

12세 아동 우식경험영구치아수 예측을 위한 머신러닝 알고리즘의 적용

양용훈, 김지수, 정승화

부산대학교 치의학전문대학원 예방과사회치학교실

Prediction of dental caries in 12-year-old children using machine-learning algorithms

Yong-Hoon Yang, Ji-Soo Kim, Seung-Hwa Jeong

Departments of Preventive & Community Dentistry, Pusan National University School of Dentistry, Yangsan, Korea

Received: February 5, 2020

Revised: March 4, 2020

Accepted: March 7, 2020

Corresponding Author: Seung-Hwa Jeong
Departments of Preventive & Community
Dentistry, Pusan National University
School of Dentistry, 49 Busandaehak-ro,
Meulgeum-up, Yangsan 50612, Korea
Tel: +82-51-510-8222
Fax: +82-51-510-8221
E-mail: jsh0917@pusan.ac.kr
<https://orcid.org/0000-0001-5173-2859>

Objectives: The decayed-missing-filled (DMFT) index is a representative oral health indicator. Prediction of DMFT index is an important basis for the development of public oral health care projects and strategies for caries prevention. In this study, we used data from the 2015 Korean children's oral health survey to predict DMFT index and caries risk groups using statistical techniques and four different machine-learning algorithms.

Methods: DMFT prediction models were constructed using multiple linear regression and four different machine-learning algorithms: decision tree regressor, decision tree classifier (DTC), random forest regressor, and random forest classifier (RFC). Thereafter, their accuracies were compared.

Results: For the DMFT predictive model, the prediction accuracy of multiple linear regression and RFC were 15.24% and 43.27%, respectively. The accuracy of DTC prediction was 2.84 times that of multiple linear regression. The important feature of the machine-learning model, which predicts DMFT index and the caries risk group, was the number of teeth with sealants.

Conclusions: Using data from the 2015 Korean children's oral health survey, which is considered big data in the field of oral health survey in Korea, this study confirmed that machine-learning models are more useful than statistical models for predicting DMFT index and caries risk in 12-year-old children. Therefore, it is expected that the machine-learning model can be used to predict the DMFT score.

Key Words: Decayed-missing-field-teeth, Decision tree algorithm, Machine learning, Prediction, Random forest algorithm

서론

치아우식은 대표적인 치과질환 중의 하나로, 병원성 치면세균막에서 발생된 산(acid)에 의해 치아법랑질과 상아질이 탈회되고, 계속 진행되면 결국 치질이 파괴될 뿐만 아니라 치아신경조직인 치수와 치조골의 염증과 통증을 유발하여 치아를 발거하게 되는 주된 원인 중의 하나이다. 치아우식은 아동기에 발생할 수 있는 가장 흔한 만성 질환

중에 하나이며, 전체 생애에 걸쳐 삶의 질에 큰 영향을 미치게 되므로 중요한 공중보건 문제이다^{1,2)}. 우식위험치아나 우식치아를 조기에 발견한다면 간단한 예방치료와 수복치료로 해결이 가능하겠지만, 이를 방지하여 상아질이나 치수까지 침범할 경우 광범위한 치질 삭제와 근관치료, 발치 후 보철치료와 같은 침습적이고 외과적인 처치를 받게 된다. 개개인이 본인의 구강건강에 큰 관심을 갖고 치과에 방문하여 전문가에게 정기적인 관리를 받는다면 이상적이겠지만 사회경제적 여

건이나 구강건강의 중요성에 대한 인식의 차이로 자각 증상이 생기기 전까지 예방이나 검진을 소홀히 하는 경우가 많다.

우식경험영구치아수(DMFT)는 대표적인 구강건강지표로서, 한 사람이 보유한 우식경험영구치의 수를 일컫는 말이다. DMFT는 일반적으로 개인의 구강건강관리습관, 식습관, 치과의료이용, 사회경제적 요인 등에 의해 영향을 받는다고 알려져 있다. 인구집단의 DMFT와 그 관련 요인을 파악하고, DMFT를 예측하는 것은 구강보건사업을 개발하고, 개인의 우식예방전략을 수립하는 데 중요한 근거가 될 수 있다. 구강건강실태조사는 「구강보건법」 제9조에 의거, 국민의 구강건강상태 및 관련 행태, 치과의료이용 등을 조사함으로써 우리나라의 구강보건사업 목표 개발과 사업 계획 수립, 사업 우선순위 결정에 필요한 기초자료를 확보하기 위해 2000년부터 매 3년마다 수행되어 왔으며, 2015년에는 12세 아동 27,568명을 조사하여, 구강역학 분야의 빅데이터로 인정된다. 구강건강실태조사에서 DMFT는 연령이 증가할수록 높아지는 경향이 있으나, 2000년에서 2015년으로 오면서 차츰 낮아지고 있는 추세이다³⁾.

공학적으로 널리 알려진 머신러닝(machine learning)의 개념은 경험을 통해 데이터를 학습하는(learn from experience) 컴퓨터 프로그램으로 특정 업무를 수행함에 있어 경험을 반영해 그 성과를 향상시키고 이를 평가하는 코드를 가진 프로그램을 말한다⁴⁾. 머신러닝의 역사는 1950년대에 시작되어 1980-90년대까지 발전 후 답보 상태였지만, 2000년대 중반에 들어 인터넷에 의해 축적된 방대한 양의 빅데이터와 이를 처리할 수 있는 하드웨어 성능의 발달로 비약적인 발달을 이루었다⁵⁾. 최근에는 머신러닝을 이용한 이미지 인식, 음성 인식, 번역 등의 분야에서 주목할 만한 성과가 이뤄지고 있다.

머신러닝 기술은 최근 의료 분야에서도 다양하고 폭넓게 시도되고 있다. 우리나라 치의학 분야에서는 파노라마 및 치과용 컴퓨터 단층촬영, 구내 사진을 이용한 딥러닝 분석 연구가 수행 중인 것으로 파악되고 있으며^{6,7)}, 향후 연구 성과가 기대된다. 하지만 기술의 중요성과 유용성에 비해 치의학 분야에서의 활용은 아직 제한적이기 때문에 더 다양한 분야에 활용하려는 시도가 필요하다.

전문가에 의한 정밀진단을 수행하기 전에 설문조사나 기본적인 정보만으로 인구집단의 우식경험영구치아수를 예측한다면 구강검진에 소요되는 인력이나 시간, 비용을 줄이고 우식고위험자를 분류하여 전문가의 정확한 진단과 필요한 치료를 받게 하는데 도움이 될 것이다. 또한, 예측 모델에서 DMFT에 많은 영향을 미치는 기여 요인을 파악한다면, 그 요인을 조절하여 우식 예방에 큰 도움이 될 것이다. 따라서 이번 연구에서는 2015년 아동구강건강실태조사 자료를 활용하여 통계적 기법과 다양한 머신러닝 알고리즘을 이용한 우식경험영구치아수(DMFT) 예측 모델을 구축하고 회귀분석과 머신러닝 모델의 성능을 비교, 분석하는 것이다.

연구대상 및 방법

1. 연구대상

보건복지부가 2015년 수행한 아동구강건강실태조사(2015 Korean Children's Oral Health Survey) 원시자료를 이용하여 분석하였다. 아동구강건강실태조사는 구강보건법에 의거하여 국민의 구강건강상태와 구강건강의식 등을 정기적으로 조사하는 것으로 생명윤리 및 안전에 관한 법률상 심의 면제 대상이다³⁾. 2015년 아동구강건강실태조사에 참여한 만12세 대상자 수는 총 27,568명이었고, 그 중 구강검

Table 1. Distribution of study subjects by characteristics

Characteristics	Classification	N	%
Gender	Men	11,942	50.4
	Women	11,760	49.6
Region	City	18,496	78.0
	Rural area	5,206	22.0
Number of pit and fissure sealant	0	10,345	43.6
	1	2,546	10.7
	2	2,937	12.4
	3	2,138	9.0
	4	3,285	13.9
	5	636	2.7
	6	573	2.4
	7	306	1.3
	8	347	1.5
	9	114	0.5
	10	109	0.5
	11	100	0.4
	12	98	0.4
	13	40	0.2
	14	55	0.2
	15	38	0.2
	16	35	0.1
Perceived oral health status	Very good	1,349	5.7
	Good	9,034	38.1
	Fair	10,932	46.1
	Poor	2,231	9.4
	Very poor	156	0.7
Dental treatment demand for the past one year	Yes	15,267	64.4
	No	8,435	35.6
Experience of toothache for the past one year	Yes	5,046	21.3
	No	18,656	78.7
Frequency of snack intake per day	No intake	2,780	11.7
	Once	7,852	33.1
	2 times	7,797	32.9
	3 times	3,692	15.6
	4 and over	1,581	6.7
Number of oral hygiene auxiliaries in use	0	14,428	60.9
	1	6,787	28.6
	2	2,002	8.4
	3	418	1.8
	4	59	0.2
	5	8	0.0
Total		23,702	100.0

Data source from 2015 Korean Children's Oral Health Survey.

진을 수행하지 않았거나 관심변수의 결측치가 존재하는 대상자 3,866명을 제외한 23,702명의 자료를 최종 분석에 활용하였다(Table 1).

2. 연구방법

2.1. DMFT 예측을 위한 다중선형회귀 모델 구축

DMFT 예측 통계 모델을 구축하기 위하여 다중선형회귀분석을 이용하였다. 독립변수들을 선정하기 이전, 종속변수인 DMFT와의 상관분석을 시행하였으며, 각 변수들 간의 다중 공선성을 고려함과 동시에 종속변수인 DMFT와 상관관계에 있는 독립변수 8가지를 선정하였다. 선정한 독립변수는 성별, 거주지역, 치면열구전색치아 수, 주관적 구강건강상태, 최근 치과방문여부(지난 1년간 치과진료 여부), 치통 여부(지난 1년간 치통 경험 여부), 1일 우식성간식 섭취횟수(어제 하루 동안), 이용 중인 구강위생보조용품 수는(치실, 치간칫솔, 구강세정액, 전동칫솔, 혀클리너, 기타 용품 중 복수선택)이었다. 다중선형회귀분석에서 각 변수들은 연속형 변수로 취급되었다.

다중선형회귀분석을 통해 도출된 함수에 각 대상자의 변수값을 입력하여 대상자별 예측 DMFT 값을 도출하였다. 산출된 예측값은 소수 첫째 자리에서 반올림하여 정수 변환하였으며, 음의 예측 값은 0으로 처리하였다. 통계분석은 IBM SPSS 23.0® (IBM Co., Armonk, NY,

USA)을 이용하였고, 유의수준은 제1종 오류 0.05로 판정하였다.

2.2. DMFT 예측을 위한 머신러닝 알고리즘의 활용

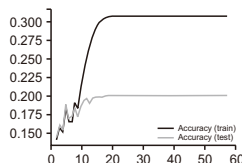
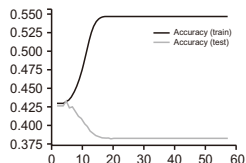
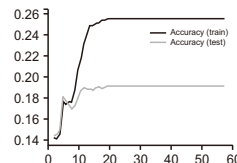
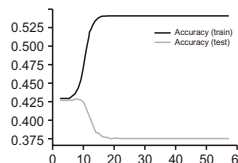
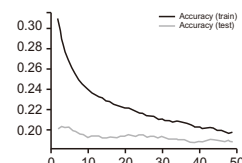
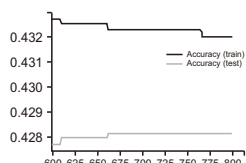
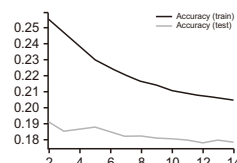
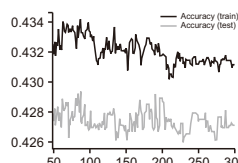
(1) 의사결정나무(Decision tree)

의사결정나무(Decision tree)는 데이터마이닝, 머신러닝에서 사용하는 예측 모델링 방법 중 하나로 여러 가지 규칙을 순차적으로 적용하면서 독립변수 공간을 분할하는 분류 모델이다. 즉, 주어진 입력 값에 대하여 출력값을 예측하는 모델인데 나무형태의 그래프를 띄고 있다. 분석 대상 자료의 탐색과 추론을 위한 모델 구성의 2가지 특성을 동시에 달성할 수 있는 기법으로 많은 분야에서 다양하게 적용하고 있다. 이번 연구에서는 파이썬 프로그래밍 언어(version 3.7.0, <https://www.python.org>)의 대표적인 머신러닝 라이브러리인 싸이킷런(scikit-learn, <https://scikit-learn.org>)의 DecisionTreeRegressor (이하 DTR)와 DecisionTreeClassifier (이하 DTC) 알고리즘을 활용하였다. Classification은 대상이 어느 범주에 속하는지 예측하는 반면에, Regression은 대상과 관련이 있는 연속형 값을 예측하는 알고리즘이다.

(2) 랜덤 포레스트(Random forest)

랜덤 포레스트(Random Forest)는 2001년 Breiman에 의해 개발한 분류기법으로²⁸⁾, 전통적인 의사결정나무 기법을 하나가 아닌 여러 개의

Table 2. Evaluation of validity and accuracy of machine learning algorithm for prediction of DMFT

	Decision tree regressor	Decision tree classifier	Random forest regressor	Random forest classifier
Self accuracy	0.279808	0.526285	0.240064	0.521813
Model 1	Train: 0.309204 Test: 0.201238 Whole: 0.276812	Train: 0.546983 Test: 0.382225 Whole: 0.497553	Train: 0.255379 Test: 0.191253 Whole: 0.236140	Train: 0.540896 Test: 0.374912 Whole: 0.491098
Model 2	 Max_depth=23 Train: 0.309204 Test: 0.201238 Whole: 0.276812	 Max_depth=7 Train: 0.440540 Test: 0.423991 Whole: 0.435575	 Max_depth=20 Train: 0.254958 Test: 0.191253 Whole: 0.235845	 Max_depth=7 Train: 0.435477 Test: 0.427929 Whole: 0.433212
Model 3	 Min_samples_split=3 Train: 0.289856 Test: 0.203066 Whole: 0.263817	 Min_samples_split=675 Train: 0.432283 Test: 0.428210 Whole: 0.431061	 Min_samples_split=2 Train: 0.254958 Test: 0.191253 Whole: 0.235845	 Min_samples_split=88 Train: 0.434211 Test: 0.429335 Whole: 0.432748

Random forest regressor, Decision tree regressor: random_state=37. Random forest classifier, Decision tree classifier: random_state=42. Blue (Orange) line: train (test) dataset accuracy by each hyperparameter. Model 1: After random train-test data set split (7:3), the accuracy of each data set was obtained. Model 2: max_depth adjusting in Model 1. Model 3: min_samples_split adjusting in Model 2.

나무로 확장시킨 의사결정나무의 메타학습(meta-learning) 형태를 갖고 있는 머신러닝 기법이다⁸⁾. 랜덤 포레스트를 구성하는 각 의사결정나무는 무작위로 선택된 학습 데이터와 입력변수들에 의해 형성되는데, 이 경우 각 개별 의사결정나무의 정밀도는 떨어질 수 있으나, 이들을 종합하여 예측을 수행하게 되는 숲(랜덤 포레스트)의 정도와 안정성은 높아지게 된다. 랜덤 포레스트의 경우, 대수의 법칙에 의해 숲의 크기(나무의 수)가 커질수록 일반화 오류가 특정 값으로 수렴하게 되어 과적합(over fitting)을 피할 수 있으며, 각 개별 의사결정나무들을 학습시킬 때 전체 학습용 자료에서 무작위로 복원 추출된 데이터를 사용하고 있어 잡음이나 이상치로부터 크게 영향을 받지 않는다^{8,9)}. 본 연구에서는 파이썬 프로그래밍 언어(version 3.7.0)의 머신러닝 라이브러리인 싸이킷런의 RandomForestRegressor (이하 RFR)와 RandomForestClassifier (이하 RFC) 알고리즘을 활용하였다.

(3) 머신러닝 알고리즘 예측 모델의 적용

DMFT를 종속변수(label), 8개의 변수를 설명변수(feature)로 설정한 뒤, 파이썬 프로그래밍 언어의 머신러닝 라이브러리인 Scikit-learn 소프트웨어의 DTR, DTC, RFR, RFC 알고리즘 각각에 대하여 23,702명의 자료를 학습시킨 뒤, 구축된 학습모델에 기존의 설명변수 값들을 재적합(fit)하여 예측값들의 정확도를 산출하였다. Regressor 알고리즘(DTR, RFR)에 의해 산출된 예측값은 소수 첫째 자리에서 반올림하여 정수 변환하여 처리하였다.

모델의 타당성 평가를 위해 전체 데이터셋을 학습용(train data set)과 평가용(test data set)으로 구분하여 7대 3의 비율로 무작위로 쪼갬 뒤, hold out validation을 시행하여 train data set과 test data set에서의 예측 정확도를 산출하였다(Table 2, model 1). 예측 모델의 정확도 향상을 위해 알고리즘을 미세 조정하는 과정인 하이퍼파라미터 튜닝(hyperparameter tuning, 초매개변수 조정)을 시행하였다. 튜닝한 하이퍼파라미터는 max_depth와 mean_samples_split였다. Test data-set의 정확도가 가장 높게 나타나는 hyperparameter 값을 모델에 적용하여, 최종적인 학습모델을 구성하였다(Table 2, model 3). 이 학습모델에 기존의 설명변수 값을 적합시켜 얻어진 예측값을 통

해 모델의 최종적인 정확도를 산출하였다.

연구 성적

1. 다중선형회귀분석 모델을 통한 DMFT 예측

DMFT 예측을 위한 회귀 모델을 도출하기 위하여, 선택된 8개의 변수를 독립변수로 한 다중선형회귀분석을 시행한 결과, 모델에 포함된 모든 변수가 통계적으로 유의하였으며, 설명력은 11.9%로 나타났다($R^2=0.119$). 치면열구전색 치아수($\beta=-0.224$), 최근치과방문 여부($\beta=0.175$), 주관적 구강건강상태($\beta=0.117$)가 상대적으로 DMFT에 큰 영향을 미치는 것으로 나타났다(Table 3).

도출된 회귀모델을 통해 각 대상자별 예측값을 다시 산출하였다. 예측값이 음수로 나온 206개의 사례는 0으로 변환하였다. 실제 값과 예측값이 동일한 경우는 3,611명으로 전체 대상자수(23,702명) 대비 15.24%의 예측 정확도를 나타냈다. 회귀분석이 DMFT 0으로 예측한 경우, 그 값이 실제 0일 확률(진성 예측도)은 64.9%였고, DMFT 1로 예측한 경우, 그 값이 실제 1일 확률은 14.9%, DMFT 2로 예측한 경우, 그 값이 실제 2일 확률은 12.7%로 나타났다(Table 4).

2. 4가지 머신러닝 알고리즘의 DMFT 예측 비교

이번 연구 데이터셋에 대하여 머신러닝 알고리즘을 적용한 결과, DMFT 예측 정확도는 DTC와 RFC 알고리즘이 각각 43.11%, 43.27%를 나타냈으며, DTR, RFR 알고리즘은 각각 26.38%, 23.58%를 나타냈다(Table 2). Regressor 알고리즘보다 Classifier 알고리즘의 정확도가 뚜렷하게 높았으며, RandomForest 알고리즘과 Decision-Tree 알고리즘의 정확도는 큰 차이가 없었다. Hyperparameter 튜닝을 통해 test dataset의 정확도를 증가시켰다. 다중선형회귀분석의 예측정확도(15.24%)에 비해 DTR 알고리즘은 1.73배, RFC 알고리즘은 2.84배 높게 나타났다.

각 알고리즘의 설명변수 중요도(feature importance)는 서로 상이하였으며, 모든 알고리즘에서 치면열구전색치아수의 비중이 가장 높

Table 3. Multiple linear regression model for prediction of DMFT in 12-year-olds

Model	Unstandardized		Standardized	t	Sig.
	coefficient		coefficient		
	B	Std. Error	Beta		
(Constant)	.302	.075		4.006	.000
Gender	.519	.032	.098	16.030	.000
Region	.220	.039	.034	5.632	.000
Number of pit and fissure sealant	-.235	.006	-.224	-36.394	.000
Perceived oral health status	.407	.022	.117	18.515	.000
Dental treatment demand for the past one year	.966	.034	.175	28.128	.000
Experience of toothache for the past one year	.350	.041	.054	8.550	.000
Frequency of snack intake per day	.035	.015	.014	2.329	.020
Number of oral hygiene auxiliaries using	.074	.022	.021	3.388	.001

R Square: 0.119, Dependent variable: DMFT, Variables Entered with Enter method.

Gender: Men=0, Women=1 / Regime: City=0, Rural area=1 / Number of pit and fissure sealant: 0-16 / Perceived oral health status: Very good-Very poor=1-5 / Dental treatment demand for the past one year: Yes=1, No=0 / Experience of toothache for the past one year: Yes=1, No=0 / Frequency of snack intake per day: No intake=1, once=2, 2 times=3, 3 times=4, 4 and over=5 / Number of oral hygiene auxiliaries using: 0-5.

Table 4. Accuracy (%) of the predicted DMFT in each machine learning algorithm

DMFT	Frequency of DMFT	Predicted DMFT	MLR	DTR	DTC	RFR	RFC
0	10,140	0	64.9	87.0	48.2	86.8	54.7
1	2,932	1	14.9	21.0	20.9	19.3	57.6
2	2,773	2	12.7	16.6	17.0	15.3	51.7
3	1,764	3	9.3	12.6	5.1	10.8	45.7
4	2,963	4	20.8	27.5	13.0	23.3	37.1
5	866	5	0	19.8	4.6	7.9	50.2
6	798	6		30.4	10.2	12.7	44.7
7	381	7		32.2	12.7	3.8	43.7
8	384	8		52.8		3.8	44.1
9	185	9		68.4		4.3	71.4
10	135	10		78.6		11.1	72.2
11	98	11		100.0		0	56.3
12	94	12		66.7		0	77.8
13	66	13		100.0		50.0	35.3
14	54	14		100.0			66.7
15	25	15		100.0			44.4
16	34	16		100.0			100.0
17	4						
18	1						
19	4						
20	1						
Total	23,702						

MLR, Multiple linear regression; DTR, Decision tree regressor; DTC, Decision tree classifier; RFR, Random forest regressor; RFC, Random forest classifier.

Table 5. The feature importance of each machine learning algorithm

	Decision tree regressor	Decision tree classifier	Random forest regressor	Random forest classifier
Gender	0.0480	0.0455	0.0501	0.0275
Region	0.0492	0.0902	0.0603	0.0620
Number of pit and fissure sealant	0.3460	0.2298	0.3132	0.3438
Perceived oral health status	0.1109	0.1232	0.1089	0.1342
Dental treatment demand for the past one year	0.1313	0.0372	0.1117	0.0447
Experience of toothache for the past one year	0.0441	0.0635	0.0598	0.0405
Frequency of snack intake per day	0.1527	0.2170	0.1642	0.1868
Number of oral hygiene auxiliaries using	0.1178	0.1936	0.1318	0.1604

았으며, 성별, 치통 여부, 거주지역은 비중이 낮게 나타났다(Table 5).

머신러닝 알고리즘이 DMFT=0으로 예측한 경우, 그 값이 실제로 DMFT=0인 비율은 DTR 87.0%, DTC 48.2%, RFR 86.8%, RFC 54.7%였다. 머신러닝 알고리즘이 DMFT=1로 예측한 경우, 그 값이 실제로 DMFT=1인 비율은 DTR 21.0%, DTC 20.9%, RFR 19.3%, RFC 57.6%였다. 머신러닝 알고리즘이 DMFT=2로 예측한 경우, 그 값이 실제로 DMFT=2인 비율은 DTR 16.6%, DTC 17.0%, RFR 15.3%, RFC 51.7%였다(Table 4). 본 데이터셋에 대하여 회귀 알고리즘(DTR, RFR)은 DMFT=0의 예측 정확도가 약 87%로 높지만, 이외의 DMFT 예측은 20% 대로 낮았다. 반면, 분류 알고리즘인 RFC는 DMFT=0부터 8까지의 예측 정확도는 평균적으로 약 48%로 고르

게 나타났다(Table 4).

고 안

2015년 아동구강건강실태조사 자료를 활용하여, 우식경험영구치아수(DMFT)에 많은 영향을 미치는 변수를 찾고, 이를 토대로 DMFT 예측하는 회귀분석과 머신러닝 모델을 구축하고, 각각의 예측 정확도를 비교하였다. 머신러닝을 통한 예측은 매우 다양한 분야에서 활용되고 있다. 예를 들면 주변교통이나 편의시설, 유휴시설, 교육시설 등을 학습시켜 지가나 주택가격을 예측한다거나 기상요소나 사회경제적 요소를 학습시켜 호우피해를 예측하거나, 노년층의 뇌졸중을 예측하는

등 다양한 분야에서 도입되고 있다¹⁰⁻¹³. 치의학 분야에서도 전문가에 의한 정밀 진단을 수행하기 전에 설문조사나 기본적인 정보만으로 인구집단의 DMFT를 예측하여 우식 고위험 집단을 분류할 수 있다면, 구강검진에 소요되는 인력이나 시간, 비용을 크게 줄일 수 있을 것이고, 일반인들이 치과에 내원하여 필요한 치료를 받을 수 있도록 하는데도 도움이 될 것이다. 또한, DMFT에 많은 영향을 미치는 기여요인을 특정할 수 있다면 그 요인을 조절하여 우식 예방에 활용할 수도 있을 것이다. 이번 연구는 머신러닝을 도입하여 23,702명의 방대한 데이터를 가지고 성별, 농어촌 거주 여부, 치면열구전색치아수, 주관적 구강건강상태, 최근 치과방문 여부, 최근 1년간 치통 경험 유무, 1일 우식성 간식섭취 횟수, 이용 구강위생보조용품 종류 수의 설명변수로 DMFT를 예측하는 모델을 구축했다는 데에 의미가 있다.

DMFT 예측 정확도는 다중선형회귀분석이 15.24%인 반면, 회귀(Regressor) 알고리즘인 RFR, DTR 알고리즘은 23.58%, 26.38%로 예측정확도는 다소 증가하였으며, 분류(Classification) 알고리즘인 DTC, RFC 알고리즘은 43.11%, 43.27%로 가장 높게 나타났다(Table 2). 회귀분석이란 변수들 간의 관련성을 파악하기 위하여 수학적 모델을 세우고 선정된 변수의 실측 데이터로부터 모델을 추정하는 통계방법이다¹⁴. 즉, 선형 회귀 모델은 다른 변수에 영향을 주는 설명(독립)변수와 영향을 받는 종속변수 간의 관계가 직선적이라는 것을 가정한다. 이러한 직선적 관계를 수식($Y=a+bX$)으로 나타낸 것이 선형 회귀모델이며, 다중선형 회귀모델은 종속변수(Y)의 총 분산을 더 많이 설명하기 위해 여러 개의 설명변수(X_1, X_2, \dots, X_n)를 투입하는 통계 방법이다¹⁴. 독립변수와 설명변수의 관계로부터 도출된 가장 적합한 회귀식이라 하더라도, 도출된 회귀선 상의 예측값과 실측치 간의 차이(잔차, residual)가 존재하며, 이러한 잔차의 크기가 커질수록 추정된 회귀식의 설명력(R^2) 또는 정확도를 낮아진다. 반면, 머신러닝 알고리즘은 전체 자료 분석으로부터 일반화된 하나의 회귀식을 도출하는 것이 아니라, 동일한 특성(label, 관심 종속변수)을 나타내는 다수의 개별 데이터의 잔차를 반복 학습하고 기존 학습 결과와 최소의 차이를 나타내는 최적의 예측값을 도출하는 과정이라 할 수 있다. 따라서 머신러닝 알고리즘은 관심 개별 데이터 영역의 자료 특성을 더 잘 반영할 수 있으며, 이러한 특성으로 인해 회귀분석보다 머신러닝의 예측정확도가 높게 나타난 것으로 여겨진다.

우식 예측을 위한 머신러닝 알고리즘 모델을 활용한 선행 연구들은 다음과 같다. Montenegro 등¹⁵은 브라질 5세 이하 3,864명을 대상으로 구강검사와 부모 설문을 통해 총 193 종류의 설명변수(feature)를 수집하고, 이 중 우식 존재 여부(종속변수)와 관련있는 15개의 변수를 선정하여 다양한 머신러닝 예측 모델을 구축하였다. 이 연구에서 활용한 알고리즘은 decision tree, MLP (Multi Layer Perceptron) neural network, kNN, SVM이었으며, 10-fold cross-validation error 비교를 통해 각 모델의 성능을 비교하였다. 그 결과 MLP neural network 알고리즘의 error가 22.75%로 가장 낮게 나타나 가장 우수한 성능을 보였으며, 모델에서 치통경험 여부, 우식진단 여부, 소득 수준, 과거 우식 경험에 현재 우식 유무와 관련이 있음을 보고하였다. Tamaki 등¹⁶은 일본 5-8세 학교아동 500명의 매년 정기검진과 3년 팔로우업을 통해 수집한 자료 중 타액 속 mutans streptococci

(MS)와 lactobacilli (LB) 레벨, 3분 자극성타액 분비량, 타액 pH, 불소 사용, 간식 및 음료 섭취 횟수를 설명변수, 새로운 우식발생 유무를 종속변수로 하는 모델을 구성하였으며, 로지스틱회귀분석, Neural network, C 5.0 Decision analysis 모델의 성능을 10-fold cross-validation을 통해 비교하였다. 새로 우식 발생한 74명에 대하여 무작위 선정 70명을 선정하여 민감도와 특이도를 비교하였으며, 이와 같은 무작위 선정과정을 10번 반복하여 최적의 모델을 구축하였다. 로지스틱회귀분석의 민감도와 특이도는 0.618, 0.698이었으며, Neural network는 0.838, 0.458, Decision analysis는 0.730, 0.773으로 의사결정나무(Decision analysis) 알고리즘이 가장 우수한 성능을 보였다. 반면, Gansky¹⁷는 Rochester Careis Study 자료를 활용하여 우식 예측 모델을 비교한 결과 neural network 알고리즘이 로지스틱회귀분석, decision analysis보다 우수한 예측 성능을 나타냈다고 보고하였다. Ito 등¹⁸은 1993년부터 2008년까지 치과외원에 내원한 환자 중 442명으로부터 DMFT, 타액 속 MS, LB 레벨, 예방프로그램의 순응도를 조사하고 Classification and Regression Trees (CART) 알고리즘을 활용하여 일차우식 발생유무, 이차우식 발생유무를 각각 종속변수로 하는 모델을 각각 구축하여, 우식 고위험, 저위험군을 예측하였다. 그 결과 CART가 이차 우식 발생 유무를 분류함에 있어 유용한 모델(민감도 0.718, 특이도 0.730)임을 제시하였다. 한편, 병동 환자를 대상으로 병세의 악화를 예측한 Churpek 등¹⁹의 연구에서도 회귀분석 예측 모델보다 랜덤포레스트 예측 모델의 예측 정확도가 더 높게 나타났다. 이와 같은 선행 연구 고찰 결과를 통해 통계 회귀분석 모델보다 머신러닝모델의 예측 성능이 뛰어남을 알 수 있으며, 각 연구에서 사용된 분석 자료의 특성에 따라 머신러닝 알고리즘 별 성능의 정확도는 달라질 수 있음을 알 수 있다. 본 연구에 사용된 데이터세트는 2015년 우리나라 12세 아동의 국가실태조사자료를 기반으로 하고 있으며, 선행연구와 마찬가지로 회귀분석보다 머신러닝 알고리즘의 예측성능이 우수함을 확인하였다.

이번 연구 데이터세트를 이용한 머신러닝 알고리즘 간의 비교 결과, 예측정확도는 회귀 알고리즘(DTR, RFR)보다 분류 알고리즘(DTC, RFC)에서 높게 나타났다(Table 2). DMFT는 우식경험영구치아수로 이론적으로 0부터 28 (사랑니 제외)의 값을 가질 수 있으며, 치아수 관련 변수는 여러 연구에서 연속형의 변수로 정의하여 분석하기도 한다^{20,21}. 이번 연구에서도 DMFT를 연속형 변수로 가정하고 다중선형회귀분석을 시행하였고, 머신러닝 알고리즘에서도 회귀 알고리즘을 적용하였다. 이를 통해 도출된 예측값은 소수값으로 나타났으며, 실질적인 정확도 비교를 위해 소수값으로 도출된 예측값을 소수 첫째 자리에서 반올림하여 정수변환 한 뒤, 실제값과 정수 변환된 예측값을 비교하였다. 이러한 분석과정은 근거와 이론에 기반한 타당한 분석 과정이지만, 실제 DMFT의 분포는 0, 1, 2, 3, 4에 편중(Table 4)되어 있으며 21 이상의 DMFT는 존재하지 않기 때문에 진정한 연속형 변수라 할 수 없다. 또한 소수값을 반올림하여 정수로 변환하는 과정에서 발생하는 결과값의 손실은 예측정확도에 영향을 미칠 수 있다. 반면 분류 알고리즘은 각 DMFT 값을 하나의 집단으로 가정하고, 예측한 결과를 제시하기 때문에, 인위적인 자료 변환의 개입 없이 단순한 정확도 비교가 가능하며 직관적이다. 이로 인해 이번 연구데이터세트에 대하여 회귀보

다 분류 알고리즘의 정확도가 높게 나온 것으로 판단되며, DMFT 예측을 위해서는 분류 알고리즘의 활용이 적절할 것이다. 하지만, 분류 알고리즘이 회귀 알고리즘보다 전체적인 정확도는 높게 나타났지만, 예측 DMFT가 0일 때의 진성예측도는 DTR이 87.0%, RFR이 86.7%로 상당히 높게 나타났지만, DTC와 RFC는 각각 48.2%, 54.7%로 낮게 나타났다(Table 4). 따라서 회귀 알고리즘은 특이도(질병이 없을 때 없다고 예측할 가능성)는 높지만, 민감도(질병이 있을 때 있다고 예측할 가능성)는 낮은 것으로 예상할 수 있으며, 우식이 없는 건강한 아동(DMFT=0)을 예측할 목적에는 회귀 알고리즘이 적절하며, 실질적인 예측값을 산출하기 위한 목적에는 분류 알고리즘, 그중에서도 DTC보다는 RFC가 적절할 것이다.

이번 연구에서는 다양한 머신러닝 알고리즘 중 랜덤포레스트(Random Forest) 알고리즘과 의사결정나무(Decision Tree) 알고리즘을 선택하여 모델을 구성하였다. 의사 결정 나무는 데이터의 분리 규칙을 찾아내고 여러 단계의 의사결정 과정을 거쳐 결과값을 도출하여 그 과정을 나뉘어 가지 형태로 표현할 수 있는 알고리즘이다. 의사결정 나무는 범주형 데이터를 예측하는 classifier와 숫자형 결과를 예측하는 regressor로 나눌 수 있다. 의사결정나무의 단점은 새로운 데이터에 대한 일반화 성능이 좋지 않아 과적합되기 쉽다는 것인데, 이를 보완하는 것이 랜덤포레스트이다. 랜덤포레스트는 여러 개의 의사결정 나무를 만들고 그들의 다수결로 결과를 선택하는 방법이다^{8,22}. 일반적으로 의사결정나무보다 랜덤포레스트 알고리즘의 예측정확도 성능이 더 좋다고 알려져 있다. 본 연구에서는 DMFT 예측에 있어 두 알고리즘 간의 전체 정확도의 뚜렷한 차이를 보이지 않았지만, 각 예측값에 대한 진성예측도 측면에서 비교했을 때는 RFC가 DTC보다 높게 나타남을 확인하였다(Table 4). 결과적으로 본 연구데이터세트에 대해서 DMFT 예측에 있어 가장 효과적인 알고리즘은 랜덤포레스트 분류(RFC) 알고리즘으로 생각된다.

머신러닝 알고리즘 모델에 활용된 8개의 변수의 설명변수 중요도(feature importance)를 도출하였다. 이 개념은 회귀분석의 표준화 회귀계수와 같은 개념으로 각 변수들 간의 상호작용을 고려한 변수들의 상대적 중요성을 비교하는데 도움이 된다¹⁴. DMFT 예측에 있어서 회귀분석의 표준화 회귀계수(베타)는 치면열구전색치아 수(-0.224), 최근 치과방문여부(0.175), 주관적 구강건강상태(0.117), 성별(0.098), 치통 여부(0.054), 거주지역(0.034), 이용 구강위생보조용품 종류 수(0.021), 1일 우식성간식 섭취횟수(0.014) 순으로 나타난 반면(Table 3), RFC의 변수 중요도는 치면열구전색치아 수(0.3438), 1일 우식성간식 섭취횟수(0.1868), 이용 구강위생보조용품 종류 수(0.1604), 주관적 구강건강상태(0.1342), 거주지역(0.0620), 최근 치과방문여부(0.0447), 치통 여부(0.0405), 성별(0.0275)로 나타났다(Table 5). 치면열구전색치아 수는 회귀분석과 머신러닝 모두 가장 중요한 변수로 선정한 반면, 회귀분석이 가장 중요도를 낮게 본 [이용 구강위생보조용품 종류 수], [1일 우식성간식 섭취횟수] 변수는 RFC에서는 치면열구전색치아수 다음으로 3번째와 2번째로 중요한 변수로 채택되었다. 이처럼 변수 간의 중요성을 달리 계산하는 각 모델의 특성은 예측정확도의 차이의 원인이 될 수도 있을 것이다. 본 연구의 머신러닝 모델에서 중요한 비중을 차지하는 변수인 치면열구전색치아 수와 1일 우식

성간식 섭취횟수는 기존의 여러 연구에서도 우식 발생 관련성이 높다고 보고되었으며²³⁻²⁶, 향후 연구에서도 우식예측 변수 선정에 있어 우선적으로 고려될 필요가 있을 것이다.

머신러닝 알고리즘은 학습 자료(training data set)의 설명변수와 종속변수 기반으로 반복학습을 통해 모델을 구축하며, 이 모델에 학습 자료의 설명변수를 재적합하여 새로운 예측값을 도출하면, 새로운 예측값의 과적합(over-fitting) 문제가 제기된다^{9,27}. 머신러닝 알고리즘의 학습 횟수를 증가시키면 증가시킬수록 모델의 학습자료에 대한 정확도는 증가(과적합)하지만, 현실 자료에 적용했을 때의 예측의 정확도는 낮아질 수 있다. 따라서 학습자료를 통해 구축된 모델의 과적합 여부를 평가하기 위해, 머신러닝 데이터사이언스 분야에서는 모델의 타당성(validation) 평가를 위해 기존 학습자료를 쪼갬 뒤(train-test split) 모델에 재적합시키는 방법을 사용한다. 데이터를 7:3, 8:2로 무작위로 나눈 뒤 평가하는 hold-out validation과 학습자료의 양에 따라 자료를 10% 또는 20%를 무작위로 선별한 뒤, 선별된 자료와 나머지 자료를 비교하고 반복하여 평균값으로 나타내는 cross validation 방법이 대표적이다. 전자는 간편하고 빠르게 모델의 과적합도를 평가할 수 있는 반면, 후자는 모델의 성능을 보다 향상시킬 수 있지만, 고성능의 하드웨어와 오랜 작업시간을 필요로 한다. 본 연구에서는 7:3 hold-out validation을 통해 모델의 과적합 정도를 평가하였으며, 이를 통해 학습(train) 자료와 평가(test) 자료, 그리고 전체(whole) 자료의 예측 정확도가 비슷한 값을 산출한 것을 확인하였으며, 모델의 성능을 효과적으로 검증하였다. 향후 연구에서는 cross validation 방법을 통해 모델의 예측 성능을 높이는 시도를 할 수 있을 것이다.

이번 연구에서는 알고리즘의 적합도 향상을 위해 max_depth와 min_samples_split 하이퍼파라미터만을 조정하였고 다른 하이퍼파라미터에 대한 고려를 하지 않았다. max depth를 조정하였을 때 예측 정확도가 상당히 상승한 반면 mean sample split을 조정하였을 때 test data set에 대한 정확도의 차이가 크지 않았다. 이외에도 각 알고리즘별로 다양한 하이퍼파라미터가 존재하므로 이들을 조정하면서 정확도를 산출하는 일을 반복한다면 보다 높은 예측 정확도를 갖는 모델을 구축할 수 있을 것이다. 또한 본 연구에서 적용한 랜덤포레스트 이외에도 좋은 성능을 나타내는 알고리즘(Gradient Boosting, Light GBM 등)이 존재한다. 따라서 향후 연구에서는 우식위험군 예측 정확도 향상을 위한 알고리즘의 보정 및 새로운 알고리즘의 적용도 시도해 볼 필요가 있을 것이다.

이번 연구에서는 모델의 정확도 평가를 위해 실제값과 예측값의 일치도(accuracy)를 이용하였다. 정확도 평가 지표는 직관적인 장점이 있기 때문에 분류 예측 모델에서 널리 활용되며, 특히 2집단 예측(예: 암 발생 유, 무)에 주로 활용된다. 하지만, 전체 연구집단에서 암 발생자의 비율이 적은 경우와 같이 imbalanced data set에는 적절하지 않다. 이 경우 집단을 구분하여 True Positive, True Negative, False Positive, False Negative False 산출하는 Confusion matrix를 사용하지만, 결과를 하나의 값으로 제시하지 않기 때문에 직관적이지 않다. 결과값을 하나로 제시하여 비교가 편리한 Per Class Accuracy, 즉 민감도, 특이도를 사용할 수도 있다. 앞서 소개한 이러한 평가지표는 종속변수를 유무로 구분한 경우 유용하게 사용되는 경우이다. 종속

변수의 구분이 여러 개로 되어 있는 경우에는 log loss라는 평가 지표가 추천된다. log loss는 imbalanced dataset에도 적용 가능하며, 각 종속변수의 확률을 기준으로 계산되는 값으로 정확도가 높아질수록 0에 가까워지며, 정확도가 낮아질수록 기하급수적으로 증가한다. log loss는 accuracy에 비해 직관적이지 않은 단점이 있지만, 본 연구 데이터셋에는 보다 적절한 평가지표일 수 있다. 따라서 추후 연구에서는 모델의 정확도 평가를 위해 accuracy 뿐만 아니라, log loss를 적용해 볼 수 있을 것이다.

한편, 성능이 우수하다고 알려진 랜덤포레스트 분류 알고리즘을 적용하였음에도 불구하고 본 연구의 정확도는 DMFT 예측에 있어 약 43%, 우식위험군 예측에 있어 약 52%로 그리 만족스러운 값은 아니다. 이러한 예측 정확도는 본 연구데이터셋의 특성에 기인한 것으로 여겨진다. 23,702명의 방대한 자료를 사용하였지만, DMFT 값은 0-20 사이에만 존재하고, 그 값이 0이 대상자는 10,140명으로 전체 대상자의 약 절반(42.8%)에 이르고 0-4 사이에 치우쳐 있다. 즉, 아무런 모델 구현 없이 모두 0으로 예측하여도 42.8%는 맞출 수 있다는 가정이 성립한다. 또한 각 대상자의 8개의 설명변수 값은 개인에 따라 다양한 값을 가진다. 우식과 관련된 요인에는 생물학적 요인, 사회경제적 요인, 구강건강행동요인 등 다양한 요인이 관여하기 때문에, 본 연구에서 선정한 8개의 설명변수만으로 완벽한 모델을 구현하기에는 한계가 존재한다. 이로 인해 성능이 우수한 머신러닝 모델임에도 예측 정확도가 높게 나타나지 않은 것으로 여겨지며, 머신러닝의 우식 예측 정확도 향상을 위해서는 우식과 관련 있는 다양한 요인을 선정하고 양질의 자료 수집이 선행될 필요가 있을 것이다. 또한 수집된 자료의 이상치에 대한 자료 검토를 통해 양질의 데이터셋을 구축할 필요가 있다.

비록, 머신러닝 알고리즘 모델의 정확도가 비록 만족스럽지는 못하였지만, 그동안 치의학분야에서 우식 예측을 위해 우선적으로 고려됐던 회귀분석보다 머신러닝 알고리즘(RFC)이 2.84배 이상의 정확도를 나타냈다는 것은 의미 있는 결과이다. 이러한 연구 결과를 바탕으로 향후 머신러닝 알고리즘 모델 구성에 있어서 설명변수를 수정(제거, 추가, 범주 변경)하거나 기존의 알려진 알고리즘(Support Vector machine, K-Nearest Neighbors, Naive Bayes 등)뿐만 아니라, 새롭게 개발된 성능 좋은 알고리즘(예 Gradient Boosting, LightGBM 등) 적용, 비교하고, 하이퍼파라미터 조정과 cross validation 과정을 통해 정확도가 최고로 향상된 최적의 머신러닝 모델을 구축할 수 있을 것이다.

머신러닝을 기반으로 설문과 구강건강 조사 자료를 바탕으로 치아 우식 발생을 예측하는 연구는 거의 수행되지 않았다. 지금까지 대부분의 조사자료는 샘플 수가 작았고 통계분석만으로 가설을 증명 수준에 머물러 있었다. 치의학 분야, 특히 공중구강보건학 분야에서 머신러닝 분석 방법론을 시도한 사례는 매우 드물며, 국내에는 전무했다. 머신러닝이 다양한 분야에서 활용되어 뛰어난 성능을 보여주고 있는 만큼, 치아 우식 발생을 예측하여 예방과 치료에 적극 활용하려는 시도는 의미가 있으며, 앞으로 이 분야에 대한 활발한 연구가 필요할 것으로 생각된다.

결론

이번 연구는 우리나라 구강의학분야에서 빅데이터로 간주되는 2015년 아동구강건강실태조사 자료를 활용하여, 12세 아동의 DMFT와 우식위험군을 예측하는 통계적 회귀분석 모델과 네 가지 서로 다른 머신러닝 모델을 구축하고, 예측 정확도를 평가하였으며, 주요 연구 결과는 다음과 같았다.

1. DMFT 예측 모델에서 다중선형 회귀분석과 RandomForest-Classifer (RFC) 알고리즘의 예측정확도가 각각 15.24%, 43.27%로 RFC의 예측 정확도는 회귀분석보다 2.84배 높았다.

2. DMFT를 예측하는 머신러닝 모델에서 가장 비중이 높은 설명변수는 치면열구전색치야 수였다.

이번 연구의 결과로 통계적 회귀분석 모델보다 머신러닝모델이 DMFT와 우식위험군 예측에 더 유용한 모델임을 확인하였다. 따라서 머신러닝 모델을 활용하여 개인 또는 집단의 DMFT를 예측하고 우식 위험군으로 분류하는 데 활용할 수 있을 것으로 기대한다. 향후 연구에서는 예측의 정확도를 높이기 위해서 학습데이터를 가공하고 알고리즘을 수정하는 과정이 필요할 것이다.

ORCID

Yong-Hoon Yang, <https://orcid.org/0000-0002-7403-1200>

Ji-Soo Kim, <https://orcid.org/0000-0003-1571-4762>

References

- Petersen PE, Bourgeois D, Ogawa H, Estupinan-Day S, Ndiaye C. The global burden of oral diseases and risks to oral health. Bull World Health Organ 2005;83:661-669.
- Fejerskov O, Nyvad B, Kidd EAM. Dental caries: what is it. In: Fejerskov O, Nyvad B, Kidd EAM. Dental caries. 3th ed. West sussex: John Wiley & Sons, Ltd;2015:7-10.
- Ministry of Health & Welfare. 2015 Korean Children's Oral Health Survey. Sejong: Ministry of Health & Welfare;2015.
- Seul MS. Current status and future developments of machine learning artificial intelligence in law: focusing the cusp of machine learning in U.S. and discourses over legal profession and law school education. The Justice 2016;156:269-302.
- National library of Korea. National library of Korea digital collection. Leading the fourth industrial revolution-artificial intelligence and deep learning [Internet]. [cited 2019 Jan 02]. Available from: <http://nlcollection.nl.go.kr/front/search/searchList.do?facet=&indent=&query=%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5%2C+%EB%8D%B0%EC%9D%B4%ED%84%B0&facetField=true&wt=json&searchPageType=main&searchKeyword=%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5%2C+%EB%8D%B0%EC%9D%B4%ED%84%B0&searchSelect=all&searchFacet=&solrStart=0&solrEnd=20&solrRows=20>.
- Lee JH, Kim DH, Jeong SN, Choi SH. Detection and diagnosis of dental caries using a deep learning-based convolutional neural network algorithm. J Dent 2018;77:106-111.
- National Science & Technology Information Service. Artificial intelligence, dentistry [Internet]. [cited 2019 Dec 17]. Available from: <https://www.ntis.go.kr/ThSearchTotalList.do?sort=RANK%2FDESC&>

- ntisYn=&searchWord=%EC%9D%B8%EA%B3%B5%EC%A7%80%EB%8A%A5%2C+%EC%B9%98%EA%B3%BC.
8. Kim SJ, Ahn HC. Application of random forests to corporate credit rating prediction. *The Journal of Business and Economics* 2016;32:187-211.
 9. Han EJ. Screening test data analysis for cataract happening prediction model using random forest [master's thesis]. Seoul:Yonsei University;2005. [Korean].
 10. Yoo JH, Hong SH, Park HG, Kim DM, Kim SJ, Park SJ. Utilization of elderly stroke disease prediction using machine learning method. *Korean Society for Emotion and Sensibility 2017 Annual spring conference program* 2017:2.
 11. Won SH, Lee CG, Park JM. A study on the prediction of land price with machine learning technique. *Journal of the Korean Association of Professional Geographers* 2017;51:347-355.
 12. Bae SW, Yu JS. Estimation of the apartment housing price using the machine learning methods: the case of gangnam-gu, seoul. *Journal of the Korea Real Estate Analysts Association* 2017;1:293-309.
 13. Choi CH, Park KH, Park HK, Lee MJ, Kim JS, Kim HS. Development of heavy rain damage prediction function for public facility using machine. *J Korean Soc Hazard Mitig* 2017;17:443-450.
 14. Lee HY, Noh SC. Chapter 8. Linear regression. In: Lee HY, Noh SC. *Advanced statistical analysis*. 2nd ed. Seoul:Moonwoosa;2013:250-339.
 15. Montenegro RD, Oliveira ALI, Cabral GG, Katz CRT, Rosenblatt A. A comparative study of machine learning techniques for caries prediction. *2008 20th IEEE International Conference on Tools with Artificial Intelligence* 2008;2:477-481.
 16. Tamaki Y, Nomura Y, Katsumura S, Okada A, Yamada H, Tsuge S, et al. Construction of a dental caries prediction model by data mining. *J Oral Sci* 2009;51:61-68.
 17. Gansky SA. Dental data mining: potential pitfalls and practical issues. *Adv Dent Res* 2003;17:109-114.
 18. Ito A, Hayashi M, Hamasaki T, Ebisu S. Risk assessment of dental caries by using Classification and Regression Trees. *J Dent* 2011;39:457-463.
 19. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Crit Care Med* 2016;44:368-374.
 20. Chung SY, Cho JW, Jung YS, Kim HY, Kim JY, Choi YH, et al. Association between unmet needs for dental treatment and the DMFT index among Korean adults. *J Korean Acad Oral Health* 2017;41:267-273.
 21. Shin HE, Kim HJ, Cho MJ, Choi YH, Song KB. Relationship between cancer and oral health in Korean adults determined using data from the 6th (2013-2014) Korea National Health and Nutrition Examination Survey. *J Korean Acad Oral Health* 2017;41:16-21.
 22. Nyanye. [Internet]. [cited 2019 Nov 26]. Available from: <https://nyanye.com/machine-learning/2017/01/18/Decision-tree/>.
 23. Ahn SH, You HY, Kim MJ, Han DH, Kim JB, Jeong SH. Caries preventive effect of permanent teeth using pit and fissure sealant program and community water fluoridation program. *J Korean Acad Oral Health* 2012;36:289-296.
 24. Oulis CJ, Berdouses ED, Mamai-Homata E, Polychronopoulou A. Prevalence of sealants in relation to dental caries on the permanent molars of 12 and 15-year-old Greek adolescents. A national pathfinder survey. *BMC Public Health*. 2011;11:100.
 25. van Loveren C, Lingstöm P. Chapter 8. Diet and dental caries. In: Fejerskov O, Nyvad B, Kidd, E. *Dental Caries: the Disease and Its Clinical Management*. 3rd ed. Oxford:Wiley Blackwell;2015:133-154.
 26. Hausen H, Baelum V. Chapter 23. How accurately can we assess the risk for developing caries lesions? In: Fejerskov O, Nyvad B, Kidd E. *Dental Caries: the Disease and Its Clinical Management*. 3rd ed. Oxford:Wiley Blackwell;2015:423-438.
 27. van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol* 2014;14:137.
 28. Breiman, L. Random Forests. *Machine Learning* 2001;45:5-32.