https://doi.org/10.5392/JKCA.2022.22.03.354

머신러닝을 활용한 식품소비에 따른 대사성 질환 분류 모델

Metabolic Diseases Classification Models according to Food Consumption using Machine Learning

홍준호*, 이경희, 이혜림***, 정환석***, 조완섭****

㈜현대자동차 데이터인텔리전스팀*, 충북대학교 경영정보학과**, 농촌진흥청 디지털농업추진단***

Jun Ho Hong(jjang7ok2002@gmail.com)*, Kyung Hee Lee(lee.kyunghee@gmail.com), Hye Rim Lee(leehr26@korea.kr)***, Hwan Suk Cheong(xpertstone@korea.kr)***, Wan-Sup Cho(wscho@cbnu.ac.kr)****

요약

대사성 질환은 국내의 경우 유병률이 26%에 이르는 질환으로 복부비만, 고혈압, 공복혈당장애, 고중성지방, 낮은 HDL 콜레스테롤 5가지 상태 중 3가지를 동시에 가진 상태를 말한다. 본 논문은 농촌진흥청의 소비자패널 데이터와 건강보험공단의 진료 데이터를 연계하여 식품 소비 특성을 통해 대사성 질환자군과 대조군으로 나누는 분류 모델을 생성하고 차이를 비교하고자 한다. 기존의 국내외에서 연구된 많은 대사성 질환과 식품 소비 특성 관련 연구는 특정 식품군이나 특정 성분의 질환 상관성 연구이며, 본 논문은 일반 식사에서 포함하는 모든 식품군을 고려한 로지스틱 회귀를 이용한 분류 모델, 의사결정나무 기반 분류 모델, XGBoost를 활용한 분류 모델을 생성하였다. 세 가지 모델 중 정확도가 높은 모델은 XGBoost 분류 모델이지만, 정확도가 0.7 미만으로 높지 않았다. 향후 연구로 환자군의 식품 소비 관찰 기간을 5년 이상으로 확대하고 섭취한 식품을 영양적 특성으로 변환한 후 대사성 질환 분류 모델 연구가 필요하다.

■ 중심어 : | 식품 소비 | 대사성 질환 | 분류 모델 | 머신러닝 | 데이터 연계 |

Abstract

Metabolic disease is a disease with a prevalence of 26% in Korean, and has three of the five states of abdominal obesity, hypertension, hunger glycemic disorder, high neutral fat, and low HDL cholesterol at the same time. This paper links the consumer panel data of the Rural Development Agency(RDA) and the medical care data of the National Health Insurance Service(NHIS) to generate a classification model that can be divided into a metabolic disease group and a control group through food consumption characteristics, and attempts to compare the differences. Many existing domestic and foreign studies related to metabolic diseases and food consumption characteristics are disease correlation studies of specific food groups and specific ingredients, and this paper is logistic considering all food groups included in the general diet. We created a classification model using regression, a decision tree-based classification model, and a classification model using XGBoost. Of the three models, the high-precision model is the XGBoost classification model, but the accuracy was not high at less than 0.7. As a future study, it is necessary to extend the observation period for food consumption in the patient group to more than 5 years and to study the metabolic disease classification model after converting the food consumed into nutritional characteristics.

■ keyword : | Food Consumption | Metabolic Disease | Classification Model | Machine Learning | Data Linkage |

접수일자 : 2022년 02월 10일
수정일자 : 2022년 02월 25일
심사완료일 : 2022년 02월 25일
교신저자 : 조완섭, e-mail : wscho@cbnu.ac.kr

I. 서론

대사성 질환은 비만이나 운동 부족, 과잉영양 등의 생활습관이 원인이 되어 나타나는 병을 통칭하는 것으로 주로 당뇨병, 고혈압, 고지혈증, 심장병 등이 여기에 속한다. 대사성 질환의 위험인자 중 식품 생활습관이 가장 중요한 위험인자로써 식생활패턴이 대사성 질환에 영향을 주는 것으로 제시되고 있다[1].

식품은 탄수화물, 단백질, 지질 같은 영양소의 공급원으로서도 중요하지만, 이들 식품에 함유된 각종 생리 활성 성분은 다양한 생체기능 조절작용을 통해서 질병 예방에 이바지할 수 있다[2]. 최근 들어 식품 성분이 단순히 영양소를 공급하는 일차적인 차원을 넘어 질병을 예방하고 치료효율을 높이는 생체조절기능을 한다는 것이 많은 연구에서 증명되고 있다[3]. 대사성 질환은 심혈관 질환 및 당뇨의 발병과 관련된 다양한 위험인자의 군집으로 정의되며[4], 고령 인구가 증가하면서 대사증후군 관리가 중요한 건강관리 방식으로 활용되고 있다.

본 논문은 농촌진흥청의 소비자패널 데이터와 건강보험공단의 진료 데이터를 연계하여 식품 소비 특성을 통해 대사성 질환 환자군과 대조군으로 나누는 분류 모델을 생성하고 차이를 비교하고자 한다. 기존의 국내외에서 연구된 많은 대사성 질환과 식품 소비 특성 관련 연구는 특정 식품군이나 특정 성분의 질환 상관성 연구가 대부분이다. 본 논문은 개인이 섭취한 다양한 식품군을 모두 고려하여 대사성 질환과 상관성 있는 식품 소비 특성을 찾기 위하여 최대한 많은 변수를 활용하는 통계 분석 및 머신러닝(Machine Learning) 분석기법을 적용하고자 한다.

II. 관련 연구

1. 식품 섭취와 질환의 상관성 연구

질병의 발생과 영양 요인의 관련성에 대한 연구 분야인 영양 역학은 전통적으로 식품 단위 또는 영양소 단위로 질병에 대한 위험요인을 연구했다[5]. 식품 소비와 질환과의 연관성 연구결과를 살펴보면 서구식 식사습관이 있는 사람들이 대장 직장암 발병확률이 높다는 연구가 있었고, 일본에서는 남성을 대상으로 한 연구에서 유제품, 과일, 채소를 많이 먹고 알코올을 적게 섭취하는 식사습관이 결장암의 위험을 감소시키는 연구결과를 보고하였다[6]. 과일, 채소, 콩, 생선, 닭고기, 전곡류를 많이 섭취하는 건강을 고려한(Prudent) 식사습관은 심혈관질환의 위험을 감소시켰으며, 붉은 고기, 당이 많은 음식, 디저트, 튀긴 감자, 정제된 곡류를 많이 섭취하는 서구식(Western) 식사습관은 위험도를 증가시킨다는 결과가 보고되었다[7]. 그 외에도 식사 패턴에 따라 2형 당뇨, 고혈압, 대사증후군 등의 위험도가 증가하거나 감소했다는 연구결과가 보고되었다[8].

2. 머신러닝을 활용한 의학 연구

머신러닝 활용 의학연구는 딥러닝 기법의 발전과 함께 주로 영상의학 분야에서 활발하게 이루어졌고, 당뇨병 및 내분비질환 임상연구에서도 최근 적극적으로 머신러닝을 활용하려는 연구자들이 늘어나고 있다[9]. 병원 자료 기반으로 5년 내 당뇨 발생률을 예측하는 모델을 제시한 연구에서는, 전자의무기록 기반 28개 변수를 추출하여 머신러닝 모델을 만들고 당뇨병 발생 여부를 예측하였다. 다만 예측력에 있어서 기존 예측모델을 상회하는 성능을 보여주지는 못하여, 기존 병원 자료에서 얻을 수 있는 예측력의 한계를 넘기 위해서는 추후 환자의 생활방식이나 식습관, 운동 등 병원 밖에서 수집될 수 있는 특성변수의 활용이 중요할 것으로 보인다[10]. 최근 활용이 증가하고 있는 연속혈당측정기(continuous glucose monitoring, CGM)에서 얻을 수 있는 혈당 변화 시계열 데이터는 머신러닝을 활용하기에 적절한 자료이다. 한 국내 연구에서, 연구진은 30분 이내 저혈당이 발생할 것을 예측하는 모델을 구축하였다. 랜덤포레스트(Random forest) 모델은 roc_auc_score (area under the receiver-operating characteristics curve) 0.966, 민감도 89.6%, 특이도 91.3%로 좋은 예측 능력을 보여 추후 CGM 및 인공췌장 개발, 고도화에 도움을줄 가능성이 제시하였다[11].

3. 머신러닝 기반 분류분석 모델

로지스틱 회귀분석(Logistic Regression Analysis)은 질병 여부를 직접 예측하는 것이 아니라, 독립변수

에 따라 종속변수가 일어날 확률을 추정하는 분석방법이다. 종속변수의 예측값은 항상 0과 1 사이의 확률값을 갖게 되며 값이 0.5보다 크면 질병에 걸렸을 확률이 높고, 0.5보다 작으면 대사성 질환에 걸렸을 확률이 낮다고 예측할 수 있다.

의사결정나무(Decision Tree)는 의사결정 규칙을 나무구조로 표현하여 관심 대상이 되는 집단을 소집단으로 분류하거나 특정 값을 예측하는 데에도 사용되는 분석방법이다. 의사결정나무는 직관적으로 결과를 도식화하여 볼 수 있기 때문에 분리 규칙을 알 수 있다는 장점이 있다. 그러나 분류되는 단계가 많을수록 이해하기가 어려우며, 데이터에 따라 결과가 안정적이지 못할 수 있다. 그러므로 최종 알고리즘을 도출하기 전 주요한 변수를 탐색하는 차원에서는 활용하기 적합하다.

또한, 의사결정나무는 모델의 특성상 최하단의 노드까지 완전탐색으로 학습하므로 과적합의 문제가 생긴다. 이러한 데이터의 과적합 문제는 트리의 깊이를 줄이는 가지치기로 어느 정도 방지할 수 있다[12].

XGBoost는 앙상블 기법의 하나로, 앙상블(Ensemble)은 정확도 향상을 위해 여러 모형을 연결하여 더 강력한 모형을 만드는 기법이다[13]. 앙상블 모형 M_k은 일련의 k개의 기반 분류 모형(Base Classifier)의 조합으로 구성되며, 각각의 분류기는 데이터 세트 D로부터 만들어진 k개의 훈련 세트 $D_1, D_2, \cdots, D_k$로부터 분류기 $M_1, M_2, \cdots, M_k$을 만든다[13].

Ⅲ. 연구대상 및 방법

1. 연구대상 및 변수 정의

본 연구는 보건복지부 지정 공용기관생명윤리심의위원회의 승인(IRB승인번호:P01-202012-21-010)을 받았으며, 연구에 참여 동의한 463명의 연구대상자의 농촌진흥청 농식품 소비자 패널데이터와 국민건강보험공단의 명세서데이터와 건강검진 데이터를 연계하여 구축한 맞춤형 데이터(건강보험공단 연구관리번호: REQ0000 46603-001)를 데이터셋으로 활용하였다. 구축한 데이터는 농진청 소비자 패널 데이터는 2015년~2020년 가구별 구매한 농식품 내역 데이터이며, 건강보험공단의 데이터는 2002년~2020년까지 건강검진 내역과 건강보험청구 데이터이다. 사용된 변수는 농식품 소비자 패널의 식품구매 내역을 농식품군별로 월단위의 구매금액을 집계하였고 건강보험공단의 건강검진 내역(건강검진설문, 신체계측, 혈액 및 요검사 수치) 및 건강보험 명세서의 진료 일자, 질환 코드, 처방일수이다.

2. 대사성 질환과 식품 소비 관계성 연구를 위한 데이터 전처리

대사성 질환자는 건강보험공단 명세서데이터에 다음 [표 1]의 상병코드가 있는 사람으로 하였으며, 본 연구에서 다루는 대사성 질환 상병코드는 [표 1]과 같다.

표 1. 대사성 질환 상병코드

질환명	진단 코드(ICD-10)
당뇨	E11, E13, E14
비만	E66
이상지질혈증	E78
고혈압성 질환	I10-I15
허혈압성 질환	I20-I25

농식품 소비와 건강과의 관계성 분석을 위하여 신규 대사성 질환자는 2002년부터 2019년까지 건강보험공단 명세서 데이터(T20)의 상병코드 변수에서 [표 1]에 명시된 코드가 처음 청구된 시점을 '최초 진단일'로 추정하여, 최초 진단일이 2015년~2019년 사이에 존재하는 사람으로 정의하였다. 대조군은 2002년부터 2019년까지 [표 1]의 코드가 명세서에 없는 사람이면서 환자군과 가족이 아닌 사람으로 하였다. 이와 같은 방법으로 반응변수인 '대사성 질환 유무'는 [표 1]에서 정의한 대사성 질환 상병코드가 건강보험 명세서에 1건이라도 있으면 대사성 질환 유무 필드에 '1'을 표시하고, 그렇지 않으면 '0'으로 정의하였다.

전처리한 데이터는 [그림 1]과 같이 구매한 식품의 14개 식품군 변수, 식품군 변수의 값은 식품의 월평균 소비량이며, 환자군 대조군을 나누는 클래스 변수(y)로 구성된다.

	fruits	cereals	other_proc_food	noodle	bread	sea_meat_proc	seafood	vegetables	milk_product	drink	seasoned	alcohol	meat	snack	y
0	10.416667	1.041667	4.062500	0.562500	1.583333	1.479167	1.333333	2.687500	3.000000	0.562500	0.750000	0.104167	1.708333	1.958333	1
1	5.500000	0.500000	8.104167	0.458333	0.833333	4.250000	0.708333	1.979167	4.895833	0.458333	0.375000	0.041667	2.250000	4.125000	1
2	0.916667	0.194444	2.722222	1.000000	2.833333	4.333333	2.055556	2.305556	7.472222	1.666667	0.444444	0.972222	0.750000	5.305556	1
3	0.083333	0.194444	2.083333	0.694444	1.222222	2.388889	5.027778	3.722222	2.527778	1.583333	1.638889	0.138889	2.555556	1.888889	1
4	6.055556	0.750000	2.166667	0.166667	1.694444	2.027778	0.388889	2.666667	3.305556	0.916667	2.666667	1.222222	1.166667	1.500000	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
323	4.875000	0.450000	3.588095	0.864286	2.455952	2.142857	0.761905	3.170238	7.017857	0.716667	1.546429	0.079762	1.832143	4.063095	0
324	3.211268	0.476056	1.425352	1.301408	1.267606	1.456338	2.090141	2.516197	2.814085	0.515493	1.036620	0.219718	2.225352	1.504225	0
325	1.854167	0.166667	3.895833	1.708333	1.395833	2.500000	0.812500	1.458333	4.395833	4.750000	0.729167	0.854167	0.958333	6.104167	0
326	7.583333	0.277778	8.305556	0.611111	2.111111	3.111111	0.861111	1.777778	2.527778	2.388889	1.166667	0.833333	27.416667	2.805556	0
327	8.338983	0.329096	4.162429	1.841808	2.799435	5.381356	3.372881	6.074859	8.766949	3.016362	1.285311	2.025424	3.343220	5.587571	0

그림 1. 전처리된 데이터의 형태

14개 식품군은 소비자패널데이터의 기존 대분류를 연구 목적에 맞게 재구성한 것으로 자세한 내용은 참고문헌[14]에 기술하였다. 이때 패널(가구) 단위 농식품 소비내역 조사 데이터를 이용하여 가구구성원의 농식품 소비데이터를 산출하였다, 가족 구성원의 식품 선호는 농촌진흥청의 소비자 패널데이터에 포함되지 않아서 이를 세세하게 반영한 개인별 농식품 소비데이터를 구분하는 데 한계가 있었다.

3. 식품 소비에 따른 대사성 질환 분류분석

본 논문에서는 종속변수의 유무에 영향을 미치는 변수를 찾기 위해 환자군 52명과 대조군 276명을 가지고 식품의 소비패턴에 따라 환자군과 대조군을 분류하는 분류분석을 진행하였다. 분석에 사용한 툴은 파이썬의 데이터분석용 라이브러리를 활용하였으며, 사이킷런(scikit-learn) 패키지의 함수를 이용했다.

3.1 SMOTE를 활용한 클래스 균형화

머신러닝 알고리즘으로 예측을 수행하는 경우에 종속변수인 클래스의 균등한 분포가 매우 중요하며, 클래스 분포의 불균형이 심한 경우, 분류기는 다수 클래스에 압도되어 소수 클래스를 올바르게 분류하지 못한다.

본 연구에서는 환자군-대조군 분류할 때 흔히 나타나는 클래스 불균형이 심한 데이터에 오버샘플링(SMOTE[15])을 적용하여 클래스의 균형을 맞추었다. [표 2]는 오버샘플링 후 데이터 사이즈를 보여준다.

표 2. 오버샘플링 후 데이터 사이즈

구분	원본 size	오버샘플링 후
X_train	229	386
y_train	229	386
환자군 수	36	193
대조군 수	193	193

3.2 로지스틱 회귀를 이용한 분류 모델

본 연구에서는 로지스틱 회귀분석을 하기 위해 StandardScaler를 통해 독립변수를 표준화시켜 스케일(Scale)을 맞추어 주었다. 그 후 변수 선택법을 통해 유의하지 않은 변수들을 제거하고 분류분석을 진행하였으며, 그 결과 [그림 2]와 같다.

로지스틱 회귀분석의 roc_auc_score 값은 0.61로 분류 정확도가 높지 않았다. 이는 데이터가 학습하기에 충분하지 않을뿐더러 로지스틱 회귀분석은 독립변수의 다중공선성이 있으면 변수 자체를 제거하기에 단순한 모형이 아닌 이상 모델의 정확도가 떨어지기 때문이다.

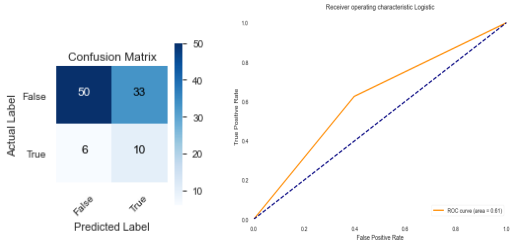

그림 2. 로지스틱 회귀분석 결과

이를 보완하기 위해 의사결정 나무와 앙상블 중 정확도가 높다고 알려진 XGBoost를 사용하여 분류 모델을 생성하였다.

3.3 의사결정나무를 이용한 분류 모델

의사결정 나무는 매개변수를 조정하여 주지 않으면 train_data를 모두 학습하므로, 과대 적합이 일어난다. 과대 적합을 방지하기 위해 가지치기를 해주어야 한다. 최적의 모델을 찾기 위해 그리드 서치(완전탐색)을 통해 해당 모델의 매개변수인 max_depth ,min_samples_leaf의 최적값을 찾았으며, 모델을 의사 결 정 나 무 분 류 기 (m a x _ d e p t h = 3 , min_samples_leaf =4)로 학습하였다. 그 결과는 [그림 3][그림 4]와 같다.

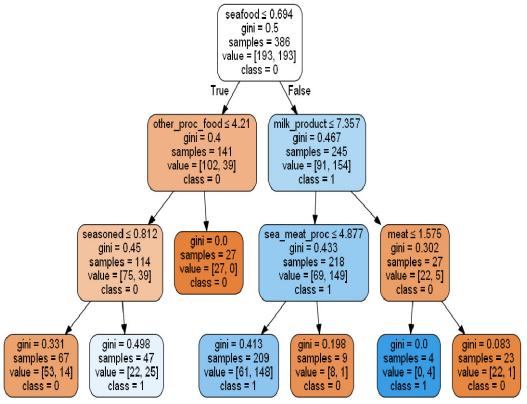

그림 3. 의사결정나무 분석 과정 시각화

[그림 3]에서 보는 것과 같이 수산물(seafood), 유제품(milk_product), 수산축산가공식품(sea_meat_prod) 순으로 대사성 질환을 분류하는데 주요하게 작용하는 것으로 분석되었다.

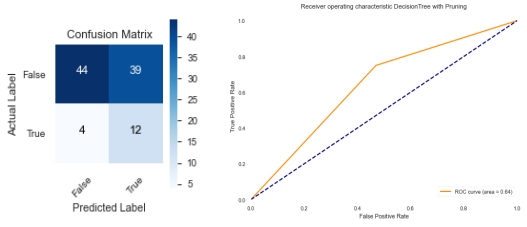

그림 4. 의사결정나무 분석 결과

의사결정 나무의 roc_auc_score 값은 0.64로 로지스틱보다는 좋지만, 결론적으로 좋지 않은 결과를 나타내었다. 이를 보완하기 위해 의사결정 나무와 앙상블 중 정확도가 높다고 알려진 XGBoost를 사용한 분류 모델을 생성하였다.

3.4 XGBoost를 이용한 분류 모델

본 논문은 의사결정나무의 대부분 단점을 보완하는 XGBoost를 사용하여 대사성 질환에 미치는 변수를 찾고자 하였다. 의사결정 나무와 마찬가지로 매개변수를 완전탐색으로 조정하였으며, 데이터에 최적이 되는 매개변수인 learning_rate = 0.49, max_depth = 5로 지정하였다. 학습결과는 [그림 5]와 같다.

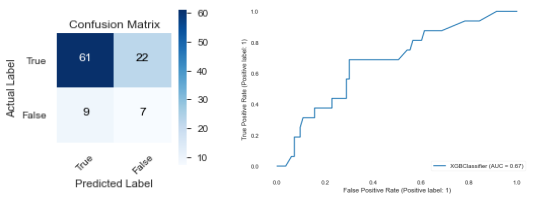

그림 5. XGBoost 분석결과

XGBoost 알고리즘을 사용한 결과 roc_auc_score를 0.67까지 올릴 수 있었다. 하지만 최종 모델도 0.7을 넘지 못하는 스코어를 가지고 있음으로 해당 데이터로는 대사성 질환의 유무에 영향을 주는 변수를 찾기에는 어려움을 발견하였다. 이는 [그림 4]의 결과처럼 환자군을 잘 예측하지 못하는 결과 때문이다. 원본 데이터의 사이즈가 워낙 작고 오버샘플링을 하더라도 원본 데이터의 부재로 인해 대사성 질환 분류의 정확도가 낮은 것으로 보인다[16].

3.5 분류 모델의 평가

지도학습 기법의 하나인 분류분석을 사용하여 대사성 질환 발병에 영향을 미치는 주요 식품 변수를 찾고자 하였다. 종속변수는 대사성 질환 여부로 설정하고 독립변수로 14개의 식품군별 월별 식품 소비량으로 설정하였다. 분석결과 [표 3]과 같다.

표 3. 머신러닝 분류 모델 분석결과 비교

구분	사용 모델	roc_auc_score
1	LogisticRegression	0.61
2	Decision Tree	0.64
3	XGBoost	0.67

클래스의 불균형을 해결하기 위해 SMOTE 기법을 사용하였다. 또한, 다양한 머신러닝 기법을 사용하고 최적의 매개변수 탐색까지 하였지만, 해당 모델의 성능을 높이는 데에는 한계가 있었다[16].

Ⅳ. 결론

본 논문에서는 건강보험 데이터를 이용하여 대상자가 관찰 기간 안에 당뇨(E11, E13~14) 또는 혈압성 질

환(I10-I15, I20~25) 또는 비만(E66) 또는 이상지질혈증(E78)으로 진단을 받았으면 환자군으로 그렇지 않으면 대조군으로 정의하고 로지스틱 회귀모형, 의사결정나무 모형, XGBoost를 활용한 분류모형을 구축하였다. 세 가지 분류모형 중에서 XGBoost가 가장 높은 정확도를 나타냈지만 0.7 미만의 ROC 점수를 보였다.

본 연구에서는 환자군-대조군에서 특징 변수인 식품소비를 가지고 다양한 연구방법을 제안하였다. 또한, 기존 많은 연구가 하나의 식품이나 특정 기능성 성분의 섭취가 질환 발생이나 관리에 미치는 영향을 분석하였지만, 본 연구는 일반 식사에 포함되는 모든 식품군을 대상으로 환자군-대조군의 식품 소비 특성을 비교하고자 하였다.

연구의 성과는 6년간의 농식품 구매내역 데이터를 기반으로 개인의 식품소비 특성을 분석하고 이러한 식품섭취 특성을 활용하여 대사성 질환자 여부를 분류하는 모델을 구축했다는 것이다. 이번 연구에서 높은 정확도의 분류 모델을 생성하지는 못했지만 향후 연구를 통해 대사성 질환을 예방하는 식품 소비 특성 연구, 영양섭취 제안 또는 국민건강관리를 위한 정책 발굴에 활용할 수 있을 것이다.

대사성 질환은 2~3년간으로 짧은 시간 동안 발생되는 질환이 아니고 식습관을 포함한 다양한 환자의 생활양식을 통해 장기간에 의해 진행되다가 발병하게 된다. 하지만 본 연구는 환자군의 식품 소비 관찰 기간이 2015년부터 2020년으로 질환 진단 전의 관찰기간은 4년 미만으로 단기간이기 때문에 분류 모델의 정확도가 낮았다. 본 연구의 한계점은 대사성질환 진단 전의 식품 소비 특성을 파악하기에 관찰기간이 짧다는 점과 가구 단위의 식품 소비데이터를 토대로 개인의 식품 소비로 간주하여 보았다는 점이다. 향후 연구에서는 연구대상자 수와 농촌진흥청 농식품 소비데이터의 시간 범위를 확대하고, 가구구성원마다 선호식품군을 조사하여 개인 섭취 식품을 산출할 때 반영하는 방법을 연구할 필요가 있다. 그리고, 구매한 농식품 상품의 영양 정보를 연계하여 식품 소비를 통해 섭취한 열량, 탄수화물, 지방, 단백질, 식이섬유, 비타민, 무기질, 미네랄 등 영양소 특성이 각종 질환발생에 어떠한 영향을 주는지 분석하는 연구를 수행하고자 한다.

참 고 문 헌

[1] 원진기, 한두봉, “식품영양표시와 운동이 고혈압 집단의 식생활패턴에 미친 영향,” 농촌경제, Vol.42, No.3, pp.55-84, 2019.

[2] 강지혜, 유리나, “비만성 염증/대사질환 제어를 위한 기능성 식품성분의 활용 가능성,” 대한비만학회지, Vol.21, No.3, pp.132-139, 2012.

[3] B. Annie, P. Ann-Marie, R. Iwona, L. Simone, and C. Patrick, and V. Marie-Claude, “Associations between dietary patterns and gene expression profiles of healthy men and women: a cross-sectional study,” Nutrition journal, Vol.1, No.12, pp.1-13, 2013.

[4] Czekajło, Różańska, Zatońska, Szuba, Regulska-Ilow, “Association between dietary patterns and metabolic syndrome in the selected population of Polish adults—results of the pure Poland study,” European journal of public health, Vol.2, No.29, pp.335-340, 2019.

[5] 정미미, 엄한주, “Two-way ANOVA 분석절차 및 사후검증방법의 이해,” 한국체육측정평가학회지, Vol.13, No.2, pp.1-15, 2011.

[6] Safari, Shariff, Kandiah, Rashidkhani, Fereidooni, “Dietary patterns and risk of colorectal cancer in Tehran Province: a case-control study,” BMC Public Health, Vol.13, No.222, 2013. https://doi.org/10.1186/1471-2458/13/222

[7] T. Mizoue, T. Yamaji, S. Tabata, K. Yamaguchi, S. Ogawa, M. Mineshita, and S. Kono, “Dietary patterns and glucose tolerance abnormalities in Japanese men,” The Journal of nutrition, Vol.136, No.5, pp.1352-1358, 2006.

[8] 유소영, 홍혜숙, 이현숙, 최영주, 허갑범, 김화영, “제 2 형 당뇨병 환자에서 인슐린저항성과 심혈관질환 위험요인 및 식이요인과의 관계,” 한국영양학회지, Vol.40, No.1, pp.31-40, 2007.

[9] 홍남기, 박혜정, 이유미. “특집 : 당뇨병 및 내분비질환 분야 머신러닝 활용,” 당뇨병(JKD), Vol.21, No.3, pp.130-139, 2020.

[10] B. G. Choi, S. W. Rha, S. W. Kim, J. H. Kang, J. Y. Park, and Y. K. Noh, “Machine learning for the prediction of new-onset diabetes

mellitus during 5-year follow-up in non-diabetic patients with cardiovascular risks," Yonsei Med J, Vol.60, pp.191-199, 2019.

[11] W. Seo, Y. B. Lee, S. Lee, S. M. Jin, S. M. Park, "A machinelearning approach to predict postprandial hypoglycemia," BMC Med Inform Decis Mak, Vol.19, p.210, 2019.

[12] 조우균, "한국 남자 당뇨환자의 식품 섭취 실태 조사," 韓國食品營養學會誌, Vol.6, No.3, pp.143-157, 1993.

[13] Andreas C. Müller and Sarah Guido, "Introduction to machine learning with Python: a guide for data scientists," O'Reilly Media, Inc, 2016.

[14] 홍준호, 오민지, 조용빈, 이경희, 조완섭, "다차원 데이터의 군집분석을 위한 차원축소 방법: 주성분분석 및 요인분석 비교," 학술지명 삽입, Vol.5, No.2, pp.135-143, 2020.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of artificial intelligence research, Vol.16, pp.321-357, 2002.

[16] 홍준호, *머신러닝과 통계적 방법을 이용한 대사성 질환과 식품 소비와의 관계성 연구*, 충북대학교, 석사학위논문, 2022.

저 자 소 개

홍 준 호(Jun Ho Hong) 정회원

- 2015년 2월 : 고려대학교 정보통계학과(이학사)
- 2022년 2월 : 충북대학교 빅데이터협동과정(공학석사)
- 2022년 1월 ~ 현재 : ㈜현대자동차 데이터인텔리전스팀 매니저

〈관심분야〉 : 데이터 마이닝, 데이터 분석

이 경 희(Kyung Hee Lee) 정회원

- 1999년 2월 : 충북대학교 전자계산학과(이학석사)
- 2022년 2월 : 충북대학교 빅데이터협동과정(이학박사)
- 2020년 3월 ~ 2021년 12월 : ㈜힐링소프트 데이터사업부(이사)
- 2022년 1월 ~ 현재 : 충북대학교 경영정보학과 초빙교수

〈관심분야〉 : 데이터 분석, 데이터 거버넌스, 블록체인

이 혜 림(Hye Rim Lee) 정회원

- 2008년 2월 : 동국대학교 통계학과(이학사)
- 2010년 2월 : 동국대학교 통계학과(이학석사)
- 2013년 1월 ~ 현재 : 농촌진흥청 농업연구사

〈관심분야〉 : 농업 빅데이터 분석 활용

정 환 석(Hwan Suk Cheong) 정회원

- 2014년 8월 : 경북대학교 컴퓨터공학과(공학석사)
- 2017년 2월 : 전남대학교 정보보안협동과정(이학박사)
- 2019년 7월 ~ 현재 : 농촌진흥청 전산사무관

〈관심분야〉 : 데이터, 개인정보보호, 스마트팜, 스마트시티

조 완 섭(Wan-Sup Cho) 정회원

- 1987년 : KAIST 전산학과(공학박사)
- 1996년 ~ 현재 : 충북대학교 교수

〈관심분야〉 : 빅데이터, 블록체인, 빅데이터거버넌스