KNN 알고리즘을 기반으로 하는 질병 예측 및 건강기능식품 추천 알고리즘에 관한 연구

(Research on Disease Prediction and Health Supplement Recommendation Algorithm Based on KNN Algorithm)

추용주*

(Yong-Ju Chu)

요 약

본 논문에서는 최근 고령화 사회로 진입하면서 건강기능식품에 높은 관심과 머신러닝의 발달로 질병을 고려한 맞춤형 건강기능식품을 추천할 수 있는 알고리즘을 제시하였다. KNN 알고리즘을 적용하여 질환에 대한 분석과 공개된 건강기능식품 정보, 국가 공공데이터의 매칭 기법을 적용하여, 맞춤형 건강기능식품 추천에 대한 플랫폼의 기초 워크프레임을 제시하였다. 신뢰성 높은 질환 대비 건강기능식품 사이의 매칭을 위해서, 상관관계를 분석하고, KNN알고리즘의 고도화를 위한 변수의 적절성과 정확도를 분석하고, 향후 공개되는 정보와 학습 모델의 개선을 통해 제안하는 시스템의 개선 방향에 대해서 도출하였다.

■ 중심어 : KNN알고리즘 ; 질병 예측; 건강기능식품 ; 상관관계 ; 공공데이터

Abstract

In this paper, we propose an algorithm that can recommend personalized health functional foods considering diseases due to the high interest in health functional foods and the development of machine learning as society enters an aging phase. By applying the KNN algorithm, we presented a foundational framework for a platform for personalized health functional food recommendations through disease analysis, matching techniques of publicly available health functional food information, and national public data. To ensure reliable matching between diseases and health functional foods, we analyzed correlations, assessed the appropriateness and accuracy of variables for enhancing the KNN algorithm, and derived improvement directions for the proposed system through the improvement of learning models and information to be disclosed in the future.

■ keywords : KNN algorithm ; disease prediction ; health functional foods ; correlation ; Public data

Ⅰ. 서 론

현재 대한민국은 초고령화 사회로 진입하고 있으며, 보건 복지 및 의료 수준이 높아짐에 따라 개인 건강관리에 관심이 많아지고 있다. 국내·외 대형 온오프라인을 통해 형성된 국내 건강기능식품 시장의 매출 규모는 6조원 이상이며, 꾸준하고 성장하고 있다. 지속적인 시장의 양적 성장에서 질적 성장으로 변화되고 있으며, 개인 맞춤형 건강기능식품 제공 서비스가 이뤄지고 있다.

과거 약사의 처방, 권유 등을 통해 추천되던 시스템은 오남용의 발생이 높으며, 버려지는 의약품은 수질, 토양 등 환경 오염의 주요 원인이 된다. 현재는 빅데이터, 인공지능 등의 발달로 개인 맞춤형 건강기능식품 추천 서비스가 빠르게 발전하고 있으며, 웹 기반 추천, 휴대폰 애플리케이션 추천 등과 같은 다양한 서비스를 시행하고 있다.

본 논문에서는 공공데이터 및 오픈 데이터를

* 정회원, 신한대학교 사이버드론봇군사학과
이 논문은 2023년도 신한대학교 학술연구비 지원으로 연구되었음

접수일자 : 2024년 06월 27일
수정일자 : 2024년 08월 04일
게재확정일 : 2024년 08월 09일

활용하여, 건강기능식품의 선택과 개인의 건강 및 질환과 상관관계 가능성을 확인하고자 한다. 개인 건강 상태와 건강기능식품의 상관관계를 모르는 상태에서 임의의 섭취는 효과를 저해할 수 있고, 부작용을 초래할 수 있다. 또한, 수요자의 필요 영양성분은 연령, 성별, 질환 유무, 생활 습관, 식습관, 생체 활동 수준 등 제 3의 요인이 작용할 수 있다. 위와 같은 문제점을 해소하고 수요자가 자신의 건강 상태를 고려하여, 필수 영양성분을 적절하게 섭취한다면, 장기적인 관점에서 건강 유지에 많은 도움을 줄 수 있다. 최근 국민건강보험관리공단이나 건강보험심사평가원 등을 통해 국민 건강 기초건강 데이터의 수집과 활용이 이뤄지고 있으며, 개인정보 보호 처리가 된 데이터를 활용하여 공공의 연구에 활용할 수 있도록 제공하고 있다. 이와 같은 수집된 기초건강 데이터를 활용해 다양한 보건 및 의료 분야에 연구가 진행되고 있다.

본 논문과 같이 기초건강 데이터에 머신러닝 알고리즘을 적용하여 개인의 건강 상태를 분석하고 질환을 예측하는 다양한 연구가 진행되고 있다. 특히, 군집화와 분류에 강점이 있는 KNN 알고리즘이 주로 적용되고 있다[1].

본 논문에서는 머신러닝 방법 중 KNN 알고리즘을 이용해 개인의 질환 상태를 예측하고, 해당 질환 도움이 되는 맞춤형 건강기능식품을 추천하는 서비스를 제안하고 해당 서비스의 활용 및 발전 가능성에 대해 논의하고자 한다. 과거 오랜 역사가 있는 분류 알고리즘인 KNN 알고리즘은 구현이 간단하면서, 뛰어난 접근성을 가지고 있다. 먼저 KNN은 테스트 샘플에 대해 k개의 가장 가까운 훈련 샘플을 선택한 후, k개의 가장 가까운 주요 데이터들로 테스트 모집단의 속성을 예측하는 알고리즘 특성에 따라, 모집단의 크기에 따라 선형 시간 복잡성이 상승한다. 다만, 이와 같은 작동 방식은 다양한 인자가 복합적으로 작동하는 환경에서는 높은 에러율 및 낮은 분류 정확성을 가진다. 이 문제점을 보완한 k-평균 클러스터링이라는 방법은 데이터를 여러 부분으로 분리한 후 가장 가까운 클러스터를 훈련 샘플로 선택하고 분류를 수행한다. 클러스터링의 대표적인 예시로 랜드마크 기반 스펙트럴 클러스터링(LSC)이 있다. LSC는 선형 조합으로 원래 샘플을 표현하는 방식으로 데이터를 분석한다.

빅데이터 분석에 맞춰 발전한 KNN 알고리즘은 다양한 분야에 활용될 수 있으며, 예측 및 추천 시스템으로 많은 분야에서 활용되고 있다[2, 3]. 특히, 의료 및 헬스케어 분야에서도 KNN 알고리즘으로 환자의 건강 상태나 질환 유무를 예측할 수 있고, 관련 연구가 다수 진행되고 있다[4-6].

최근에는 수집된 환자의 기초건강 데이터를 활용해 질환을 예측하는 연구가 활발하게 진행되고 있다. 심현 등은 당뇨병 환자들의 데이터 전처리 과정들을 거친 기초 건강데이터에 KNN 알고리즘을 적용하여 당뇨병 진단 결과를 예측하고, KNN 알고리즘 등이 당뇨병 진단 예측에 활용될 수 있음을 제시하였다[6].

또한, 김성수 등은 KNN 알고리즘이 파킨슨병을 조기에 예측하는데 효과적일 수 있음을 보여주고 있으며, 해당 연구에 따르면 일반적인 분류 알고리즘을 사용한 파킨슨병 예측율은 86.7%이며, KNN으로 예측한 결과값은 91.8%로 5.1% 웃도는 수치를 나타내고 있어, KNN 알고리즘이 다양한 질환 예측에 사용될 수 있음을 제시하였다[5]. 최근에는 빅데이터 분석 기법의 발달로 협업 필터링 기술로 사용자-상품 행렬 생성, 이웃 집단 탐색, 추천 목록 생성의 3단계로 구성된 코사인 유사도를 활용한 확장된 유사도를 제시하는 기법도 발전했다[7-9].

Ⅱ. 추천 및 분석 시스템 구축

1. 질환과 건강기능식품 상관관계 및 맞춤형 추천 알고리즘

본 논문에서 제시하는 개인 맞춤형 건강기능식

품 추천 알고리즘 구현을 위해, 분석하고자 하는 질환에 대한 정의가 필요하다. 일반적으로 국내 한국표준질병사인분류에 따라 7대 중, 발병 인자가 상대적으로 명확한 당뇨, 뇌졸중, 고혈압으로 분류되는 질환을 선정하였다. 이는 선행 연구된 공공데이터 및 연구자료들이 높은 신뢰성을 가지고 있으며, 질환군의 하위 분류에는 질환군에서 세분화되는 여러 질환군의 분류에 관한 사전 연구가 이뤄져 있다. 해당 질환과 상관관계를 나타내는 건강기능식품도 당뇨, 뇌졸중, 고혈압에 대한 예방 또는 치료에 직·간접적인 도움이 될 수 있음을 제시하고 있다. 단, 건강기능식품과 의약품의 차이인 용량 및 작용 기저의 차이로 인해 개선의 가능성으로 볼 수 있고, 치료 이외의 목적을 가지지는 않는다.

본 논문에서 제시하고 있는 당뇨, 뇌졸중, 고혈압 개선을 위한 추천 프로세스는 질환과 상관관계를 가지고 있는 개인 문진을 기준으로 추천할 수 있도록 그림 1과 같이 프로세서를 정립하였다. 그림 1과 같은 프로세서는 유재준 등과 같은 KNN 알고리즘 예측 프로세스와 유사한 방식을 참고하였다[2, 10]. 그림 1에서 제안하는 프로세스를 적용하기 이전에 사용자의 기초건강데이터를 수집하여 분류 기준을 정립할 필요성이 있다. 본 논문에서는 병원 문진의 기준을 적용하여, 체중, 키, 혈액형, 혈당 등 본인의 건강 기초 상태 지표를 결정할 수 있는 자료를 적용하였다. 이때 사용자 식별을 위해 ID, 성별, 나이 등에 관한 정보를 수집하고, DBMS에 식별자로써 저장하였다. 또한, 일반적인 웹기반 서비스와 DBMS와의 연결, 통신은 Spring 기반으로 구현된다. 질환 및 건강기능식품 추천 시스템은 별도의 서버(저장장치)에 저장되며, 예측과 추천의 상관관계를 통해 도출된 결과값은 별도의 데이터 처리 분석을 위한 NoSQL, DBMS에 저장하였다.

추후, 사용자가 설문조사를 완료하면, 서버로 결과값 출력을 요청하고, NoSQL DBMS에서 웹기반 서비스 서로로 반환한다.

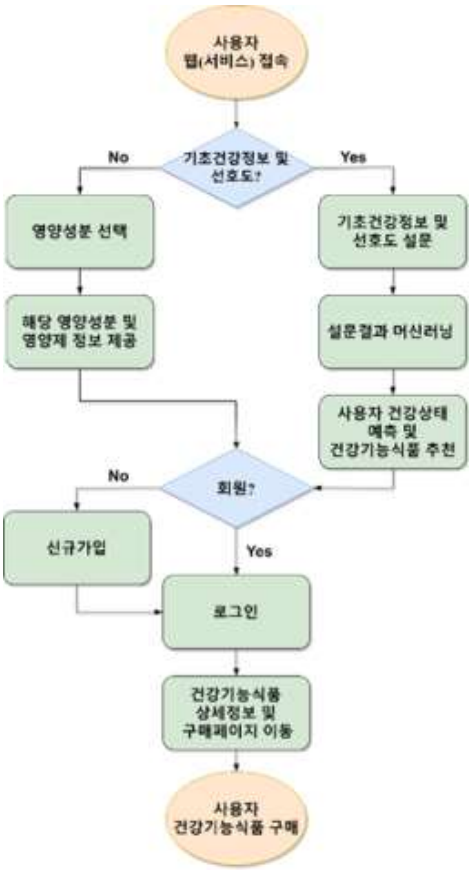


그림 1 KNN알고리즘 기반의 맞춤형 건강기능식품 추천 시스템 프로세스

2. 개인화 데이터 획득

본 논문에서는 KNN 분류 알고리즘을 활용한 질환 분류 및 예측 시스템과 건강기능식품 추천 서비스 구현을 위해 공공데이터를 활용하였다.

최근의 과학 기술 발전과 사회적인 변화로 맞춤형 건강관리에 대한 소비자들의 관심이 증가하고 있고, 맞춤형 건강 기능식품에 대한 법적 규제 측면과 개인정보 보호의 균형이 필요하여 민감하게 다뤄야 한다[8, 11]. 따라서, 건강보험심사평가원에서 제공하는 제한적인 공공데이터를 API 연동을 통해 분절되어 있는 데이터를 결합하였다.

본 논문에서 질환과 건강식품 간 상관관계 구성에 대한 정의가 필요하다. 첫 번째로 정부에서 제공하는 공공데이터포털*을 통해 연령대별 건강검진내역을 수집하여, 주요 7대 질환과 연령간 상관관계를 규명하였다. 또한, 건감보험심사평가

* 공공데이터포털의 웹사이트 주소는 www.data.go.kr이다.

원*을 통해 질환 정보 및 건강검진내역으로 질환간 상관관계의 정밀도를 높였다. 또한 KOSIS 통계청에 제공하는 공유서비스**를 통해 인적사항이 비공개 처리된 건강검진내역을 활용하였다. 또한, 전세계적인 빅데이터 연구자들에 의해 후처리되고 정량화된 데이터를 얻을 수 있는 Kaggle***을 통해, 질환 상관관계 정보, 건강 상태 정보 등을 제공받았다. 해당 데이터는 DB구축 및 결과 데이터의 기준 마련과 질환 예측 시스템의 상호 보완을 위해 적용하였다.

두 번째로 건강기능식품 데이터 획득을 위해, 식약처에서 제공하고 있는 식품안전나라****에서 건강기능식품의 영양 정보 API를 연결하고, 해당 질환과 자동으로 연결했다. 다만, 해당 질환에 필요한 영양소 정보는 전문가 의견에 따라 상관관계를 사전 반영하였다.

본 논문에서 수집한 데이터는 대한민국 7대 질환 중에서 상대적으로 노출, 발현 빈도 등이 높은 고혈압, 당뇨, 뇌혈관질환을 중심으로 데이터를 획득 및 가공하였다. 획득한 데이터 중 질환 예측 시스템 구현에 요구되는 데이터셋은 일반인과 환자의 일반건강검진 등의 의료검진 결과이며, 신장, 체중, 혈압, 혈당, 총콜레스테롤, 혈색소 등의 세부 검진 내역을 포함하고 있다.

또한, 질환 예측 시스템 구현의 질환 판별 기준을 마련하기 위해 병원 일반 문진표 및 진단 관련 정보 데이터도 결합하였다[2, 11, 12].

최종적으로 건강기능식품 추천 시스템 구현을 위해 각 건강식품의 영양과 효능 등의 내용이 있는 건강기능식품 데이터를 오픈 API 호출을 통해 획득하였다. 원활한 데이터 전처리 수행 및 머신러닝 알고리즘 학습을 위해 Kaggle의 추천수 200이상 질환 예측 관련 데이터셋과 공개된 기본 구조를 참조하였다[13].

* 건강보험심사평가원의 웹사이트 주소는 www.opendata.hira.or.kr 이다.
** KOSIS 통계청 공유서비의 웹사이트 주소는 www.kiosis.kr 이다.
*** Kaggle의 웹사이트 주소는 www.kaggle.com이다.
**** 식품안전나라의 웹사이트 주소는 www.foodsafetykorea.go.kr 이다.

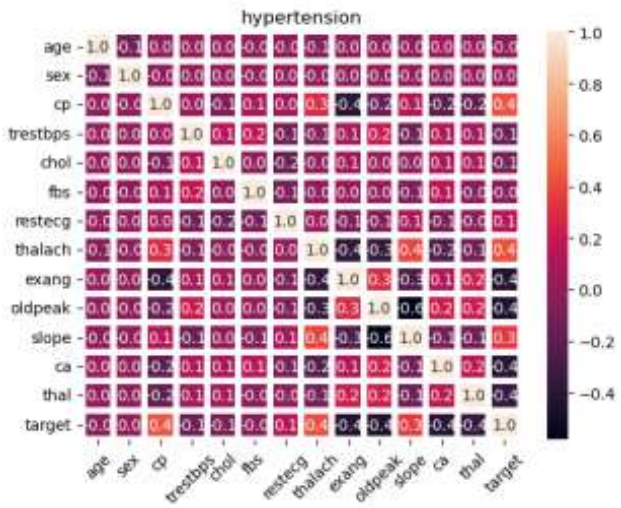
3. 데이터 전처리 및 DB 구축

본 논문에서 제안하고자 하는 개인 맞춤형 건강기능식품 추천 알고리즘을 구현하기 위한 공공데이터는 각각 포맷과 구성이 다르므로 병합하여 사용하면, Null 값 또는 중복되는 문제를 가지고 있다. 기본 원시데이터는 건강정보 이외에 종합검진을 통해 시력, 청력, 충치 등 질환 예측과 다소 상관관계가 낮은 데이터도 있다.

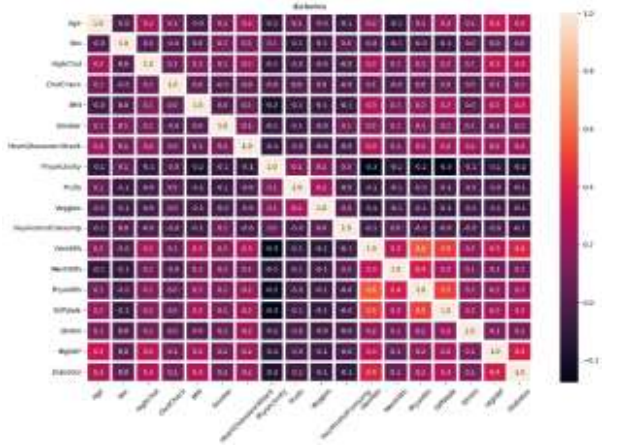
이에 따라, 추천 알고리즘에 원시데이터를 직접 사용하기 전에 전처리를 수행한다. 각 변수 사이의 높은 상관관계를 도출하기 위해 탐색적 데이터 분석으로 전처리를 진행하여, 원시데이터를 분석하고 시각화함으로써, 데이터 포맷, 형식, 속성 변수의 상관관계를 높일 수 있다. 다만, 지나친 개입은 데이터의 순수성, 상관관계의 순수성을 낮출 수 있는 문제점을 가지고 있다.

본 논문에서는 질환 유무는 종속변수로 설정하고, 발병 인자 및 관련 변수를 독립변수로 설정하였다. 이를 info()를 기법으로 전체적인 포맷을 검색 및 분석하였다. 이에 따라, 성별, 음주 등 범주형 변수는 0과 1로 One-Hot 인코딩을 수행하였다. 그 결과 앞서 언급한 종합검진을 통해 얻을 수 있는 시력, 청력, 충치 등과 사회적 변수와 같은 상관관계가 낮은 변수는 Drop() 기법을 적용하여 제외하기로 했다. 각각의 성격이 다른 데이터의 정규화를 위해 본 논문에서는 Standard Scaler를 이용하여 변수 정규 분포화를 진행하였다. 이때 수치형 데이터의 결과값은 평균값으로, 범위형 데이터의 결과값은 중앙값으로 변경하였다.

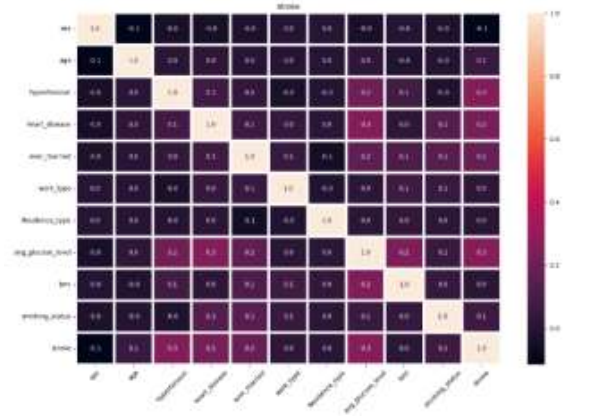
본 논문에서는 전처리 데이터와 건강기능식품 데이터를, MySQL을 이용하여 저장 및 서버화함으로써 데이터베이스를 지속해서 업데이트할 수 있는 구조를 만들었다. 본 데이터베이스의 테이블을 데이터 분석용, 건강기능식품 분석용, 개인정보 관리 및 서비스 제공용으로 분류하여 데이터를 저장하였다. 또한, 각 테이블 간 종속변수를 제거하여 독립변수로 활용될 수 있도록 정규화를 추가로 진행하였다.



(a) 고혈압 질환 인자 사이의 상관계수 분석



(b) 당뇨 질환 인자 사이의 상관계수 분석



(c) 뇌졸중 질환 인자 사이의 상관계수 분석

그림 2 질환별 상관관계 분석 (a) 고혈압 질환 인자 사이의 상관계수 분석, (b) 당뇨 질환 인자 사이의 상관계수 분석, (c) 뇌졸중 질환 인자 사이의 상관계수 분석

Ⅲ. 질환과 건강기능식품 추천 상관관계 및 결과

1. KNN 알고리즘을 적용한 질환 인자와 상관관계 분석

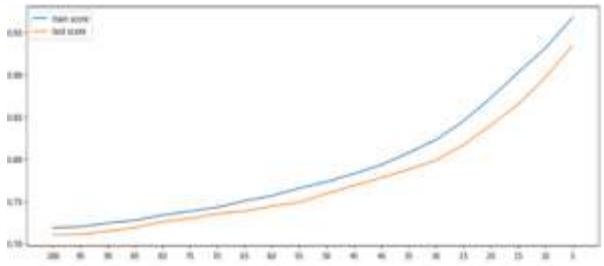
본 논문은 범용 KNN알고리즘을 사용하여 각 변수간 상관관계를 회기분석으로 중요 인자를 도출한 후, KNN 알고리즘을 구현할 수 있도록 했다. 학습 모델 생성 전, 데이터별 상관관계가 높은 변수를 중심으로 학습 모델 분석을 진행하였다[14, 15].

그림 2은 고혈압, 당뇨, 뇌졸중, 심장병에 대한 상관계수 분석 히트맵이다. 고혈압 관련한 질환은 그림 2(a)에서 확인할 수 있듯이 이완기혈압을 학습 모델로 정했고, 그림 2(b)를 통해 당뇨 질환과 상관관계가 0.4에서 0.6 이상으로 높게 나타나는 열을 배합하여 학습을 진행하였다. 일반적으로 당뇨 질환에서 0.7 이상이 강한 상관관계로 볼 수 있다[15]. 그러나, 그림 2(c)와 같은 뇌졸중은 상관계수가 낮게 나타났다. 이는 뇌졸중과 같은 질환은 다양한 인자들의 복합적인 상관관계로 진행되기 때문에 일반적으로 규명된 비만, 흡연, 당뇨 등과 같은 데이터와 연동할 수 있도록 전처리하였다.

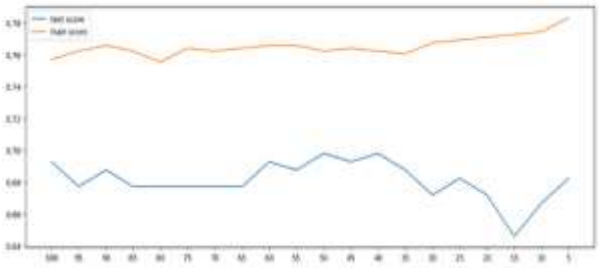
2. 질환별 학습 모델 분석

본 논문에서는 사전 상관관계 분석을 통해 학습 모델 과정에서 사용할 데이터를 선별하였다. 학습과 평가 데이터 비율을 8:2로 정하고, 랜덤 함수를 활용하여 무작위로 분류될 수 있도록 한 후, 질환별 정확도를 분석하였다.

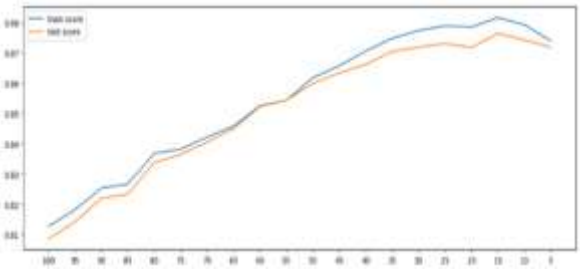
본 논문에서 정확도 분석은 상관계수 및 상관관계를 중심으로 참고하지만, 상관계수가 높아도 반드시 인과관계로 이어지는 것은 아니다. 예측 모델 생성 과정에 인과관계가 불분명한 질환에서 과대 또는 과소 적합 문제가 발생했다.



(a) 고혈압 질환 모델 정확도 분석



(b) 당뇨 질환 모델 정확도 분석



(c) 뇌졸중 질환 모델 정확도 분석

그림 3 질환별 학습 모델 정확도 분석 (a)고혈압 질환 모델 정확도 분석, (b)당뇨 질환 모델 정확도 분석, (c)뇌졸중 질환 모델 정확도 분석

위의 그림 3(a)는 고혈압 질환 모델의 정도 분석으로써 비만도를 제거하여 0.95수준의 신뢰도를 확보하였다. 즉, 비만도라는 단일 변수가 상관관계 및 신뢰도에 차지하는 상관관계가 낮음을 의미한다. 그림 3(b)에서 확인할 수 있듯이, 당뇨 질환은 비만, 혈당, 생활 습관, 성별 다양한 인자들이 작용하고 있으나, 학습 모델은 0.75의 신뢰도를 보이고, 테스트 모델은 0.7의 신뢰도를 보인다. 이는 앞서 언급한 독립 인자들의 영향로 판단된다. 그림 3(c)는 뇌졸중 질환의 신뢰도를 나타내는데, 연령, 성별을 제거하여 0.9 이상의 높은 신뢰도를 보이는 것을 확인할 수 있다.

3. 질환 예측 알고리즘 모델 평가 및 결과

본 논문에서 구축하고자 하는 시스템은 질환과 건강 상태의 상관관계에서 건강기능식품을 추천하는 것이다. 각 질환의 회복이나 치료에 효과가 큰 필수 영양소를 구별하고, 우선순위를 부여하였다[16]. 개인의 질환 예측 후, 결과 변수를 저장한다. 이를 바탕으로 식품안전나라와 연동 API에 따라 건강기능식품 데이터베이스에 저장된 영양소 종류를 기준으로 검색, 추출, 반환한다. 이때, 예측된 질환의 영양소별 섭취 요구량에서 각 제품의 영양소 함유량의 차를 오름차순으로 정렬하였다. 이 추천 알고리즘을 지속해서 반복 재사용 및 학습하여 정확도 및 신뢰도를 높인다. 다만, 영양소 간 상극 관계 및 음식과 상극 관계 등으로 기능을 저해하거나, 부작용 우려가 발생할 수 있다.

이를 해결하기 위해서 Chat-GPT 4.0과 API 연동을 통해 영양소 간 상극 조합 및 섭취 주의사항을 공지할 수 있도록 연동했다. 즉, 건강기능식품의 영양소 간 상극 조합이 발생하면, 해당 키워드를 추출해 데이터베이스에서 배제하는 방식을 선택했다.

본 논문에서 제안하는 KNN알고리즘 기반 예측 모델을 정확도, 정밀도, 재현율로 평가하였다. 그리고 F-1 Score를 Scikit-learn 모듈을 이용하여 계산하였다. 일반적으로 정밀도와 재현율의 조화평균으로 계산되는 F-1 Score는 높을수록 모델 성능이 우수한 것으로 평가된다.

본 논문에서 적용한 모델 데이터를 진양성(True Positive, TP), 진음성(True Negative, TN), 위양성(False Positive, FP), 위음성(False Negative, FN)으로써, F-1 Score는 다음과 같이 식(1) ~ (4)와 같이 표현한다.

$$Accuracy = \frac{Number\,of\,Correct\,\mathrm{Pr}edictions}{Total\,\mathrm{Pr}edictions} \quad (1)$$

$$\mathrm{Pr}ecision = \frac{TP}{TP+FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

$$F-1\,Score = \frac{1}{\frac{1}{Rrecision}+\frac{1}{Recall}} \quad (4)$$

본 논문에서는 시험 평가를 위해 학습과 평가 비율을 8:2으로 설정하였다. 아래의 그림 4(a)와 같이 뇌졸증 질환은 근접 이웃점 15에서 정밀도, 0.99, 재현율 0.75, F-1 Score 0.85, 지지도 48679, 정확도 75%로 나타났다. 또한, 그림 4(b)와 같이 당뇨병 질환은 근접 이웃점 40에서 정밀도 0.95, 재현율 0.66, F-1 Score 0.78, 지지도 42795, 정확도 64%로 나타났다.

그림 4(c)에서 고혈압 질환은 근접 이웃점 15에서 정밀도 0.73, 재현율 0.68, F-1 Score 0.7, 지지도 28810, 정확도 67%로 나타났다.

```
Training Accuracy: 0.7532324187953326
Test Accuracy: 0.7536266162093976
              precision    recall  f1-score   support

         0.0       0.99      0.75      0.85     48679
         1.0       0.11      0.74      0.20      2057

    accuracy                           0.75     50736
   macro avg       0.55      0.75      0.53     50736
weighted avg       0.95      0.75      0.83     50736

F1-Score: 0.5251302323973107
Precision Score: 0.5492015578173965
Recall: 0.7468265088180033
```

(a) 뇌졸중 질환 예측 결과

```
Training Accuracy: 0.6391861794386628
Test Accuracy: 0.6419110690633869
              precision    recall  f1-score   support

         0.0       0.95      0.66      0.78     42795
         1.0       0.03      0.33      0.06       944
         2.0       0.35      0.57      0.44      6997

    accuracy                           0.64     50736
   macro avg       0.45      0.52      0.42     50736
weighted avg       0.85      0.64      0.72     50736

F1-Score: 0.4249078913354436 5
Precision Score: 0.4455614954580675
Recall: 0.5209615878571365
```

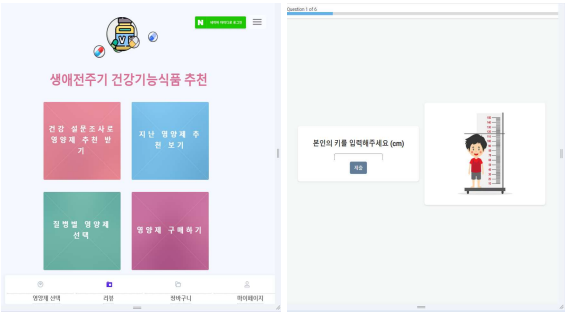
(b) 당뇨 질환 예측 결과

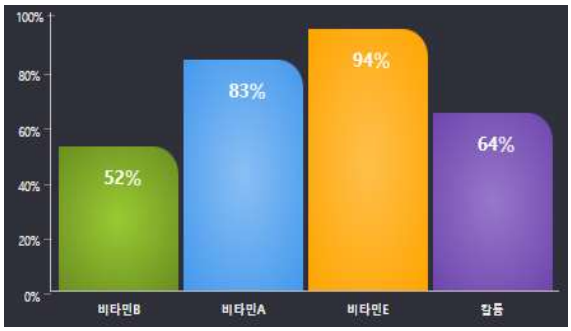
```
Training Accuracy: 0.7875226663513087
Test Accuracy: 0.6745111952065594
              precision    recall  f1-score   support

         0.0       0.73      0.68      0.70     28810
         1.0       0.61      0.67      0.64     21926

    accuracy                           0.67     50736
   macro avg       0.67      0.67      0.67     50736
weighted avg       0.68      0.67      0.68     50736

F1-Score: 0.671642811956612
Precision Score: 0.6714849586588312
Recall: 0.674242387927991
```

(c) 고혈압 질환 예측 결과

그림 4 질환별 학습 모델 예측 결과 (a)뇌졸증 질환 예측 결과, (b)당뇨 질환 예측 결과, (c)고혈압 질환 예측 결과

본 논문에서는 개인정보와 공공데이터의 질병 정보 기반 예측을 통해, 그림 5같은 생애전주기 데이터를 기반으로 질환이 예상되는 수요자에게 최소한의 영양소를 추천한다.





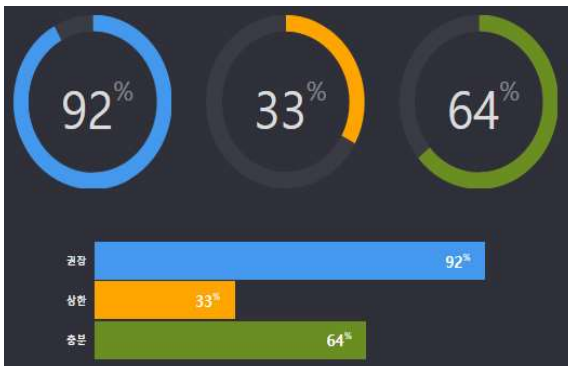


그림 5 생애 전주기 기반의 건강기능식품 추천 결과(예시)

Ⅳ. 결 론

본 논문에서 살펴본 KNN 알고리즘을 적용하여 뇌졸중, 고혈압, 당뇨와 같은 독립인가 뚜렷한 질환에서 복합 인자가 작용하는 질환까지 분류기를 적용해 보았고, 건강기능식품 추천의 가능성을 확인하였다. 향후, 개인의 일반건강검진결과와 연동하여 자동으로 맞춤형 추천 서비스를 개발할 수 있는 기초 시스템으로 활용될 수 있다. 또한, 일반적인 KNN 알고리즘을 적용해도, 질환 발병 인자의 분석과 사전 분류로 높은 정확도를 얻을 수 있음을 확인하였다.

이와 같은 시스템의 고도화 및 사전 데이터 및 학습 모델이 생성된다면, 소비자가 간단한 건강 데이터를 입력하면 해당 서비스가 연령대 및 질환 이력 등 개인 요소들이 고려하여 질환 유무를 예측할 수 있다.

KNN 알고리즘 스코어 및 높은 F-1 Score를 통해 예측 시스템의 비교적 높은 신뢰도를 확인할 수 있다. 질환 예측 결과를 토대로 해당 질환에 도움이 되는 건강기능식품에 대한 데이터베이스를 통해 소비자에게 추천하고 선택을 도울 수 있다.

본 논문에서 구축한 기초건강 데이터베이스와 KNN 알고리즘 기반 추천 시스템은 건강 기초 플랫폼으로 확장이 가능하다. 또한, 선진국과 같이 빠르게 고령화 사회로 진입하고 있는 국가 노령인구의 건강관리를 위한 부가적인 서비스로 개발될 수 있다. 본 논문에서 제안한 플랫폼과 IoT 기술 결합한다면, 헬스케어 및 의료 분야의 사각지대에 있는 노인들과 취약계층도 쉽게 접근할 수 있는 플랫폼으로 개발될 수 있다. 더 나아가 지역사회 의료 서비스에 보조적인 소임을 수행에 질환 예방 및 건강 증진에 큰 도움을 줄 수 있을 것으로 기대된다. 아울러, 현재는 키워드에 의존한 추천 시스템을 건강기능식품 및 제약정보, 영양소 함량, 주의 사항 등을 복합적으로 고려할 수 있는 가중치에 관한 연구를 진행하면 본 추천 시스템의 고도화가 가능할 것으로 판단된다.

후기

본 논문은 2023년도 신한대학교 학술연구비 지원으로 연구되었음

REFERENCES

[1] Deng, Z., Zhu, X., Cheng, D., Zong, M., and Zhang, S. “Efficient KNN classification algorithm for big data,” *Neurocomputing,* Vol. 195, pp. 143-148, Jun. 2016.

[2] 유재준, 유준영, 조재춘, “KNN알고리즘 기반의 교양과목 추천 모델 연구,” *한국컴퓨터교육학회 학술발표대회논문집,* 제24권, 제1호, 107-109쪽, 2020년 01월

[3] 유영중, 문상호, 박성호, “도심 도로의 속도 예측을 위한 KNN 알고리즘 분,” *예술인문사회융합 멀티미디어 논문지,* 제7권, 제2호, 245-253쪽, 2017년 02월

[4] 은상남, “A Study on Unsupervised Outlier Detection Algorithms in Network Intrusion Detection”, *건국대학교 박사학위논문*, 2019년 2월

[5] 김선호, 최낙훈, 오종석, “k NN과 앙상블을 이용한 항결핵약 캡슐제의 용출 예측 연구,” *한국산학기술학회,* 제24권, 제1호, 531-537쪽, 2023년 01월

[6] 김성수, 진훈, “파킨슨 질병의 다변량 수치 데이터 속성을 고려한 분류기 선택,” *대한전자공학회 추계학술대회,* 346-350쪽, 2013년 01월

[7] 심현, 김현욱, “빅데이터 기반 2형 당뇨 예측 알고리즘 개발,” *한국전자통신학회 논문지*, 제18권, 제5호, 999-1008쪽, 2023년 10월

[8] 김형일, 김형진, 신영성, 장재우, “kNN Query Processing Algorithm based on the Encrypted Index for Hiding Data Access Patterns,” *한국정보과학회*, 제43권, 제12호, 1437-1457쪽. 2016년 12월

[9] 이지성, 박종무, 박태환, 이경복, 이수주, 조용진, 이준영, “한국인을 위한 뇌졸중 발생 예측 모형 개발,” *대한신경과학회지*, 제 28권, 제1호, 13-21쪽, 2010년 01월

[10] 전유민, 원정현, 이승우, 홍수연, 백유진, 조비룡, 이형기, “개인맞춤형 건강기능식품 추천용 설문지 개발,” *대한임상건강증진학회*, 제22권, 제1호, 26-39쪽, 2022년 01월

[11] 이희성, 김은태, 김동연, “KNN 규칙과 새로운 특징 가중치 알고리즘을 결합한 패턴 인식 시스템,” *전자공학회논문지-CI*, 제42권, 제4호, 43-50쪽, 2005년 07월

[12] Ham Jun-hyuk, Son Moon-ki, “Regulatory Review and Insights for Personalized Dietary Supplements,” *식품법과 정책*, 제1권, 제4호, 243-279쪽. 2023년 12월

[13] 홍세인, 정의주, 김재경, “확장된 사용자 유사도를 이용한 CF-기반 건강기능식품 추천 시스템,” *지능정보연구학회지*, 제29권, 제3호, 1-17쪽, 2023년 03월

[14] 정주호, 이나은, 김수민, 서가은, 오하영, “환

자 IQR 이상치와 상관계수 기반의 머신러닝 모델을 이용한 당뇨병 예측 메커니즘," *한국정보통신학회논문지,* 제25권, 제10호, 2021년 10월
[15] 이지성, 박종무, 박태환, 이경복, 이수주, 조용진, 이준영, "한국인을 위한 뇌졸중 발생 예측모형 개발," *대한신경과학회지,* 제28권, 제1호, 13-21쪽, 2010년 01월
[16] 김경진, 김은주, 송유리, 김유진, 전상현, 김지연, "맞춤형 건강기능식품의 국내외 현황," *식품산업과 영양학회지,* 제25권, 제2호, 20-37쪽, 2020년 02월

저 자 소 개

추용주(정회원)

2005년 중앙대학교 기계공학과 학사 졸업
2007년 고려대학교 기계공학과 석사 졸업
2019년 서강대학교 기계공학과 박과 박사 졸업.

<주관심분야 : 로보틱스, 전기자동차 배터리 응답 특성 예측 모델, 인공지능(전이학습), 스마트팩토리>