JKAOH
Journal of Korean Academy of Oral Health 2021 December 45(4):218-226
https://doi.org/10.11149/jkaoh.2021.45.4.218
Original Article

치주 영역에서 오렌지 데이터마이닝을 활용한 랜덤 포레스트 모델의 단계별 안내

임회정

전남대학교 치의학전문대학원 치의학교육학교실, 치의학연구소

A step-by-step guide to random forest model using orange data mining in the field of periodontitis

Hoi-Jeong Lim

Department of Dental Education, Dental Science Research Institute, School of Dentistry, Chonnam National University, Gwangju, Korea

Received: November 29, 2021
Revised: December 10, 2021
Accepted: December 14, 2021

Corresponding Author: Hoi-Jeong Lim
Department of Dental Education, Dental Science Research Institute, School of Dentistry, Chonnam National University, 77 Yongbong-ro, Buk-gu, Gwangju 61186, Korea
Tel: +82-62-530-5830
Fax: +82-62-530-5810
E-mail: hjlim@jnu.ac.kr
https://orcid.org/0000-0002-0795-8305

*This Research was supported by the National Research Foundation of Korea (NRF) grant, funded by the Korea government (No. 2018R1D1A1B07049719).

Objectives: The purpose of this study was to show a procedure for a random forest (RF) analysis which predicts periodontal disease status by using R and Orange Data Mining software, and helps us to understand how to apply the RF technique for dental research.
Methods: Oral examination data of the 7th Korea National Health and Nutrition Examination Survey were used. A RF model was adopted to analyze the data where the target variable was periodontal disease status and the features were gender, age, education level, marital status, alcohol consumption level, smoking status, brushing before sleep, hypertension, and diabetes-related variables.
Results: The important features of the RF analysis were in the order of age, marital status, and prevalence of hypertension and diabetes. The accuracy of the RF analysis was 73% which is not high enough for use in the clinical field.
Conclusions: The RF technique is an ensemble method used to predict periodontal disease status which produces higher accurate outputs than a single method. This study provides a step-by-step guide using Orange Data Mining for researchers who want to study machine learning techniques.

Key Words: Machine learning, Periodontitis, Random forest

서 론

4차 산업의 출현에 따라 머신러닝과 인공지능 기술의 급격한 발전이 이루어지고 있다. 인공지능이란 인간의 지적 능력을 인공적으로 구현한 컴퓨터 시스템이다. 이러한 인공지능은 머신러닝을 포함하고, 딥러닝은 머신러닝의 한 분야이다. 머신러닝이란 빅데이터를 통한 학습 방법으로 알고리즘을 이용해 데이터를 분석하고, 학습하며, 학습한 내용을 기반으로 판단을 내리거나 예측하는 것을 말하며, 딥러닝은 사람의 사고방식을 컴퓨터에게 가르치는 기계학습의 한 분야이다(Fig. 1). 머신러닝은 지도학습(supervised learning), 비지도학습(unsupervised learning), 강화학습(reinforcement learning)으로 나뉜다. 지도학습이란, 정답이 주어진 상태에서 데이터를 학습시키는 알고리즘을 말하며, 비지도학습은 정답이 주어지지 않은 상태에서 데이터의 특성을 학습하여 스스로 패턴을 파악하는 것을 말한다. 강화학습이란, 보상이 최대가 되게 효율적인 행동을 하도록 학습시키는 것을 말한다[1,2]. 이 중 랜덤 포레스트 모델은 지도학습에 속한다. 이러한 기법이 이제

는 컴퓨터 공학이나 통계학을 전공한 학자들만의 전유물이 아니라 모든 학문에서 적용되어야 하는 기법이므로 치의학 분야에서도 머신러닝 기법이나 딥러닝 기법을 배우고 연구 분야에 적용시켜야 할 필요가 있다.

통계 분석은 어떤 질병의 위험 요인들을 찾는 것에 중점을 두는 반면[3-7], 머신러닝은 위험 요인들을 기반으로 질병이 발생할 가능성에 대한 예측 정확도에 중점을 둔다. 예를 들어 머신러닝은 치주 질환이 있는 환자들의 생활 습관과 삶의 질 정보를 기반으로 알고리즘을 이용하여 질병 진단의 정확도를 높이는데 도움을 줄 수 있고, 미래의 환자 상태를 예측하여 치주 질환을 사전에 예방하게 할 수 있다. 이전 연구를 살펴보면, 스포츠 관련 치아 상해를 예측하기 위한 랜덤 포레스트(random forest) 모델이 89.3%로 로지스틱 회귀 분석(84.2%)보다 더 높은 정확도를 보였다[8].

머신러닝 기법 중 앙상블(ensemble) 기법은 여러 가지 모델을 조합하여 단일 학습 모델에 비해 예측력을 향상시킨다. 본 연구에서는 여러 앙상블 기법 중 배깅(bagging)에 대해 설명하고 배깅의 일종인 앙상블 모델 중 랜덤 포레스트 방법을 수행하기 위한 절차를 알아보고자 한다. 또한 치주 질환 연구에서 머신러닝 소프트웨어인 Orange datamining이 랜덤 포레스트 방법의 결과를 얻기 위해 어떻게 수행되었는지 살펴보고자 한다.

연구대상 및 방법

1. 앙상블 모델의 기본 개념

앙상블(ensemble)은 전체적인 어울림, 조화를 이루게 한다는 의미로, 앙상블 학습은 meta-learning 알고리즘이라고도 불리는데, 주어진 자료로부터 그 데이터가 어떤 클래스에 속하는지 여러 개의 예측 모형을 만들어 결합함으로써 보다 정확한 하나의 최종 예측을 도출하는 기계학습의 일종이다[9]. 앙상블 학습의 아이디어는 그 자체로는 그다지 좋은 성능을 발휘하지 못하는 약한 학습(weak learner) 모델들을 조합하여 더 나은 결과를 도출하는 강한 학습(strong learner) 모델을 만드는 것이다. 약한 학습기란 선형이고 단순한 분류기이며 연산이 빠른 반면, 강한 학습기란 비선형이며 복잡한 분류기를 말하지만 높은 정확도를 보인다.

기계학습 분류기에는 support vector machine, decision tree, neural network 등 여러 가지가 있다. 예를 들어, 서로 다른 20개 분류기 모두의 평가를 진행할 때, 이 중 과반인 10개 이상의 분류기가 잘못 분류할 경우, 이 분류기 세트 전체가 잘못 분류한 것으로 간주한다. 이때 각각의 개별 분류기의 오류율이 30%라고 가정을 하면 여러 개의 분류기 전체 오류율은 다음과 같다.

$$Error\ rate = \sum_{i=10}^{20} \binom{20}{i} 0.3^i (1-0.3)^{20-i} = 0.04796$$

분류기 각각의 개별 오류율은 30%였으나 위의 분류기 세트 전체의 오류율은 5%까지 감소했다. 이로써 개별 학습보다는 앙상블 학습으로서의 이점을 확인할 수 있다.

하나의 약한 학습기에서 생길 수 있는 과대적합과 과소적합을 잘 표현한 것이 편향-분산 트레이드오프이다. 편향(bias)이 크다는 것은 추정값들의 평균과 참값과의 거리가 멀다는 의미로 모델의 예측 결과가 실제 값에서 멀리 떨어진 상태를 뜻한다. 편향은 일반적으로 적절하지 못한 가정을 하거나 모델이 지나치게 단순할 때 발생한다. 분산(variance)은 추정값들의 평균과 추정값들 간의 거리를 의미한다.

고분산 학습 알고리즘은 훈련 데이터(training data)를 잘 표현하기는 하지만, 과대적합(overfitting)의 위험이 있다. 과대적합이란 학습 결과가 훈련 데이터에서는 매우 잘 수행되지만 새로운 데이터(test data)에서는 잘 수행되지 않는 현상이다. 반대로 고편향 학습 알고리즘은 과대적합 문제가 거의 없지만 훈련 데이터로부터 중요한 규칙성을 잘 파악하지 못하는 과소적합(underfitting) 문제가 발생한다. 낮은 편향과 낮은 분산은 모델에서 기대되는 가장 기본적인 특성이나 편향과 분산은 서로 상충 관계에 있어 동시에 감소시키기 어렵다(Fig. 2).[10] 랜덤 포레스트는 이러한 과대적합과 과소적합 문제를 극복하는 데 도움이 된다[11]. 효과적인 앙상블 방법은 많으나 여기서는 배깅과 부스팅 중 랜덤 포레스트 방법을 포함하는 배깅 방법을 간략히 소개하고자 한다.

1.1. 배깅(bagging)

배깅은 bootstrap aggregating의 약어로, bootstrapping sample을 여러 번 뽑아 약한 학습기들이 서로 독립적으로 병렬로 학습하여 여기서 얻은 결과들을 하나의 값으로 결합하는 방식이다. 배깅은 분산을

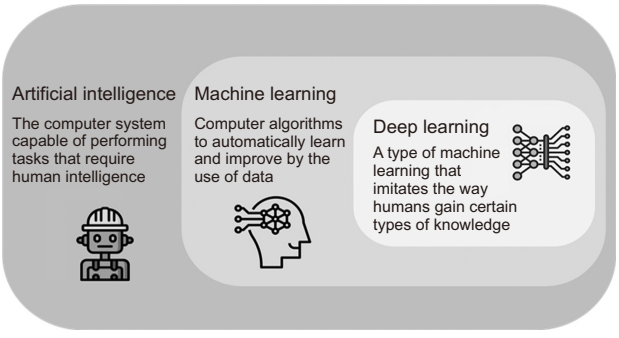

Fig. 1. Association of artificial intelligence, machine learning and deep learning.

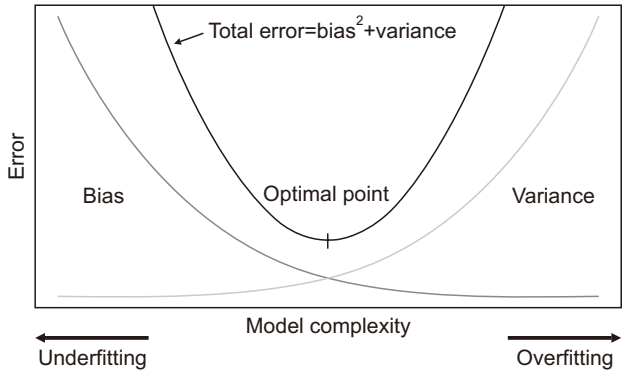

Fig. 2. Bias-variance trade-off.

감소시키는 역할을 하고, 부스팅은 편향을 감소시키는 역할을 한다[12].

배깅의 아이디어는 과대 적합을 피하기 위해 즉, 분산이 더 낮은 모델을 얻기 위해 여러 개의 독립적인 모델들을 적합시켜 예측값들을 평균하거나 과반수 득표를 얻는 방법을 이용해 하나의 결과값으로 산출하는 것을 말한다. 먼저 대표성과 독립성을 갖는 여러 개의 부트스트랩(bootstrap) 표본을 생성한다. 부트스트래핑이란 원래의 데이터 셋으로부터 관측치를 무작위 복원 추출을 반복하여 데이터 셋을 얻는 방법을 의미한다. 이후 각 표본에 대해 약한 학습기를 적합시켜 각각의 예측값을 종합한 후 voting 방식을 이용하여 상대적으로 분산이 낮은 앙상블 모델을 형성할 수 있다. 여기서 사용되는 voting 방법에는 두 가지가 있다. 분류 문제의 경우, 각 모델별로 투표로 산출된 class 중 과반수 득표를 얻은 class가 최종적인 결과값으로 산출되는 것을 Hard-voting이라고 한다(Fig. 3A). 반면, 각각의 모델에 대한 class들의 확률들을 평균한 후 더 높은 확률을 갖는 class가 최종적인 결과값으로 산출되는 투표 방식을 soft-voting이라고 한다(Fig. 3B).

1.2. 랜덤 포레스트(random forest)

랜덤 포레스트는 배깅의 일종으로, 여러 개의 약한 의사결정나무(decision tree)를 종합하여 하나의 강한 forest를 형성하는 기법이다. 앞에서 얘기했듯이 의사결정나무는 일반화 성능이 낮고 과대적합 문제가 발생할 수 있기 때문에 이를 해결하기 위한 방법으로 앙상블 모델 중 하나인 랜덤 포레스트를 사용한다. 랜덤 포레스트의 알고리즘을 설명해보면, 치주질환 여부를 예측하기 위해서 성별, 나이, 교육수준, 결혼여부, 흡연여부, 음주여부, 고혈압 유병여부, 당뇨병 유병여부 등 많은 요소(feature)가 필요하다. 이렇게 많은 요소를 가지고 치주질환 여부(label)를 예측한다면 과대적합이 일어날 수 있다. 예를 들어 feature가 20개라고 하고 이들 feature를 기반으로 하나의 커다란 결정 트리를 만든다면 트리의 가지가 많아져 과대적합의 결과를 야기할 수 있다. 하지만 20개 중 4개의 feature만 랜덤으로 선택(random feature subspace selection)하여 하나의 트리를 만들고, 또 20개 중 4개의 feature를 랜덤으로 선택해 또 다른 트리를 만든다. 이것을 반복하여 여러 개의 작은 트리를 만들 수 있고 다수결의 원칙에 따라 여러 결정 트리들이 내린 예측 값들 중 가장 많이 나온 값을 최종 예측값으로 정한다. 따라서 랜덤 포레스트 알고리즘은 배깅과 random feature subspace selection 개념을 결합하여 보다 견고한 모델을 만든다(Fig. 4)[13].

2. 연구대상 및 사용된 변수들

본 연구에서는 질병관리청에서 제공하는 국민건강영양조사 중 2016년에서 2018년까지의 구강검사 제7기 데이터를 사용하였다. 치주질환 유병여부를 결과변수로 사용하였고, 독립변수로는 성별, 나이, 교육수준, 결혼여부, 한 번에 마시는 음주량, 현재 흡연여부, 잠자기 전 칫솔질 여부, 고혈압 관련변수와 당뇨병 관련변수를 사용하였다. 본 연구에서 사용된 변수들은 Table 1과 Table 2에 정리되어 있다. Table 1에는 고혈압과 당뇨병 관련 변수를 제외한 독립변수들과 결과변수가 포함되어 있고, Table 2에는 고혈압, 당뇨병과 관련된 독립변수들에 대한 설명이 포함되어 있다.

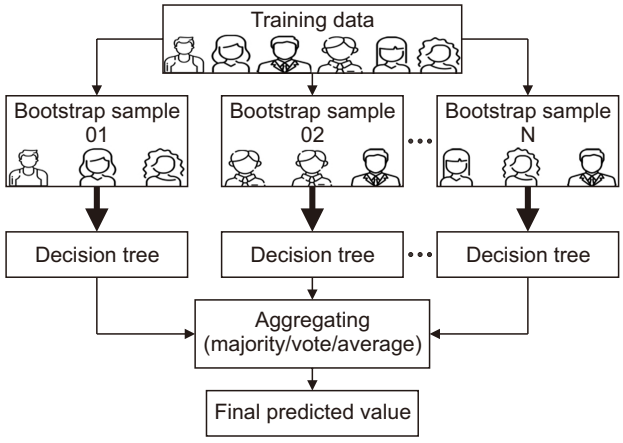

Fig. 4. Random forest.

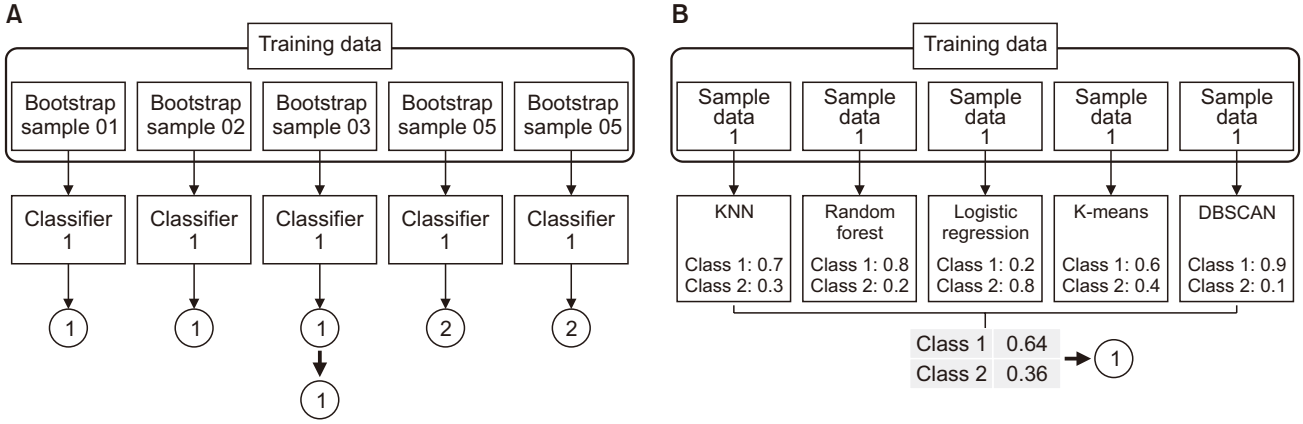

Fig. 3. Voting classifiers. (A) Hard voting. (B) Soft voting.

Table 1. Target variable and features excluding hypertension and diabetes-related variables used in this study

Variable	Description	Value
NO_CPI_34	Periodontal disease status	0. No 1. Yes
Sex	Gender	1. Male 2. Female
Age1	Age	1. 10~49 yr 2. ≥50 yr
Edu	Education	1. ≤Elementary school 2. Middle school 3. High school 4. ≥College
Marri_1	Marriage status	1. Married 2. Single 9. Unknown
BD2_1	(≥12 yr) amount of alcohol consumed at once	1. 1-2 cup 2. 3-4 cup 3. 5-6 cup 4. 7-9 cup 5. ≥10 cup 8. 0cup for recent 1 year 9. Unknown
BS3_1	Smoking status	1. Smoking everyday 2. Smoking sometimes 3. Ex-smoker 8. Non-smoker 9. Unknown
BM1_8	Brushing before sleep	0. No 1. Yes 8. No brushing 9. Unknown

3. 연구 방법

3.1. 변수 생성 과정

고혈압 유병 여부(HE_HP), 당뇨병 유병 여부(HE_DM) 데이터들의 결측값들을 살펴본 결과 고혈압, 당뇨병 관련 변수들을 가지고 고혈압과 당뇨병 유병 여부의 결측값들은 아래와 같이 추정할 수 있다.

고혈압 유병 여부(HE_HP1)라는 변수는 최종 수축기 혈압(HE_sbp)이 140 mmHg 이상 또는 최종 이완기 혈압(HE_dbp)이 90 mmHg 이상이거나 고혈압 의사 진단(HE_HPdg)을 받았거나 검진 당일 고혈압 약을 복용(HE_HPdr)한 경우 고혈압으로 정의하였고, 최종 수축기 혈압이 120 mmHg 이상 140 mmHg 미만 또는 최종 이완기 혈압이 80 mmHg 이상 140 mmHg 미만인 경우 고혈압 전단계로 정의하였으며, 최종 수축기 혈압이 120 mmHg 미만 이고 최종 이완기 혈압이 80 mmHg 미만인 경우 정상으로 정의하였다(Table 2).

당뇨병 유병 여부(HE_DM1)라는 변수는 8시간 이상 공복자 중 공복혈당(HE_glu)이 126 mg/dL 이상이거나 당화혈색소(HE_HbA1c)가 6.5% 이상이거나 당뇨병 의사 진단(HE_DMdg)을 받았거나 당뇨병 혈당 관리 차원에서 인슐린 주사(DE1_31)를 맞거나 당뇨병 약(DE1_32)을 먹는 경우 당뇨병으로 정의하였고, 공복혈당이 100 mg/dL 이상 126 mg/dL 미만 또는 당화혈색소가 5.7% 이상 6.5% 미만인 경

Table 2. Hypertension and diabetes related variables

Hypertension-related variables		
HE_HPdg	Diagnostic status of hypertension	0. No 1. Yes
HE_HPdr	Hypertension medication on the checkup day	0. No 1. Yes
HE_sbp	Final systolic blood pressure (average of 2nd and 3rd values)[1)]	□□□ mmHg
HE_dbp	Final diastolic blood pressure (average of 2nd and 3rd values)[1)]	□□□ mmHg
HE_HP	Hypertension status (≥19 yr)	1. Normal 2. Prehypertension 3. Hypertension
HE_HP1	Adjusted Hypertension status (≥19 yr)[1)]	1. Normal 2. Prehypertension 3. Hypertension
Diabetes-related variables		
HE_DMdg	Diagnostic status of diabetes	0. No 1. Yes
DE1_31	Blood glucose management for Diabetes_ Insulin injection	0. No 1. Yes 8. No blood glucose management 9. Unknown
DE1_32	Blood glucose management for Diabetes_ Diabetes medicine	0. No 1. Yes 8. No blood glucose management 9. Unknown
HE_glu	Fasting Glucose[2)]	□□□.□ mg/dL
HE_HbA1c	Glycated hemoglobin[2)]	□□.□ %
HE_DM	Diabetes status (≥19yr)	1. Normal 2. Impaired fasting glucose 3. Diabetes
HE_DM1	Adjusted diabetes status (≥19 yr)[2)]	1. Normal 2. Impaired fasting glucose 3. Diabetes

[1)]① Normal: Not ②,③, and a person whose systolic blood pressure (HE_sbp) is less than 120 mmHg and diastolic blood pressure (HE_dbp) is less than 80 mmHg. ② Prehypertension: Not ③, and a person with systolic blood pressure of 120 mmHg or more and less than 140 mmHg, and diastolic blood pressure of 80 mmHg or more and less than 90 mmHg. ③ Hypertension: A person with systolic blood pressure of 140 mmHg or more, diastolic blood pressure of 90 mmHg or more, or diagnosed with hypertension (HE_HPdg), or took hypertension medicine (HE_HPdr).

[2)](Among of persons with an empty stomach for more than 8 hours) ① Normal: Not ②,③ and a person whose fasting glucose (HE_glu) is less than 100 mg/dL and glycated hemoglobin (HE_HbA1c) is less than 5.7%. ② Impaired Fasting Glucose: Not, ③ and a person with fasting glucose of 100 mg/dL or more and less than 126 mg/dL, or glycated hemoglobin of 5.7% or more and less than 6.5%. ③ Diabetes: A person with fasting glucose of 126 mg/dL or more glycated hemoglobin of 6.5% or more, or diagnosed with Diabetes (HE_DMdg), or takes hypoglycemic agent (DE1_32), or receives insulin injection (DE1_31).

우 당뇨병 전단계로 정의하였으며, 공복혈당이 100 mg/dL 미만이고 당화혈색소가 5.7% 미만인 경우 정상으로 정의하였다(Table 2).

만 나이(age)는 10~49세까지와 50세 이상으로 나누어 age1이라는 변수를 생성하였다. 본 연구에 포함된 10-49세인 6,917명 중 치주질환에 걸린 사람은 969명(14.0%)이었고, 50세 이상인 6,289명 중에서 치주질환에 걸린 사람은 2,813명(44.7%)으로 카이제곱 검정 결과 유의하게 차이가 났기 때문에(P<0.01) 만 나이를 이 두 범주로 나누었다.

3.2. R studio와 orange data mining 적용

R studio 소프트웨어 설치 과정은 이전 논문[7]에 자세히 나와 있다. 또한 Orange Data Mining 소프트웨어는 머신러닝 및 데이터 시각화를 위한 소프트웨어이고 아래 링크에서 무료로 다운로드 받을 수 있다(http://orangedatamining.com).

3.3. R studio를 이용한 데이터 준비

```
> library(heaven)
> dental <- read_spss('c:/data/HNY7_OE.sav')
> write.csv(dental, 'C:/data/dental.csv')
```

먼저 작업할 데이터 파일을 작업 공간으로 불러와야 한다. 국민건강영양조사 원시자료인 'HNY7_OE.sav' 파일을 로컬 디스크 (C:) 안의 data 폴더에 저장한다. R에서 sav파일을 읽을 수 있게 해주는 라이브러리인 haven을 이용하여 sav파일을 읽은 후 dental에 저장하고 write.csv를 이용하여 C:의 data 폴더에 dental.csv로 저장한다. 이 파일은 국민건강영양조사 원시자료 (https://knhanes.kdca.go.kr/knhnanes/sub03/sub03_02_05.do) 중 2018년 데이터에서 'HNY7_OE.sav'를 다운로드 받을 수 있다. 여기서 sav 파일은 통계 소프트웨어인 SPSS 파일이다.

```
> dental <- read.csv('C:/data/dental.csv')
> dental <- subset(dental, select = c(NO_CPI_34, sex, age,
edu, marri_1, BD2_1,
+ BS3_1, BM1_8, HE_HPdg, HE_HPdr, HE_sbp,
+ HE_dbp, HE_HP, HE_DMdg, DE1_31, DE1_32,
+ HE_glu, HE_HbA1c, HE_DM))
> write.csv(dental, 'C:/data/dental2.csv')
```

read.csv를 이용하여 dental.csv를 읽어들인 후 subset 함수를 사용하여 필요한 변수들만으로 구성된 데이터셋을 만들어 dental2.csv로 저장한다. 이 데이터는 아래의 블로그에 저장되어 있다(https://blog.naver.com/nickjr66).

3.4. Orange data mining 적용

Workflow에 따라 데이터를 import하고 새로운 특성들을 생성하고 target을 지정하고 결측치를 제거하는 전처리를 거친 후, training data set을 가지고 랜덤 포레스트로 학습시키고 결과변수를 뺀 test data를 랜덤 포레스트 모델에 넣어서 예측값을 출력한 후 정확도를 측정하고 변수의 중요도를 출력한다(Fig. 5). 각각의 위젯에 해당하는 옵션은 아래 순서에 따른다.

CSV file import를 이용하여 dental2.csv를 reload한다(Fig. 6A). 총 데이터 수는 16,489개이고, 20개의 특성(feature)을 가진다(Fig. 6A). Feature constructor를 이용하여 아래와 같이 새로운 특성을 생성시킬 수 있다. 공복혈당(HE_glu), 당화혈색소(HE_HbA1c), 인슐

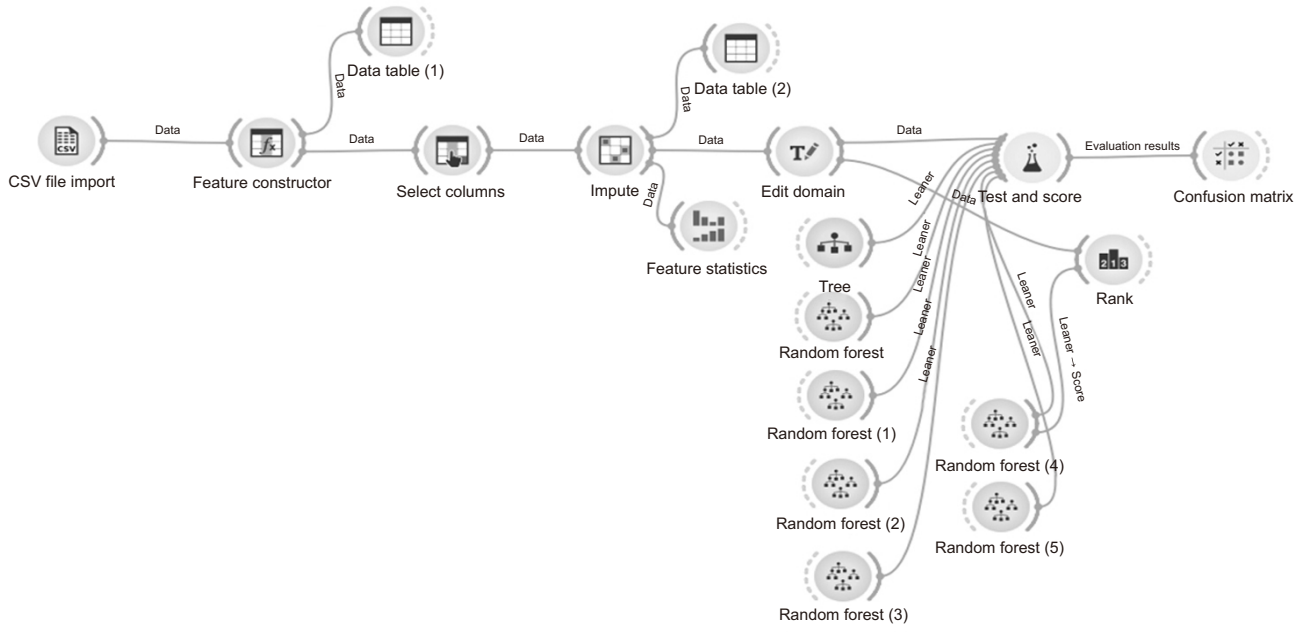

Fig. 5. Workflow.

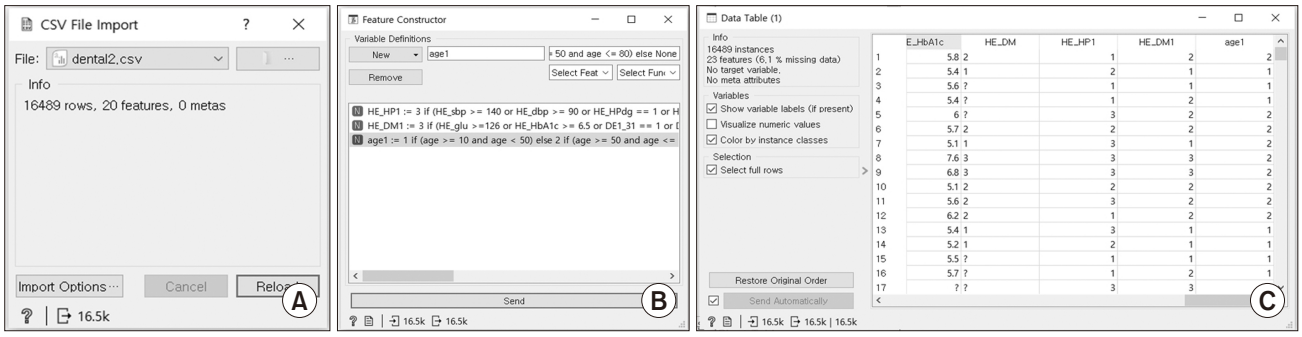

Fig. 6. Widget options in orange datamining. (A) CSV file import. (B) Feature constructor. (C) Data table.

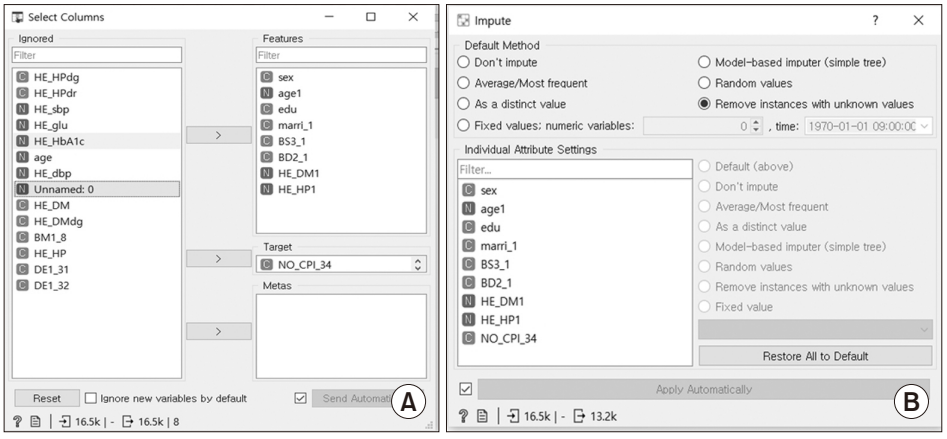

Fig. 7. (A) Select columns. (B) Impute.

린 주사(DE1_31), 당뇨병약(DE1_32), 당뇨병 의사진단 여부(HE_DMdg)를 이용해 당뇨병 유병 여부를 정의할 수 있도록 새로운 특성(HE_DM1)을 생성하였다. 공복혈당과 당화혈색소의 기준이 당뇨병 전단계(2)나 정상(1)의 조건에 만족하더라도 인슐린 주사 혹은 당뇨병 약을 먹거나 당뇨병 의사진단을 받은 경우에는 HE_DM1 == 3(당뇨병)에 포함시켰다. 마찬가지로, 최종 수축기 혈압(HE_shp), 최종 이완기 혈압(HE_dbp), 고혈압 의사진단 여부(HE_HPdg), 검진 당일 고혈압 약 복용 여부(HE_HPdr)를 이용해 고혈압 유병 여부를 정의할 수 있도록 새로운 특성(HE_HP1)을 생성하였고, 최종 수축기 혈압과 최종 이완기 혈압의 기준이 고혈압 전단계(2)나 정상(1)의 조건에 만족하더라도 검진 당일 고혈압 약을 복용하였거나 고혈압 의사진단을 받은 경우에는 HE_HP1 == 3 (고혈압)에 포함시켰다. 또한 만 나이(age)의 경우는 10세 이상 50세 미만과 50세 이상으로 두 개의 범주로 나누어 age1이라는 새로운 특성을 생성하였다(Fig. 6B). Variable definitions의 data table을 이용하여 새로운 특성들이 생성된 것을 확인할 수 있다(Fig. 6C).

(HE_DM1)

3 if (HE_glu>=126 or HE_HbA1c >= 6.5 or DE1_31 == 1 or DE1_32 == 1 or HE_DMdg== 1) else 2 if (HE_glu>= 100 and HE_glu< 126) or (HE_HbA1c >= 5.7 and HE_HbA1c <= 6.5)

else 1 if (HE_glu< 100 and HE_HbA1c< 5.7) else None

(HE_HP1)

3 if (HE_sbp>= 140 or HE_dbp>= 90 or HE_HPdg== 1 or HE_HPdr== 1)

else 2 if ((HE_sbp>= 120 and HE_sbp< 140) or(HE_dbp>= 80 and HE_dbp< 90))

else 1 if (HE_sbp< 120 and HE_dbp< 80) else None

(age1)

1 if (age >= 10 and age < 50) else 2 if (age >= 50 and age <= 80) else None

Select Columns를 이용하여 Target 변수를 치주질환 여부(NO_CPI_34)로 설정하였다(Fig. 7A). Impute 위젯을 이용하여 치주질환 여부(NO_CPI_34), 성별(sex), 나이(age1), 교육수준(edu), 결혼여부(marri_1), 현재 흡연 여부(BS3_1), 한번에 마시는 음주량(BD2_1), 보정된 당뇨병 유병여부(HE_DM1), 보정된 고혈압 유병여부(HE_HP1)에서 결측값을 포함한 모든 행들을 제거하였다(Fig. 7B).

Edit domain을 이용하여 새로운 특성의 생성을 위해 numeric으로 지정된 특성들을 본래의Categorical 특성으로 바꿀 수 있다(Fig. 8).

연구 성적

Rank는 중요한 변수의 확인이다. 정보이득(information gain)은 랜덤 포레스트 모델에서 특성의 중요도를 측정한 값으로, 전체 불순도(entropy)와 각 특성의 불순도의 차이(정보이득)가 클수록 변별력이 커지고 특성의 중요도가 커진다. 여기서는 나이(age1)>결혼여부(marri_1)>고혈압 유병여부(HE_HP1)>당뇨병 유병여부(HE_DM1)>현재 흡연 여부(BS3_1)>교육수준(edu)>성별(sex)>한번에 마시는 음주량(BD2_1) 순으로 중요한 특성이다. 특히 나이(age1)가 유의한 분류를 할 수 있는 특성임을 확인할 수 있었다(Fig. 9).

Random forest에서 포레스트에 포함할 결정 트리수인 number of trees:(ntree)는 100으로 지정해 주었고 분할에 적용되는 특성의 수인 number of attributes considered at each split:(mtry)는 4로 지정해 주었다(Fig. 10A). Tree의 옵션을 살펴보면, Induce binary tree는 feature의 범주가 3개 이상일 때, 3개 이상의 범주를 2개의 범주로 묶어주는 역할을 한다. 이 경우 3개 이상의 범주를 그대로 사용한다면 정확도가 커지기 때문에 이 옵션을 선택하지 않았다. Min. number of instances in leaves는 leaf의 instance가 지정된 숫자(Min. number) 이하이면 분할을 멈추게 한다. 지정된 숫자가 2라는 것은 거의 끝까지 분할하겠다는 의미이다. Do not split subsets smaller than은 지정된 숫자보다 작으면 분할을 멈추게 하는 것이고, limit the maximal tree depth to는 최대 분할 횟수이다. Stop when majority reaches [%]는 전체 total 값의 지정 기준값(%)에 이를 때까지 분류를 계속한다는 것이다(Fig. 10B).

Sampling에서 random sampling을 선택했는데 전 과정을 10번 반복하고 전체 데이터의 80%는 training data로 사용되고, 20%는 test data로 사용하여 decision tree와 랜덤 포레스트의 결과를 얻었다. 그 결과들은 AUC나 CA, F1 등으로 표현되는데, AUC의 x축은 1-Specificity (false positive rate)이고 y축은 sensitivity (true positive rate)이므로 1에 가까울수록 false positive보다 true positive가 많아진다는 의미이고 즉, 양성을 맞추는 비율이 높아진다는 것을 의미한다. Random forest의 결과인 AUC나 F1은 Tree의 결과보다 높았으

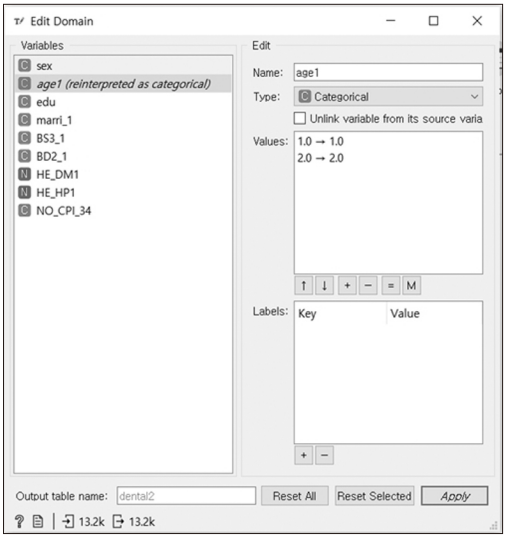

Fig. 8. Edit domain.

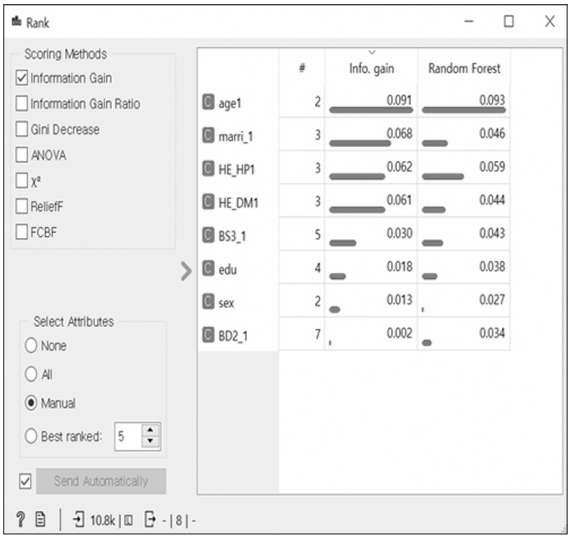

Fig. 9. Rank.

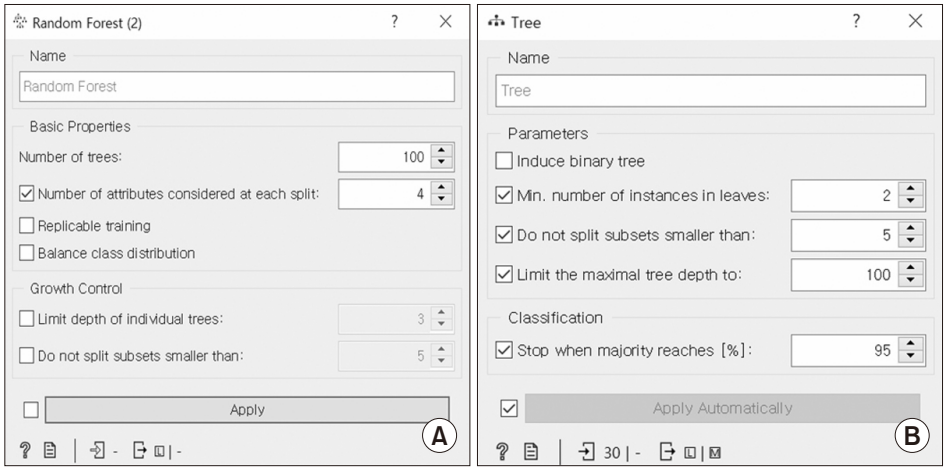

Fig. 10. Widget options in Orange Data-amining. (A) Random forest. (B) Tree.

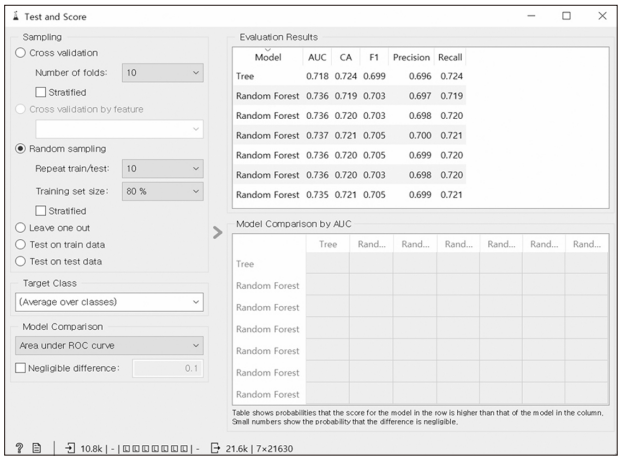

Fig. 11. Test and score.

나 CA는 다소 낮았다. 그 이유는 모델이 너무 간단하기 때문이다. 모델이 복잡해질수록 random forest의 성능이 tree의 성능보다 더 좋아진다(Fig. 11, Table 3). 모델의 정확도 73%는 실제 임상 분야에서 사용하기에는 높은 값은 아니다. 위의 정확도를 올리기 위한 다양한 option 설정을 통한 방법들이 있으나, 본 논문의 목적은 머신러닝 입문자만을 대상으로 절차를 소개해주기 위한 것이기에 더 깊이 다루지 않았다.

최적의 하이퍼 파라미터를 찾기 위한 그리드 기법을 이용하여 OOB error rate의 최소를 구할 수 있다. Grid search란 랜덤 포레스트 모델을 돌려 error rate과 accuracy를 계산하여 가장 낮은 error rate을 가진 ntree와 mtry의 조합을 발견하는 것이고, OOB error란 bootstrap 추출로 train data에 속하지 않는 데이터를 decision tree를 이용하여 예측한 값과 실제 값의 차이를 말한다. Orange data mining에는 범위를 정하면 자동으로 최적의 하이퍼 파라미터를 찾아주는 기능이 없기 때문에 여러 개의 랜덤 포레스트 위젯으로 하이퍼 파라미터를 다르게 설정하여 적절한 하이퍼 파라미터를 찾는 작업을 하였다. 하이퍼 파라미터인 ntree는 50과 100을, mtry는 3에서 5까지 수동으로 설정하여 찾은 결과, ntree는 100, mtry는 3인 경우가 최적의 하이퍼 파라미터인 것을 알았다. 그러나 가장 좋은 하이퍼 파라미터가 아닐 수 있고, OOB error rate이 더 낮아질 수 있다(Table 4).

고 안

이 연구에서 사용된 오렌지 데이터마이닝(이하 오렌지) 무료 소프트웨어는 슬로베이나 류블라나 대학교에서 개발한 소프트웨어로 기계학습과 데이터 시각화의 강력한 기능을 가지고 있다. 또한 R이나 Python과 비교하여 코딩이 부담스러운 연구자들이 아주 쉽게 적응할 수 있는 대화형식 프로그램이다. 파이썬의 장점은 광범위한 라이브러리를 포함하고 있고 확장성도 좋은 반면 오렌지에 비해 비전공자들이 파이썬을 익혀 적절히 사용하는데 시간이 걸린다는 단점이 있다. R과 파이썬 모두 명령문을 직접 입력해야 하는 소프트웨어기 때문에 초보자들에겐 다소 어려운 점이 있을 수 있다. 본 연구에서는 치의학을 연구하는 코딩 비전공자들이 손쉽게 다가갈 수 있는 오렌지를 소개함으로써 머신러닝 분석을 좀 더 용이하게 스스로 수행할 수 있도록 돕는데 목적이 있다.

Table 3. Confusion matrix

		Predicted class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

Recall=TP/(TP+FN).
Precision=TP/(TP+FP).
F1=Harmonic mean of Recall and Precision.
CA(Accuracy)=(TP+TN)/(TP+FN+FP+TN).
AUC (Area Under the ROC Curve)=X-axis: 1-Specificity, Y-axis: Sensitivity.

Table 4. Grid search for the optimal hyperparameters

Model	ntree	mtry	AUC	CA	F	Precision	Recall
Tree			0.718	0.724	0.699	0.696	0.724
Random forest	100	3	0.736	0.729	0.703	0.697	0.719
Random forest	50	3	0.736	0.720	0.703	0.698	0.720
Random forest	100	4	0.737	0.721	0.705	0.700	0.721
Random forest	50	4	0.736	0.720	0.705	0.699	0.720
Random forest	100	5	0.736	0.720	0.703	0.698	0.720
Random forest	50	5	0.735	0.721	0.705	0.699	0.721

이 연구에서 decision tree를 사용하지 않고 앙상블 방법 중 하나인 랜덤 포레스트를 사용하여 모델을 학습시켜 accuracy를 계산하였는데, 각각의 decision tree는 과대적합 되기 쉬운 경향이 있다. 그 이유는 decision tree에서 하나의 error가 발생하면 다른 노드에도 error가 전달되기 때문이다. 이러한 문제를 해결하기 위해 과대적합된 tree를 많이 만들고 앙상블 방법을 사용하여 그 결과를 평균을 낸다면 과대적합된 양을 줄일 수 있다. 이렇게 하면 decision tree 모델의 예측 성능은 유지되면서 과대적합이 줄어들 수 있다는 것을 알 수 있을 것이다.

위에서 사용된 부트스트랩 샘플링은 데이터의 분포가 불균형할 때 균형 잡힌 분포로 바꿀 수 있는 유용한 방법이다. 예를 들어, 치주질환 여부를 분류기로 training한다고 하면 training set에 치주질환이 없는 경우가 있는 경우보다 훨씬 많을 수 있다. 이런 경우 치주질환이 없는 경우만 분류했을 때도 99%의 정확도를 보일 수 있다. 이렇게 데이터의 분포가 불균형일 때 치주질환이 있는 class의 error는 무시되는 방향으로 training되기 쉽다. 이를 해결하기 위해서는 부트스트래핑을 통해 치주 질환이 있는 데이터 수를 늘리거나 반대로 치주질환이 없는 데이터의 수를 줄이는 방법도 있다[13].

오렌지는 소규모 데이터 프로젝트나 탐색적 데이터 분석, 교육을 목적으로 쓰일 때 더 적합하다. 반면, 오렌지의 단점은 대규모 데이터로 작업할 경우 안정적이지는 않다는 것이다. 대규모 데이터가 파이썬에서는 잘 작동하지만, 오렌지에서는 제한적으로 작동할 수 있다. 또 다른 단점으로는 파이썬의 효과적인 배열(array) 처리와 같은 편의성을 갖기 어렵다는 점이다. 예를 들어 grid search를 할 경우 일일이 각각의 하이퍼 파라미터에 대해 결과를 얻어야 하는 반면, 파이썬에서는

코딩 몇 줄로 한 번에 가능하다. 결과변수가 연속변수인 경우에서의 오렌지를 이용한 절차 및 정확도를 높이는 방법 등은 추후 연구에서 고려할 필요가 있다.

결 론

앙상블 학습은 많은 응용 분야에서 명백한 이점을 보여 주는 강력한 머신러닝 패러다임이다. 다수의 약한 학습기를 사용함으로써 앙상블의 일반화 성능은 단일 학습기의 일반화 성능보다 향상된다. 효과적인 앙상블 방법 중 하나인 배깅의 일환으로 랜덤 포레스트가 있고 이 방법은 의사결정나무의 과대 적합 문제를 해결하기 위해 사용되며, 약한 학습기인 의사결정나무들을 사용함으로서 하나의 강한 학습기를 형성한다. 이러한 머신러닝 기법을 연구하려는 연구자들에게 orange data mining을 이용한 단계별 방법을 알려, 스스로 분석을 수행하여 연구 결과를 얻고 해석하는데 도움이 되고자 한다.

References

1. Mitchell T. Machine Learning. New York: McGraw Hill; 1997.
2. Geron A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow Second edition: Concepts, Tools, and Techniques to Build Intelligent Systems, California: O'Reilly Media Sebastopol; 2019.
3. Park SH, Lim HJ. A step-by-step guide to Meta-analysis with dichotomous outcomes using RevMan in dental research. J Korean Dent Assoc 2018;56:18-40.
4. Lim HJ, Park SH. A step-by-step guide to Generalized Estimating Equations using SPSS in dental research. J Korean Dent Assoc 2016;54:850-864.
5. Lim HJ. Sample size determination in dental research. The Journal of the Korean dental association 2014;52:558-569.
6. Lim HJ. Meta-analysis in dental research. J Korean Dent Assoc 2014;52:478-490.
7. An H, Lim HJ. A step-by-step guide to Propensity Score Matching method using R program in dental research. J Korean Dent Assoc 2020;58:152-168.
8. Farhadian M, Torkaman S, Mojarad F. Random forest algorithm to identify factors associated with sports-related dental injuries in 6 to 13-year-old athlete children in Hamadan, Iran-2018-a cross-sectional study. BMC Sports Sci Med Rehabil. 2020;12:69.
9. Rocca J. Ensemble methods: bagging, boosting and stacking [Internet] [cited 2021 Oct 26]. https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205.
10. Neal B, Mittal S, Baratin A, Tantia V, Scicluna M, Lacoste-Julien S, Mitliagkas L. A Modern Take on the Bias-Variance Tradeoff in Neural Networks. 2019; arXiv:1810.08591.
11. Guido S, Mueller AC. Introduction to Machine Learning with Python. California; O'Reilly Media: 2016.
12. Zhou ZH. Ensemble learning. In: Li, SZ (eds) Encyclopedia of biometrics, Berlin; Springer: 2009.
13. Yang JH. Welcome. Are you new to machine learning? Seoul, Knowing More Publishing: 2016.