Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology
Vol.5, No.1, February (2015), pp. 47-56
http://dx.doi.org/10.14257/AJMAHS.2015.02.03

통계적 의사결정 모형을 이용한 우리나라 여성의 흡연 요인

변해원[1)]

A Study on Effects of Women's smoking behavior in Korea Using the Classification and Regression Trees

Haewon Byeon[1)]

요 약

이 연구는 우리나라 여성 흡연에 영향을 미칠 수 있는 다양한 요인을 고려하여 통계학적 모형을 개발하고 여성 금연을 위한 기초자료를 제공하였다. 자료원은 2010년 서울복지패널이며, 분석 대상은 조사를 완료한 20세 이상 여성 3,958명이다. 결과변수는 현재 흡연 여부(흡연자, 비흡연자)로 정의하였다. 설명변수는 연령, 최종학력, 현재 취업 상태, 가구 월 평균 총 소득, 배우자 유무, 음주 여부, 주관적 건강상태, 정기적인 운동여부, 가족관계, 교우관계, 지난 1달 간 우울증상 여부, 현재 질병 여부로 설정하였다. 분석방법은 데이터마이닝의 CART(Classification And Regression Tree) 알고리즘을 이용하였다. 한국 여성의 흡연 분류 모형을 구축한 결과, 유의미한 요인은 음주, 우울증상, 가구 월 평균 총소득, 주관적 가족관계, 배우자 유무이었다. 이 결과를 기초로 우리나라 여성 흡연자의 특성을 고려한 맞춤형 금연 프로그램 개발이 필요하다.

핵심어 : 데이터마이닝, 흡연, 위험요인, 의사결정나무모형

Abstract

The purpose of this study was to analyze the factors that affects the Women's smoking behavior in Korea. Data were from the A Study on the Seoul Welfare Panel Study 2010. Subjects were 3,958 females aged 19 and older living in the community. A prediction model was developed by the use of a classification and regression tree algorithm of data-mining approach. In the classification and regression tree algorithm analysis, alcohol consumption, depression, income, family relationship, marital status were significantly associated with smoking behavior in Korean females.

Keywords : Smoking, Classification And Regression Tree, Decision tree, Risk factor

Received (December 19, 2014), Review Request(December 22, 2014), Review Result(January 05, 2015)
Accepted(January 28, 2015), Published(February 28, 2015)

[1]506-706 Dept. Speech Language Pathology & Audiology, Nambu University, Chumdan 23, Gwansangu, Gwangju, Korea,
email: byeon@nambu.ac.kr

ISSN: 2383-5281 AJMAHS

1. 서론

현재 국민건강증진 사업의 주요 핵심 과제 중 하나는 금연 사업이다. 우리나라는 1995년 국민건강증진법이 제정된 이후 담배광고 규제, 금연구역 설정, 담배부담금 부과, 금연 교육 및 금연 홍보 등 지속적으로 금연정책을 수행해왔다. 특히, 2015년부터는 국민건강증진법이 개정되어 담배값이 평균 80% 이상 인상되는 등의 강도 높은 금연 정책이 시행되고 있다. 이처럼 지난 20년 동안 계속된 적극적인 금연 정책에도 불구하고, 여전히 우리나라의 흡연율은 경제협력개발기구(OECD) 국가 중에서 매우 높은 수준이다[1].

흡연은 대표적인 건강위험행위로서 만성적인 흡연은 만성폐쇄성폐질환, 천식 등의 호흡기질환의 주요 원인일 뿐만 아니라, 관상동맥성, 심잘질환, 뇌졸중, 구강암, 식도암, 후두질환 등의 위험요인이다[2-5]. 특히, 여성 흡연은 조기 폐경, 유방암, 자궁경부암 등 여성 질환의 위험 요인으로 작용하기 때문에 남성 흡연보다 건강에 더 부정적이다[6]. 또한, 임산부의 흡연은 유산, 조산, 저체중아 출생 등 태아의 건강에도 악영향을 미칠 수 있다.

이 같은 흡연으로 인해 손실된 우리나라의 사회경제적 비용은 2006년 한 해에만 최소 4조8860억 원에서 최대 5조 9381억 원으로 추산되는데, 사회적 손실에 여성의 흡연으로 인한 출산 저하 등의 간접적인 손실을 포함한다면 상상할 수 없을 만큼 큰 폭으로 증가될 것으로 예측된다[7].

흡연은 수정 가능한 건강행위이기 때문에, 사회적으로 금연을 위한 다양한 정책이 수립되어 왔다. 그럼에도 불구하고, 남성의 흡연율은 1989년 70.4%에서 2010년 48.3%로 지속적으로 감소하는 반면에, 같은 기간 여성 흡연율은 4.4%에서 6.3%로 소폭 증가하였다[8]. 이처럼 여성의 금연 정책이 상대적으로 남성에 비해서 효과적이지 못한 이유가 여성의 흡연에 미치는 요인이 다이어트 시도 등 남성과는 다르기 때문이라는 지적도 있다[9]. 따라서 여성의 성공적인 금연을 위해서는 여성 흡연의 실태를 파악하고 관련 요인을 규명하는 것이 중요하다.

현재까지 여성의 흡연을 주제로 한 다수의 연구들이 수행되었고, 그 결과, 여성의 흡연에 영향을 미치는 요인으로 직업, 연령, 경제수준, 학력, 음주, 우울증 등이 보고되었다[10-14]. 그러나 이 같은 선행연구들은 대부분 건강 행위 등의 개별 요인의 탐색에 머무르고 있으며, 다양한 요인을 복합적으로 규명한 연구는 매우 드물다.

흡연은 사회, 문화의 영향을 받기 때문에 국외 연구 결과를 직접적으로 적용하는 것은 한계가 있다. 따라서 우리나라 여성의 흡연 실태를 파악하고, 흡연에 영향을 미칠 수 있는 요인을 파악하기 위해서는 인구사회학적 요인을 포함한 사회적 환경, 경제적 환경, 건강수준, 건강행위 등 다양한 요인을 고려한 분석이 요구된다.

이 연구는 우리나라 여성 흡연에 영향을 미칠 수 있는 다양한 요인을 고려한 통계학적 모형을 개발하고 여성 금연을 위한 기초자료를 제공하였다.

2. 연구 방법

2.1 연구대상

이 연구에서 분석한 자료는 2010년 6월 1일부터 2010년 8월 31일까지 서울 시민을 대상으로 서울시복지재단에서 조사한 서울시복지패널조사(Seoul Welfare Panel Study)의 원시데이터의 일부이다. 서울시복지패널조사는 서울시에 거주하는 가구의 복지수준을 파악하고 복지취약계층의 실태파악 및 복지서비스 수요를 추정하기 위한 목적에서 2009년 통계청의 승인(제20113호)을 받아서 수행되었다[15]. 2005년 인구주택총조사 대상가구 중 조사 시기를 기준으로 한 서울시 소재 가구를 모집단으로 하였고, 표본추출방식은 서울시 25개 구를 대상으로 층화집락추출방법을 이용하였다. 주요 조사항목은 소득, 경제수준, 건강, 생활여건, 복지서비스 수요 등이며, 조사방법은 면접원이 조사 대상 가구를 방문하여 휴대용 컴퓨터에 구조화된 설문에 따라 응답한 내용을 입력하는 컴퓨터를 이용한 대면면접조사(Computer Assisted Personal Interviewing)방법을 이용하였다.

본 연구는 조사 완료자 7,761명 중에서 남성 3,547명, 19세 이하 256명을 제외한 여성 3,958명을 분석대상으로 하였다.

2.3 변수의 측정 및 정의

결과 변수는 현재 흡연 여부(흡연자, 비흡연자)로 정의하였다. 현재 흡연자는 세계보건기구(WHO)의 기준을 참고하여 매일 한 개비 또는 가끔 담배를 피우는 사람으로서, 평생 담배 5갑(100개비)이상 흡연한 사람으로 정의하였다. 비흡연자는 과거 담배를 피웠지만 현재는 피우지 않는 사람(과거 흡연자)과 평생 동안 담배를 피운 경험이 없거나, 담배를 5갑(100개비) 미만으로 피운 사람(비흡연자)로 정의하였다.

설명변수는 연령(20-39세, 40-59세, 60세 이상), 최종학력(초등학교 이하, 중학교, 고등학교, 대학졸업 이상), 현재 취업 상태(취업, 미취업), 가구 월 평균 총 소득(200만원 미만, 200-400만원, 400만원 이상), 배우자 유무(배우자가 있고 함께 살고 있음, 배우자가 있으나 함께 살고 있지 않음, 배우자 없음), 음주 여부(음주자, 비음주자), 주관적 건강상태(좋음, 보통, 나쁨), 정기적인 운동 여부(없음, 있음) 주관적 가족관계(좋음, 보통, 나쁨), 교우관계(좋음, 보통, 나쁨), 지난 1달 간 우울증상 여부(없음, 있음), 현재 순환기, 내분비계, 근골격계, 호흡기계, 이비인후 질환, 간질환, 비뇨기계질환 등 질병 여부(없음, 있음)를 포함하였다.

2.4 분석 방법

2.4.1 여성 흡연의 잠재적 요인 탐색

흡연 여부에 따른 집단 간의 차이는 카이제곱검정(Chi-square test)으로 분석하였다. 이 때, 유의미수준 0.5 이하인 설명변수는 여성 흡연의 잠재적 요인으로 가정하고 데이터마이닝 모형에 포함하였다[16].

2.4.2 CART 알고리즘

CART(Classification And Regression Tree)는 통계학적 의사결정분류모형의 분석 알고리즘 중 하나로서 지니 계수(Gini Index)를 이용하여 불순도(impurity)를 측정하며, 부모마디로부터 자식마디가 2개만 형성되는 이진분류(binary split)에 기반한 알고리즘이다[17]. CART는 생성되는 규칙을 해석하기 쉽고, 연속형 변수와 범주형 변수를 모두 이용할 수 있다는 장점이 있다[18].

지니 계수는 n개의 원소 중에서 임의로 2개를 추출하였을 때, 추출된 2개가 서로 다른 그룹에 속할 수 있는 확률을 의미하며, 도출과정은 각 마디에서 도수가 가장 많은 목표변수의 오분류 확률을 식(1)의 통계식을 이용하여 산출한다.

$$Gini\,Index(t) = 1 - \sum_{j}[P(j/t)]^2 \quad (1)$$

지니계수의 감소량이 계산되면, 알고리즘의 마지막 과정으로 지니 계수를 가장 감소시켜 주는 분류 변수와 최적 분리를 자식 마디로 선택한다.

이 연구의 모형에서 CART 알고리즘에 대한 의사결정규칙(decision rule)의 분리 및 병합 기준값은 0.05로 설정 하였고, 부모마디의 수는 200명, 자식마디 수는 100명, 분지가지 개수는 5개로 제한하였다.

최종 모형의 타당성 평가는 K배 교차 검증법(K-fold cross-validation method)을 이용하여 평가하였다[18]. 이 때, 모형의 신뢰도를 높일 수 있는 K의 값은 10일 경우가 적절하다고 알려져 있기 때문에 본 연구에서도, K의 값을 10으로 설정한 10배 교차 검증법을 적용하였다[19].

모든 분석은 MINITAB version 13(Minitab Inc., State College, Pennsylvania, USA)과 Decision Tree version 20.0(IBM Inc., Chicago, Illinois, USA)을 이용하였다. 유의수준은 양측검정에서 0.05로 설정 하였다.

3. 결과

3.1 연구대상의 특성과 흡연의 잠재적 요인

[표 1] 흡연 여부에 따른 일반적 특성

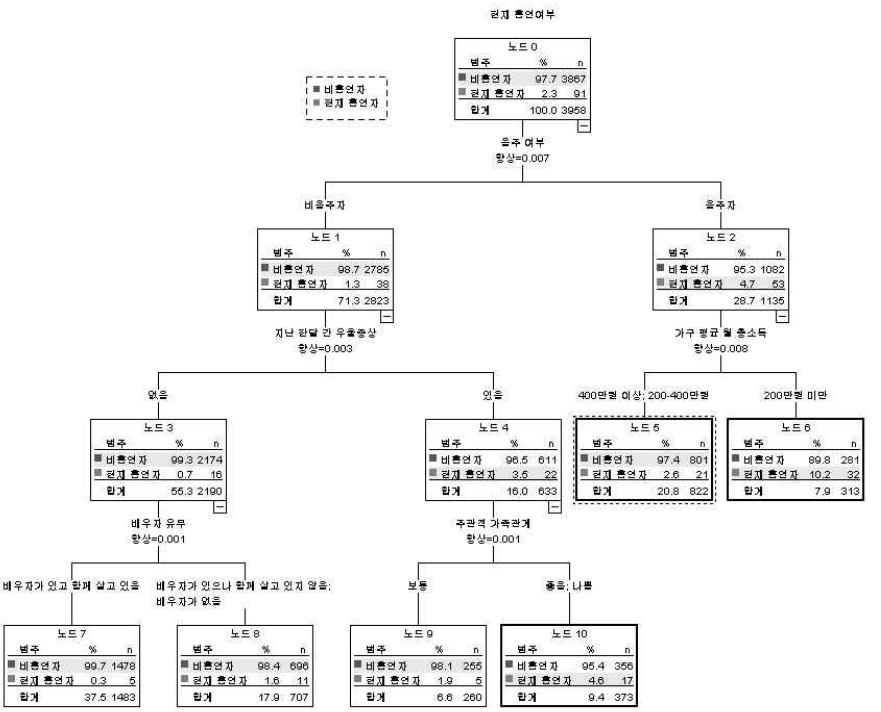
[Table 1] General characteristics of the subjects by smoking, n(%)

Variables	Smoking		p
	No (n=3,867)	Yes (n=91)	
연령			0.484
20-39세	1,303 (97.3)	36 (2.7)	
40-59세	1,359 (98.0)	28 (2.0)	
60세 이상	1,205 (97.8)	27 (2.2)	
최종학력			0.183
초등학교 졸업이하	836 (97.0)	26 (3.0)	
중학교 졸업	394 (98.0)	8 (2.0)	
고등학교 졸업	1,112 (97.4)	30 (2.6)	
대학 졸업 이상	1,525 (98.3)	27 (1.7)	
현재 취업 상태			0.200
취업	1,081 (97.2)	31 (2.8)	
미취업	2,786 (97.9)	60 (2.1)	
가구 월 평균 총 소득			<0.001
200만원 미만	1,435 (96.1)	58 (3.9)	
200만원-400만원	1,294 (98.4)	21 (1.6)	
400만원 이상	409 (98.6)	6 (1.4)	
배우자 유무			<0.001
동거	2,512 (98.8)	31 (1.2)	
별거	75 (96.2)	3 (2.8)	
없음	1,280 (95.7)	57 (4.3)	
음주 여부			<0.001
비음주자	2,785 (98.7)	38 (1.3)	
음주자	1,082 (95.3)	53 (4.7)	
주관적 건강상태			<0.001
좋음	1,667 (98.7)	22 (1.3)	
보통	1,276 (97.9)	27 (2.1)	
나쁨	924 (95.7)	42 (4.3)	
정기적인 운동 여부			0.097
안한다	2,434 (97.4)	65 (2.6)	
한다	1,433 (98.2)	26 (1.8)	
주관적 가족관계			0.001
좋음	2,595 (98.3)	44 (1.7)	
보통	995 (97.1)	30 (2.9)	
나쁨	223 (94.9)	12 (5.1)	
주관적 교우관계			0.384
좋음	1,591 (97.7)	37 (2.3)	
보통	1,854 (97.9)	40 (2.1)	

나쁨	422 (96.8)	14 (3.2)	
지난 한 달 간 우울증상			<0.001
없음	3,085 (98.4)	51 (1.6)	
있음	782 (95.1)	40 (4.9)	
현재 질병 여부			0.037
없음	1,496 (97.1)	45 (2.9)	
있음	2,371 (98.1)	46 (1.9)	

흡연 여부에 따른 대상자의 일반적 특성 및 잠재적 요인은 [표 1]에 제시하였다. 전체 대상자 3,958명 중에서 현재 흡연자는 91명(2.3%) 이었다. 카이제곱검정 결과, 현재 흡연자와 비흡연자는 가구 월 평균 총 소득, 배우자 유무, 음주 여부, 주관적 건강 상태, 주관적 가족 관계, 지난 한 달 간 우울 증상경험, 현재 질병 여부에서 통계적으로 유의미한 차이가 있었다($p<0.05$). 가구 월 평균 총 소득 200만원 미만(3.9%), 배우자 없음(4.3%), 음주자(4.7%), 주관적 건강 상태가 나쁨(4.3%), 주관적 가족 관계가 나쁨(5.1%), 현재 순환기, 내분비계, 근골격계, 호흡기계, 이비인후 질환, 간질환, 비뇨기계질환 등의 질병이 없는 집단(2.9%)에서 현재 흡연율이 높았다.

3.2 CART 알고리즘을 이용한 여성의 흡연 요인



[그림 1] 한국 여성의 흡연 행위 분류 모형

[Fig. 1] Prediction model for smoking behavior in Korean female

CART 알고리즘을 이용한 여성 흡연의 분류 모형은 [그림 1]에 제시하였다. 카이제곱 검정을 이용하여 여성 흡연의 잠재적 요인으로 설정된 변수들을 예측 모형에 포함한 후 CART 알고리즘을 이용하여 통계학적 분류모형을 구축한 결과, 유의미한 영향을 미치는 분류 변수는 음주, 우울증상, 가구 월 평균 총소득, 주관적 가족관계, 배우자 유무이었다.

가장 우선적으로 관여하는 예측 요인은 음주 여부였다. 다음으로 비음주자에서는 지난 한달 간 우울증상이 관여되는 분류 변수였고, 음주자는 가구 월 평균 총소득이 분류 변수였다. 세 번째는 우울증상이 없는 집단에서는 배우자 유무가 분류 변수였고, 우울증상이 있는 집단은 주관적 가족관계가 분류 변수였다.

[표 2]는 여성 흡연 행위 분류에 있어서 유의미한 경로를 이득율이 높은 순서 데로 제시한 CART 알고리즘 기반 예측모형의 이익 도표이다. 마디번호는 최종마디의 번호이고, Gain Index (%)는 최종 노드에 대한 이익지표이다. 이익지표를 도출했을 때, 총 3개의 유의미한 경로가 확인되었지만, 세 번째 경로는 비흡연자로 범주가 예측되었기 때문에, 여성 흡연 분류 경로를 탐색하기 위한 유의미한 경로는 2개의 노드가 확인 되었다. 먼저, 여성 흡연의 분류에 있어서 이익지표 값이 가장 큰 제1경로는 월 평균 가구소득이 200만원 미만인 현재 음주를 하는 여성으로 10.2%가 현재 흡연자로 분류되었고, 이익지표는 444.7%이었다. 제2경로는 지난 한달 간 우울증상이 있으며, 가족관계가 나쁘다고 인지하는 음주를 하지 않는 여성으로 4.6%가 현재 흡연자로 분류되었고, 이익지표는 198.2%이었다.

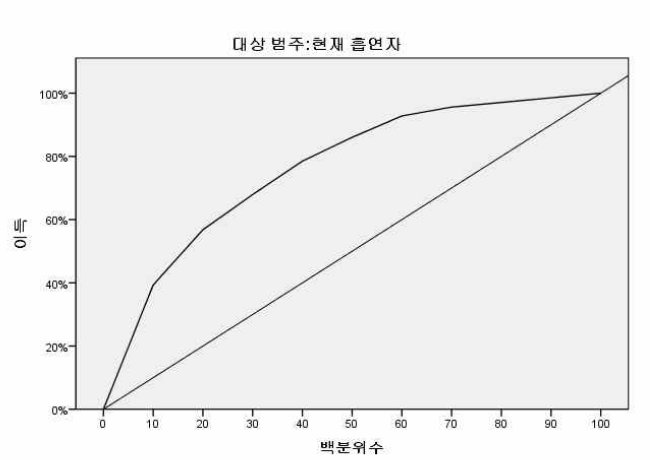
[표 2] CART 알고리즘에 의한 이득 도표
[Table 2] Gains chart of predictor variable by CART algorithm

Node no	Node n (%)[1]	Gain n (%)[2]	Response (%)[3]	Gain Index (%)[4]	특성	예측 범주
6	313 (7.9)	32 (35.2)	10.2	444.7	월 평균 가구소득이 200만원 미만인 현재 음주를 하는 여성	현재 흡연자
10	373 (9.4)	17 (18.7)	4.6	198.2	지난 한달 간 우울증상이 있으며, 주관적 가족관계가 나쁜 비음주 여성	현재 흡연자
5	822 (20.8)	21 (23.1)	2.6	111.1	월 평균 가구소득이 200만원 이상인 현재 음주를 하는 여성	비흡연자

1 Node n(%); node number, % to 3,958
2 Gain n(%); gain number, % to 91
3 Response (%): The fraction of the smoking behavior
4 Gain index (%):=444.7 in total 3 node

예측 모형의 분석이 완료되면, 개발된 예측 모형을 평가하기 위해서 10-fold 교차타당성 검정을

이용하였다. 도출된 모형의 안정성을 비교하기 위해서 10-fold 교차타당성 검정을 수행한 결과, 위험지수는 크로스 분류모형의 위험지수는 0.220, 오분류율은 22%로 도출되어, 예측모형의 위험지수 0.220 및 오분류율 22%와 동일하였다.



[그림 2] 최종 모형의 이득율

[Fig. 2] Gains percentile of final model

4. 결론

이 연구는 통계적 의사결정분류 모형의 분석 방법 중 하나인 CART 알고리즘을 이용하여 한국 여성의 흡연 요인을 분석하였다. 통계학적 분류모형을 구축한 결과, 첫째, 월 평균 가구소득이 200만원 미만인 현재 음주를 하는 여성, 둘째, 지난 한달 간 우울증상이 있으며, 가족관계가 나쁘다고 인지하는 음주를 하지 않는 여성은 현재 흡연의 가능성이 높은 집단으로 예측되었다.

본 연구의 결과를 근거로 한국 여성의 건강 증진을 위해서 여성 흡연자의 특성을 고려한 맞춤형 금연 프로그램 개발이 요구된다.

References

[1] World Health Organization, The world health report: 2006: working together for health, World Health Organization, Geneva **(2006)**.

[2] J. Banoczy and C. Squier, Smoking and disease. European Journal of Dental Education. **(2004)**, Vol.8, pp.7-10.

[3] H. Byeon. The trend of the association between amount of smoking and self-reported voice problem. Journal of the Korea Academia-Industrial cooperation Society. **(2012)**, Vol.13, No.3, pp.1246-1254.

[4] H. Byeon. The association between smoking and self-reported voice problems in the Korean Adult Population. Journal of speech & hearing disorders. **(2011)**, Vol.20, No.3, pp.17-30.

[5] H. Byeon and Y. Lee, Prevalence and risk factors of benign laryngeal lesions in the adult population. Communication Sciences and Disorders. **(2010)**, Vol.15, No.4, pp.648-656.

[6] Y. S. Kim and J. S. Jo, Smoking behavior and related factors of female smokers from public health center in Incheon. Journal of Korean Society for Health Education and Promotion. **(2008)**, Vol.25, No.3, pp.125-138.

[7] S. Kim and S. Kwon, The social cost of smoking in Korea. Korean journal of policy analysis and evaluation. **(2008)**, Vol.18, No.4, pp.119-140.

[8] Korea Centers for Disease Control and Prevention, Korea National Health and Nutrition Examination Survey 2010, Korea Centers for Disease Control and Prevention, O-Song **(2013)**.

[9] K. A. Perkins, M. D. Levine, M. D. Marcus and S. Shiffman, Addressing women' concerns about weight gain due to smoking cessation. Journal of Substance Abuse Treatment. **(1997)**, Vol.14, No.2, pp.173-182.

[10] S. Sohn, LCGM Analysis of Smoking Patterns of Korean Women. Social Science Research. **(2013)**, Vol.29, No.1, pp.219-235.

[11] M. K. Suh, Women's smoking behavior and its implications. Health and welfare policy forum. **(2009)**, Vol.152, pp73-82.

[12] K. Park, Predictors of intention to quit smoking among woman smokers in Korea. The Korean journal of fundamentals of nursing. **(2014)**, Vol.21, No.3, pp.253-263.

[13] H. Byeon, Relationship between cigarette smoking and depression symptoms of high school students. Journal of the Korea Academia-Industrial cooperation Society. **(2012)**, Vol.13, No.10, pp.4669-4675.

[14] H. Byeon and R. Lee, Prediction model for the smoking in Korean adolescent using CART algorism. Information-An International Interdisciplinary Journal. **(2014)**, Vol.17, No.12, pp.6273-6278.

[15] Seoul Welfare Foundation, Seoul Welfare Panel Study 2010, Seoul Welfare Foundation, Seoul **(2010)**.

[16] H. Byeon, The prediction model for self-reported voice problem using a decision tree model. Journal of the Korea Academia-Industrial cooperation Society. **(2013)**, Vol.14, No.7, pp.3368-3373

[17] L. Brieman, J. Friedman, R. A. Olshen, C. J. Stone, Classification and Regrression Trees, Chapman &

Hall, New York **(1984)**.

[18] H. Byeon, The factors that affects the experience of discrimination in children in multi-cultural families using QUEST algorithm : focusing on Korean language education, Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology. **(2014)**, Vol.4, No.2, pp.303-312.

[19] L. Breiman, Bagging Predictors. Machine Learning. **(1996)**, Vol.24, No.2, pp.123-140.