diastolic blood pressure, SBP: systolic blood pressure, SDWD: sleep duration during weekday, SDWE: sleep duration during weekend)

추출하였고, 추출된 변수 리스트를 다시 동일한 머신러닝에 적용하여 모델을 생성하였다. 그림 2는 고혈압과 이상지질혈증 질병 판별 모델 생성을 위해 사용된 3가지 변수 선택법들과 7가지 머신러닝 알고리즘들, 주요 성능평가지표와 세부 성능지표, 그리고 실험순서 등에 대하여 본 연구의 전체적인 실험 디자인에 대한 구체적인

내용을 제시하고 있다.

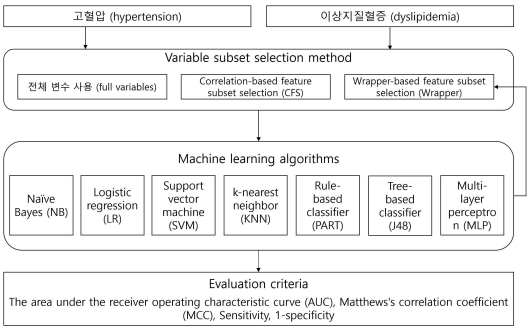

그림 2. 심혈관질환 판별 모델 생성 및 평가를 위한 실험 디자인
Figure 2. Experiment design to build and evaluate identification models of cardiovascular diseases

III. 실험 및 결과

1. 본 실험에 사용된 변수의 통계분석 결과

표 1과 2는 본 실험에 사용된 모든 변수들에 대한 통계적 유의성 분석 결과를 나타낸다. 고혈압 환자군과 정상군사이의 crude 분석에서 25가지 변수들 중 sex, education, occupation, activity, age, family number, SDWE, SBP, DBP, weight, WC, WHtR, BMI 변수가 두 그룹간에서 통계적 유의성이 존재하였다 (p = <0.05). 이상지질혈증의 crude 분석에서는 education, occupation, marriage, drinking, stress, age, family number, SDWE, depressive score, SBP, height, weight, WC, WHtR, BMI, right grip strength, left grip strength 변수가 두 그룹간에서 통계적 차이를 나타내었다 (p = <0.05).

2. 생성된 모델에 대한 성능평가 결과

표 3과 4는 본 실험에 사용된 3가지 변수 선택법과 7가지 머신러닝을 이용하여 생성된 모델들에 대한 성능평가 결과를 나타낸다.

고혈압 판별 모델에서, 전체 변수 (Full)를 이용한 모델 중에서 주요 평가지표 (MCC와 AUC)에서는 NB와 LR 모델이 다른 모델들보다 우수하였고, 세부성능지표 (sensitivity와 1-specificity)에서는 NB가 LR보다 다소 우수한 것으로 나타났다. CFS 기반 모델들에서도 NB와 LR이 가장 우수하였으나, 세부성능에서는 NB가 LR보다 우수하였다. Wrapper 기반 모델들에서는 LR 모델이 다른 모든 모델들보다 우수한 성능을 나타내었다.

표 3. 고혈압 판별 모델 성능 평가 결과
Table 3. Results of performance evaluation of identification models of hypertension

	Algo.	Class	Sens.	1-sp.	MCC	AUC
Full	NB	Normal	0.801	0.435	0.348	0.771
		Hyper.	0.565	0.199		
	LR	Normal	0.918	0.718	0.255	0.776
		Hyper.	0.282	0.082		
	SVM	Normal	0.973	0.915	0.126	0.529
		Hyper.	0.085	0.027		
	KNN	Normal	0.78	0.661	0.119	0.562
		Hyper.	0.339	0.22		
	PART	0	0.791	0.603	0.186	0.607
		Hyper.	0.397	0.209		
	J48	Normal	0.84	0.641	0.210	0.589
		Hyper.	0.359	0.16		
	MLP	Normal	0.824	0.589	0.240	0.704
		Hyper.	0.411	0.176		
CFS	NB	Normal	0.835	0.495	0.339	0.781
		Hyper.	0.505	0.165		
	LR	Normal	0.928	0.727	0.266	0.784
		Hyper.	0.273	0.072		
	SVM	Normal	1	1	-	0.500
		Hyper.	0	0		
	KNN	Normal	0.789	0.609	0.178	0.587
		Hyper.	0.391	0.211		
	PART	Normal	0.925	0.742	0.243	0.729
		Hyper.	0.258	0.075		
	J48	Normal	0.941	0.795	0.214	0.698
		Hyper.	0.205	0.059		
	MLP	Normal	0.894	0.674	0.258	0.764
		Hyper.	0.326	0.106		
Wrapper	NB	Normal	0.924	0.716	0.270	0.720
		Hyper.	0.284	0.076		
	LR	Normal	0.939	0.742	0.271	0.757
		Hyper.	0.258	0.061		
	SVM	Normal	-	-	-	-
		Hyper.	-	-		
	KNN	Normal	0.978	0.923	0.129	0.653
		Hyper.	0.077	0.022		
	PART	Normal	0.944	0.765	0.258	0.720
		Hyper.	0.235	0.056		
	J48	Normal	0.95	0.809	0.219	0.674
		Hyper.	0.191	0.05		
	MLP	Normal	0.922	0.729	0.250	0.738
		Hyper.	0.271	0.078		

Algo.: algorithms, AUC: the area under the receiver operating characteristic curve, MCC: Matthews's correlation coefficient, Sens.: sensitivity, 1-sp.: 1-specificity, Hyper.: hypertension, 표기 "-"는 wrapper 기반 변수 선택법을 적용하였을 시에 선택된 변수가 없는 것을 의미하거나, 성능지표상 점수가 매우 낮아 계산이 불가한 것을 의미함)

이상지질혈증 판별 모델에서, 전체 변수 (full)를 이용한 모델 중 NB 모델이 다른 모든 모델들보다 우수한 성능을 나타내었다. CFS 기반 모델들에서는 NB와 LR이 가장 우수하였고, 세부성능에서는 NB가 LR보다 다소 우수하였다. Wrapper 기반 모델들에서는 LR모델이

표 4. 이상지질혈증 판별 모델 성능 평가 결과
Table 4. Results of performance evaluation of identification models of dyslipidemia

	Algo.	Class	Sens.	1-spe.	MCC	AUC
Full	NB	Normal	0.801	0.578	0.217	0.697
		Dysli.	0.422	0.199		
	LR	Normal	0.957	0.877	0.140	0.695
		Dysli.	0.123	0.043		
	SVM	Normal	1	1	-	0.500
		Dysli.	0	0		
	KNN	Normal	0.777	0.701	0.076	0.539
		Dysli.	0.299	0.223		
	PART	Normal	0.81	0.704	0.109	0.559
		Dysli.	0.296	0.19		
	J48	Normal	0.867	0.754	0.131	0.595
		Dysli.	0.246	0.133		
	MLP	Normal	0.796	0.696	0.100	0.621
		Dysli.	0.304	0.204		
CFS	NB	Normal	0.94	0.834	0.160	0.711
		Dysli.	0.166	0.06		
	LR	Normal	0.971	0.921	0.107	0.712
		Dysli.	0.079	0.029		
	SVM	Normal	1	1	-	0.500
		Dysli.	0	0		
	KNN	Normal	0.901	0.815	0.112	0.611
		Dysli.	0.185	0.099		
	PART	Normal	0.988	0.966	0.069	0.578
		Dysli.	0.034	0.012		
	J48	Normal	1	1	-	0.500
		Dysli.	0	0		
	MLP	Normal	0.977	0.934	0.100	0.694
		Dysli.	0.066	0.023		
Wrapper	NB	Normal	-	-	-	-
		Dysli.	-	-		
	LR	Normal	0.993	0.965	0.105	0.629
		Dyslipidemia	0.035	0.007		
	SVM	Normal	-	-	-	-
		Dysli.	-			
	KNN	Normal	0.996	0.975	0.091	0.553
		Dysli.	0.025	0.004		
	PART	Normal	0.998	0.996	0.023	0.541
		Dysli.	0.004	0.002		
	J48	Normal	-	-	-	-
		Dysli.	-	-		
	MLP	Normal	1	1	-	0.515
		Dysli.	0	0		

Algo.: algorithms, AUC: the area under the receiver operating characteristic curve, MCC: Matthews's correlation coefficient, Sens.: sensitivity, 1-sp.: 1-specificity, Dysli.: dyslipidemia, 표기 "-"는 wrapper 기반 변수 선택법을 적용하였을 시에 선택된 변수가 없는 것을 의미하거나, 성능지표상 점수가 매우 낮아 계산이 불가한 것을 의미함)

다른 모든 모델들보다 우수한 성능을 나타내었으나, 실제적으로 다른 변수선택법 기반 모델들에 비하여 판별

표 5. 변수 선택법에서 선별된 최종 변수 리스트
Table 5. Lists of final variables extracted by variable subset selection methods

Method	Machine learning	Hypertension	Dyslipidemia
CFS	NB	age, education, SBP, WC, WHtR	age, WHtR
Wrapper	NB	education, drinking, activity, height, WHtR	-
	LR	age, income, stress, activity, pulse, WHtR, BMI	drinking, SDWD, depressive score, BMI, right grip strength
	SVM	-	-
	KNN	WC	drinking, height
	PART	age, DBP, WC	sex, education, left grip strength
	J48	age, drinking, SDWD, WC	-
	MLP	age, activity, WC, right grip strength	sex, depressive score

(WHtR: waist-to-height ratio, BMI: body mass index, WC: waist circumference, DBP: diastolic blood pressure, SBP: systolic blood pressure, SDWD: sleep duration during weekday, SDWE: sleep duration during weekend, 표기 "-"는 wrapper 기반 변수 선택법을 적용하였을 시에 선택된 변수가 없는 것을 의미함)

정확도가 매우 낮은 것으로 나타났다. 표 5는 각각의 변수 선택법에서 최종적으로 선별된 변수들의 리스트를 제시하고 있다.

결론적으로, 각 질병은 모델 생성 기법에서 사용된 변수 선택법과 머신러닝 알고리즘에 따라 모델별로 판별 성능이 다소 차이가 있는 것으로 나타났다. 또한 2가지 질병에 대한 모든 판별 모델들을 검토한 결과 wrapper 기반 모델들을 제외하고 전체적으로 NB 알고리즘을 기반으로 생성된 모델들이 다른 머신러닝 기반 모델들보다 다소 우수한 것으로 나타났다.

본 생성 모델들은 아래와 같은 제약사항을 지니고 있으므로 본 실험에서 사용된 모델 생성방법에 따라 성능차이가 항상 유사하게 나타난다고 판단하기 어렵다. 첫째, 본 실험에 사용된 샘플수는 데이터 셋 자체적인 문제와 샘플 추출 등의 과정에 기인하여 한국인 전체 샘플을 대표하지 못하며, 정상군과 질병군사이에서 샘플수 차이에 대한 편향성이 존재한다. 둘째, 본 실험에서 고려되어지지 않은 다양한 변수 선택법 메소드들과 판별 알고리즘들이 존재한다. 셋째, 각 머신러닝 알고리즘에

대한 파라미터 값들에 대하여 디폴트 값 만을 사용함에 따라 각 모델들은 파라미터 튜닝에 따라 다른 성능을 나타낼 수 있다.

Ⅳ. 결 론

본 연구에서는 한국인의 심혈관 질환중 고혈압과 이상지질혈증 판별을 위하여 데이터마이닝을 기반으로 다양한 판별 모델들을 생성하였고, 생성된 모델들에 대한 성능 비교평가를 수행하였다. 본 연구의 결과에서 두 가지 질병 모두에서 wrapper기반 변수 선택법을 제외한 대부분의 모델에서 NB가 주요성능지표와 세부성능지표의 관점에서 다른 머신러닝을 이용한 모델들보다 다소 우수한 판별 성능이 있는 것으로 나타났다. 또한, wrapper 기반 변수 선택법에서는 LR 모델이 NB 모델보다 다소 우수한 것으로 나타났다. 본 연구의 결과는 원격의료 및 대중보건분야에서 향후 한국인의 심혈관질환에 대한 판별 및 예측 모델 생성을 위한 참고자료로 활용될 수 있을 것으로 기대된다.

References

[1] H.H. Lee, S.M.J. Cho, H. Lee, J. Baek, J.H. Bae, W.J. Chung, H.C. Kim, “Korea Heart Disease Fact Sheet 2020: Analysis of Nationwide Data,” Korean circulation journal, Vol. 51, No. 6, pp. 495-503, 2021. DOI.org/10.4070/kcj.2021.0097

[2] B.J. Lee, J.Y. Kim, “A Comparison of the Predictive Power of Anthropometric Indices for Hypertension and Hypotension Risk,” PLoS ONE, Vol. 90, No. 1, pp. e84897, 2014. DOI.org/10.1371/journal.pone.0084897

[3] B.J. Lee, B. Ku, “A comparison of trunk circumference and width indices for hypertension and type 2 diabetes in a large-scale screening: a retrospective cross-sectional study,” Scientific Reports, Vol. 8, pp. 13284, 2018. DOI.org/10.1038/s41598-018-31624-x

[4] J.H. Chi, B.J. Lee, “Risk factors for hypertension and diabetes comorbidity in a Korean population: A cross-sectional study,” PLoS ONE Vol. 17, No. 1, pp. e0262757, 2022. DOI.org/10.1371/journal.pone.0262757

[5] C.F. Lin, Y.H. Chang, S.C. Chien, Y.H. Lin, H.T. Yeh, “Epidemiology of Dyslipidemia in the Asia Pacific Region,” International Journal of Gerontology, Vol. 12, No. 1, pp. 2-6, 2018. DOI.org/10.1016/j.ijge.2018.02.010

[6] B.J. Lee, “Prediction Model of Hypertension Using Sociodemographic Characteristics Based on Machine Learning,” KIPS Transactions on Software and Data Engineering, Vol. 10, No. 11, pp. 541-546, 2021. DOI.org/10.3745/KTSDE.2021.10.11.541

[7] M.H. Kim, J.H. Seo, J.Y. Lee, “Nomogram building to predict dyslipidemia using a na¨ıve Bayesian classifier model,” The Korean Journal of Applied Statistics, Vol. 32, No. 4, pp. 619-630, 2019. DOI.org/10.5351/KJAS.2019.32.4.619

[8] Y.H. Lee, E.M. Kwak, M. Jo, “Factors affecting cardiovascular disease in Korea adults: Focusing on smoking behavior including urine cotnine and health behaviors,” The Journal of the Convergence on Culture Technology, Vol. 7, No. 3, pp. 293-301, 2021. DOI.org/10.17703/JCCT.2021.7.3.293

[9] X. Z, F. Tang, J. Ji, W. Han, P. Lu, “Risk Prediction of Dyslipidemia for Chinese Han Adults Using Random Forest Survival Model,” Clinical Epidemioloy. Vol. 11, pp. 1047-1055, 2019. DOI.org/10.2147/CLEP.S223694

※ 이 논문은 2021년도 정보(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2021-0-00104), 비대면 심혈관 건강관리를 위한 디지털헬스 서비스 플랫폼 개발