ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
https://www.jiisonline.org
J Intell Inform Syst 2023 March: 29(1): 307~325 http://dx.doi.org/10.13088/jiis.2023.29.1.307

머신러닝을 활용한 수도권 약수터 수질 예측 모델 개발

임영우
국민대학교 비즈니스IT전문대학원
(duddn7244@naver.com)

엄지연
국민대학교 비즈니스IT전문대학원
(sellawldus@naver.com)

곽기영
국민대학교 경영대학/비즈니스IT전문대학원
(kykwahk@kookmin.ac.kr)

코로나19 팬데믹의 장기화로 인해 실내 생활에 지쳐가는 사람들이 우울감, 무기력증 등을 해소하기 위해 근거리의 산과 국립공원을 찾는 빈도가 폭발적으로 증가하였다. 자연으로 나온 수많은 사람들이 오가는 걸음을 멈추고 숨을 돌리며 쉬어가는 장소가 있는데 바로 약수터이다. 산이나 국립공원이 아니더라도 근린공원 또는 산책로에서도 간간이 찾아볼 수 있는 약수터는 수도권에만 약 6백여개가 위치해 있다. 하지만 불규칙적이고 수작업으로 수행되는 수질검사로 인해 사람들은 실시간으로 검사 결과를 알 수 없는 상태에서 약수를 음용하게 된다. 따라서 본 연구에서는 약수터 수질에 영향을 미치는 요인을 탐색하고 다양한 곳에 흩어져 있는 데이터를 수집하여 실시간으로 약수터 수질을 예측할 수 있는 모델을 개발하고자 한다. 데이터 수집의 한계로 인해 서울과 경기로 지역을 한정한 후 데이터 관리가 잘 이루어지고 있는 18개 시의 약 300여개 약수터를 대상으로 2015~2020년의 수질 검사 데이터를 확보하였다. 약수터 수질 적합 여부에 영향을 미칠 것으로 여겨지는 다양한 요인들 중 두 차례의 검토를 거쳐 총 10개의 요인을 최종 선별하였다. 최근 주목받고 있는 자동화 머신러닝 기술인 AutoML 기법을 활용하여 20여가지의 머신러닝 기법들 중 예측 성능 기준 상위 5개의 모델을 도출하였으며 그 중 catboost 모델이 75.26%의 예측 분류 정확도로 가장 높은 성능을 가지고 있음을 확인하였다. 추가로 SHAP 기법을 통해 분석에 사용한 변인들이 예측에 미치는 절대적인 영향력을 살펴본 결과 직전 수질 검사에서 부적합 판정을 받았는지 여부가 가장 중요한 요인이었으며 그 외 평균 기온, 과거 연속 2번 수질 부적합 판정 기록 유무, 수질 검사 당일 기온, 약수터 고도 등이 수질 부적합 여부에 영향을 미치고 있음을 확인하였다.

주제어 : 약수터, 수질, 예측 모델 개발, AutoML, SHAP

논문접수일 : 2022년 11월 15일 논문수정일 : 2023년 2월 17일 게재확정일 : 2023년 3월 11일
원고유형 : Regular Track 교신저자 : 곽기영

1. 서론

코로나19 팬데믹은 인류의 삶을 뒤바꾸어 놓았다. 사회적 거리두기 또는 자가격리로 인한 시간, 인원 등의 외출 제한은 사람들로 하여금 실내 생활의 비중을 증가시켰다. 하지만 지속적인 확진자 발생에 따른 코로나19 유행 상황 종료 시점의 불투명함은 점차 실내 생활에 지쳐가는 사람들에게 우울증과 스트레스를 야기하였다. 세간에서는 감염 위험에 대한 우려와 더불어 일상생활의 제약이 증가하면서 코로나19와 우울감(blue)을 합친 '코로나 블루(코로나 우울)'라는 신조어마저 탄생하였다. 다시 말해 코로나 블루는 코로나19 확산으로 일상에 큰 변화가 닥치면서 생긴 우울감이나 무기력증 등을 일컫는다. 이를 해소하기 위해 근거리의 산과 국립공원 등을 찾는 인구가 폭발적으로 증가하였다. 예로 수도권을 대표하는 북한산 탐방객은 코로나 이전 대비 약 20만명이 증가한 상황이다. 이러한 가운데, 수많은 사람들이 오다가다 걸음을 멈추고 숨을

돌리며 쉬어가는 장소가 있는데 바로 약수터이다. 산이나 국립공원이 아니더라도 근린공원 또는 산책로에서도 간간이 찾아볼 수 있는 약수터는 수도권에만 약 6백여개가 위치해 있다.

과거에는 물의 오염 수준이 높지 않아 주변 지하수 또는 하천수 등을 바로 마실 수 있었다. 하지만 경제개발 및 산업 발달 등의 부산물로 생긴 각종 환경오염물질은 수계를 빠르게 오염시켰고 이는 식수의 위협으로 이어졌다. 1991년 당시 1,500만명 가량이 거주하는 영남지역의 식수원을 오염시킨 낙동강 페놀 오염 사건은 전국민이 경악을 금치 못하게 했으며, 이후에도 여러 차례에 걸쳐 발생한 유해 화학물 유출 사고는 국민들에게 수돗물에 대한 불신을 심어주었다. 또한 경제적 수준의 향상과 함께 건강유지에 대한 관심이 증가하면서 양질의 물을 섭취하기 원하는 사람들은 경제적인 부담을 감수하면서 먹는 샘물을 구매하여 음용하거나 보다 건강한 물을 찾아 사찰, 등산로 등에 위치한 약수터를 이용해 왔다.

이처럼 등산객이나 동네 주민들이 약수터에서 목을 축이고 때로는 물을 길어와서 식수로 이용하는 등 이용객 수는 지속적으로 증가해왔음에도 약수터는 정부나 지자체의 관심 사각지대에 놓여 방치되어왔다. 2019년부터 환경부에서 구축하기 시작한 '스마트상수도관리체계' 등 상수도를 통해 공급되는 식수나 수돗물 등은 AI를 활용한 실시간 검사 시스템까지 도입하여 자동 관리가 이루어질 예정이지만, 약수터는 적지 않은 인구가 이용하고 수질 부적합과 관련한 기사들이 지속적으로 나왔음에도 긴 시간 동안 효과적인 개선이 이루어지지 못했다.

약수터의 수질 안전성을 확보하기 위해 서울시는 스마트서울맵에 '우리동네약수터'라는 테마형 지도를 추가하여 지역 내 존재하는 약수터의 수질 검사 현황을 알 수 있는 서비스를 제공하고 있고, 경기도 또한 테마형 서비스까지는 아니더라도 물정보시스템에 관련 항목을 만들어 수질 검사 결과를 알려주고 있다. 하지만 위 서비스들은 단순히 수질을 측정한 후 정보를 전달하는 사후 보고 형식이며 지역별 검사 현황도 제각각이다. 현재 수질 검사는 불규칙적이며 수작업으로 수행되기에 약수터마다 수질 측정 기계를 설치하지 않는 이상 상시 검사 또한 쉽지 않다. 심지어 수질 검사 후 약 한 달에서 늦으면 세 달 뒤에 결과가 게재되는 경우도 있었다. 다시 말해 약수터를 찾은 사람들은 새로 갱신된 약수터 수질 적합 여부를 모르는 상태에서 물을 음용할 가능성이 높으며 이는 자칫 잘못하면 심각한 질병을 초래할 수 있다. 따라서 사람들의 건강을 지키고 일상 생활을 비롯한 삶의 질을 향상시키며 크고 작은 병들로 인한 사회적 비용을 최소화하는 관점에서 약수터 수질 관련 문제를 학문적으로 다루는 것은 의의가 있다.

본 연구는 앞서 언급한 약수터 수질 검사의 한계 및 여러 제약 등을 고려하여 수질 오염에 영향을 줄 수 있는 요인들의 데이터를 활용한 실시간 약수터 수질 예측 모델을 개발하고자 한다. 다양한 요인들과 약수터 수질 문제와의 연관성을 밝히는 조사연구나 개선 방안에 대한 연구는 종종 수행되어 왔지만 수질 적합 유무를 사전에 예측하여 즉각적인 정보 제공을 위한 연구는 극히 제한적이다. 따라서 본 연구는 머신러닝 기법을 활용하여 수도권에 존재하는 약수터 별로 수질 적합 유무를 분류하고 이를 예측할 수 있는 분류 모델을 제안한다. 이를 통해 사람들에게 실시간으로 약수터별 수질 적합 여부에 대한 정보를 제공함으로써 지자체를 포함한 정부와 시민들의 사회적, 경제적 손실을 줄이고 삶의 질을 향상시키는데 기여하고자 한다.

2. 이론적 배경

2.1. 약수터 수질

약수(藥水)란, 사람들에게 좋은 물에 대해 물어봤을 때 생수와 함께 흔히 언급되는 물이다. 약수는 문자 그대로 해석하면 약효를 가진 물이지만 일반적으로 인위적인 처리를 통해 가공된 물이 아닌 오염되지 않은 자연 그대로의 물을 의미다. 언제부턴가 석간수 또는 용천수를 약수라고 부르기 시작했는데 이는 자연에서 생성된 물이 건강 증진에 도움이 될 것이라는 생각에서 비롯되었다(Kim et al., 1998). 이후 약수의 개념이 보다 확장되어 사람들에게 마실 수 있는 물을 공급하기 위해 개발하거나 여과 및 침전이 잘 이루어진 지표수, 산지의 등산로 등에서 자연적으로 솟아나는 모든 샘을 지칭한다(Kim et al., 2007).

사람들이 일반적으로 생각하는 약수터는 먹는물공동시설의 한 종류이다. 먹는물 관리법 제3조 제6호에서는 여러 사람에게 먹는물을 공급할 목적으로 인위적으로 개발했거나 자연적으로 형성된 약수터, 샘터, 우물 등의 시설을 모두 포함하여 먹는물공동시설로 규정하여 관리하고 있다(Shin, 2015). 하지만 샘터와 우물 등이 거의 사라진 현재 전국적으로 약수터와 먹는물공동시설을 동일하게 생각하고 있으며 본 연구에서도 이를 참고하여 두 시설의 개념적 차이를 두지 않고 약수터라는 명칭으로 통일한다.

생활수준의 향상과 함께 사람들은 맛있고, 안전하며, 오염되지 않은 질 좋은 물을 마시려는 욕구가 증가하였다. 점점 좋은 물 즉 자연 생수를 찾아 도심 근교, 산자락 등 인적이 드문 곳에 위치한 약수터를 찾는 사람들이 늘어났으며 이와 더불어 약수터의 수질 안전성 확보가 중요한 문제로 대두되었다(Kim et al., 2010). 환경부에서 종합한 결과에 의하면 2009년 전국 약수터의 약 22.8%가 수질 부적합 판정을 받았으며, 2014년에는 먹는물 수질 기준을 초과한 약수터의 비율이 약 31.8%까지 증가하였다. 하지만 음용 부적합 판정을 받은 약수터 비율의 증가 추세로 인한 걱정이 무색하게 환경부가 발표한 2019년 자료에 의하면 전국 약수터 1일 이용인구는 약 22만 명에 육박했다.

이러한 흐름 속에서 약수터의 수질 특성 및 오염에 영향을 미치는 요인들을 파악하는 선행 연구들이 꾸준히 진행되어 왔다. Song et al.(2003)은 대구 달비약수터를 대상으로 약수터의 수질과 이에 영향을 미치는 요인을 살펴보기 위해 1년에 걸쳐 데이터를 수집하였다. 먼저 약수터의 수질 특성을 파악하기 위해 수온, pH, 전기전도도 등의 이화학적 성분과 일반세균, 총대장균군 등의 생물학적 성분을 분석 항목으로 선정하였다. 또한 약수와 더불어 빗물, 계곡수, 토양에 대한 시료 채취 및 조제를 통해 수질에 미치는 영향 요인을 살펴보았으며 강수량과 약수 수질 사이 상관성을 단순회귀분석을 통해 규명하였다. 연구 결과 약수터 수질은 기온, 강수량 그리고 이용객의 증가 등 복합적인 요인들에 의해 영향을 받고 있음을 확인하였다. 특히 강수량에 의해 크게 좌우되는데 이는 비가 내리면 쓰레기, 동물의 배설물, 자연에 존재하는 오염원 등이 강수에 섞여 약수원에 유입되기에 수질이 악화되는 것으로 판단된다. 그 외 계곡수, 빗물, 토양의 성분 또한 약수터 수질에 유의한 영향을 미치는 것을 알 수 있었다. Kim et al.(2007)은 약수터 수질 오염에 영향을 미치는 요인을 규명하기 위해 인천 관내 약수터들 중 일부를 선정하여 약 7개월에 걸쳐 데이터를 수집하였다. 이 중 수질 부적합률

이 높은 3곳을 대상으로 강수, 약수, 인근 토양을 채취하여 수질에 미치는 여러 요인을 분석하였으며, 비교적 안정된 수질을 유지하는 약수터 4곳을 대상으로 주기적인 소독 유무, 주변 지역 청결 유무 등의 관리상태 등을 살펴보았다. 또한 2006년도 수질검사 결과를 바탕으로 부적합률이 낮은 곳과 높은 곳 2곳을 추가로 선정하여 약수터 위치, 계곡수, 채수 방식, 약수터 수원 등의 영향을 분석하였다. 통계적 검정을 통한 분석 대신 다수의 시료 채취를 통해 약수터 수질 변화를 비교분석한 결과 약수의 수온, 토양의 pH, 소독 유무, 시설 주변의 오염 유무 등에 따라 수질 적합 유무의 차이가 발생함을 확인할 수 있었다. 그 외 약수터의 수원(지표수/지하수), 채수 방식, 물을 일시적으로 모아 약수의 수량을 일정하게 유지시키는 시설인 집수정의 유무 및 위치 등의 요인들 또한 수질에 영향을 미치는 것을 확인하였다. 하지만 앞서 언급한 요인들은 일부 약수터에 국한하여 자료 수집이 가능한 상황이라 약수터별 필수 요인들의 종합적인 데이터셋 구축 필요성을 제시하였다. Lee et al.(2011)은 강우량이 집중되는 특정 시기의 약수터 이용객들에게 정확한 수질 정보를 제공하기 위해 무등산에 소재한 6곳의 약수터를 대상으로 강우 기간 동안의 수질변화 특성 및 미생물 검출 현황 관련 분석을 수행하였다. 강우 전후 약 1주일 동안 매일 조사를 수행하였으며 총 4회에 걸쳐 약수의 수질 특성에 대한 데이터를 수집하였고, 광주지방기상청의 기상자료를 통해 조사기간 내 강우량과 평균기온을 수집하였다. 조사 결과 Song et al.(2003)의 연구에서 수질 오염 원인으로 기온, 강수량, 탐방객 등의 복합 요인 등을 제시한 바와 달리 무등산 약수터의 경우 월별 또는 계절별 수질의 큰 차이는 없었으며, 강우 기간 또는 집중강우에도 음용 불가능할 정도의 수질 변화는 발생하지 않았음을 알 수 있었다. Choi et al.(2018)의 연구에서는 통계 패키지의 평균비교 검정을 통해 강우 및 토양 특성에 따른 약수터 수질 오염 여부에 대한 유의성을 분석하였다. 경기도 내 349개 약수터를 대상으로 2017년 4분기 수질 부적합 현황, 경기도 강수량 자료, 시료로 채취한 토양 등을 활용한 통계분석 결과 강수량과 토양 특성은 약수터 수질 오염에 통계적으로 유의한 영향을 미치고 있음을 확인하였다.

먹는물공동시설과 같은 공공식수원의 개념을 가진 국내와 달리 국외에서는 이러한 장소적 개념이 부재하여 약수터라는 특정 장소를 대상으로 진행된 연구가 극히 제한적이었다(Nam and Zoh, 2014). 다만 유사 개념의 수질을 가질 것으로 생각되는 용천수, 지하수 등의 음용수원을 중심으로 진행된 연구 사례들(Gibson et al., 2007; Fram and Belitz, 2011; Herrero-Hernandez et al., 2013; Costa et al., 2018; Khan et al., 2021)을 살펴본 결과 국내 연구가 밝힌 오염 인자들과 큰 차이가 없음을 확인할 수 있었다.

이렇듯 약수터 수질은 강수량, 토양, 빗물, 등산객, 주변 오염물질, 애완견이나 야생동물의 배설물, 약수터의 입지 요건 등 다양한 요인에 영향을 받는다고 볼 수 있다(Park et al., 2021). <Table 1>은 앞서 살펴본 연구들을 제외한 선행연구들에서 약수터 수질에 영향을 미치는 요인들을 분석한 결과를 연구 별로 정리해놓은 표이다.

약수의 위생학적 측면에 대한 연구들 외에도 미생물학적 안정성 평가, 균종별 분포 및 생화학적 조사, 약수와 건강의 관계, 물 맛 평가, 이용실태 및 만족도 등 약수터를 이용하는 시민들의 의식 조사 등 다양한 연구들이 수행되어 왔다(Moon and Park, 1998; Ryu, 2005; Woo, 2008;

〈Table 1〉 Factors of water quality contaminations of Mineral springs

Studies	Water sources	Sampling sites	Contamination factors
Han and Park(1997)	Spring water	Daejeon, Chungnam	Surrounding facilities(glass factory)
Lee et al.(2002)	Spring water	Daegu, Gyeongbuk	Heavy rain
Lee(2002)	Ground water	Hwasun, Jeonnam	Surrounding facilities(mine area)
Yang et al.(2006)	Spring water	Jeonnam	Precipitation, Season
Lee et al.(2011)	Spring water	Gwangju, Jeonnam	Heavy rain
Hyun(2011)	Ground water	Jeju island	Manure in livestock
Ok et al.(2011)	Ground water	Gangwon	Wastewater discharge

Kim et al., 2011; Yoon et al., 2013; Song et al., 2019). 하지만 약수의 수질과 관련한 연구들은 전반적으로 일정 기간 시료 채취를 통해 수질에 미치는 요인들과 약수의 상관성을 도출하고 오염원을 규명하는 사후 분석 형식에 그쳤음을 알 수 있다. 또한 약수의 오염에 미치는 다양한 요인들을 고려해볼 때 실시간으로 관리가 이루어지지 않는 이상 현재 각 시별로 민관의 협력 하에 수행되는 먹는물공동시설의 수질 검사만으로 매일 약수터를 찾는 사람들에게 약수의 안전성을 주장하기에는 신빙성이 부족하다. 따라서 약수터 수질에 영향을 미치는 다양한 요인들의 실시간 데이터를 활용 및 조합하여 약수터 수질 적합 여부를 예측할 필요성이 제기된다.

2.2. 머신러닝을 활용한 예측 모델

머신러닝 기법은 학습 데이터의 결과 값(label) 유무에 따라 크게 지도학습(supervised)과 비지도학습(unsupervised)으로 구분된다. 지도학습은 주어진 데이터와 레이블을 이용하여 값을 예측하는 방법으로 분류(classification)와 예측(prediction)에 주로 활용되며 비지도학습은 군집(clustering), 차원 축소 등의 문제에 적용 가능하다(Kim et al., 2019). 본 연구는 수질검사 결과 값(적합/부적합)을 예측할 수 있는 여러 지도학습 모델들의 성능을 비교 평가하였다. 머신러닝 기법을 활용한 예측모델에 관한 선행연구는 의학, 심리, 금융, 교육, 공학 등 다양한 영역에서 수행되어 왔다. 수질 예측에 관한 연구 또한 간간이 이루어져 왔는데, Sattari et al.(2014)은 지표수가 부족한 이란에서 농업의 지속가능한 발전을 위해 수질 오염을 방지할 수 있는 적절한 방법을 찾고자 하였다. 하지만 지표수 수질에 대한 직접적인 모니터링 및 평가는 비용과 시간이 많이 소모되어 최소한의 수화학 매개변수들을 활용하여 수질 등급을 예측할 수 있는 모형을 고안하고자 하였다. 이들은 의사결정나무 분류기법을 활용한 예측 모델을 제안하였으며, 적은 수의 매개변수를 입력변인으로 투입한 결과 높은 정확도로 예측 분류가 이루어짐을 확인하였다. Ahmed et al.(2019)은 각종 개발 및 사람들의 활동으로 인해 오염된 말레이시아를 관통하는 Johor 강의 수질 개선을 위해 수질 예측 모델의 필요성을 제안하였다. 이들은 다층퍼셉트론 신경망 기법을 활용하여 예측 모델을 만든 후 다양한 수질 매개변

수들을 입력변인으로 투입한 결과 마찬가지로 높은 성능을 보이는 것을 확인하였다. 이 외에도 국내외에서 시계열, 앙상블 등의 기법을 활용한 예측 연구가 수행되었다(Faruk, 2010; Stidson et al., 2012; Muniruzzaman and Pedretti, 2021).

Kim and Jung(2021)은 약 6년간 동작구에서 발생한 범죄데이터를 바탕으로 머신러닝 기법을 활용한 범죄예측모델을 구축해보고자 하였다. 먼저 전통 이론 및 선행연구들을 바탕으로 범죄를 예측하는 데 영향을 미칠 수 있는 변수들을 도출하였다. 추가로 시·공간적 요소를 함께 고려하여 예측을 진행하였으며 Decision Tree, Random Forest, SVM, K-NN 등 총 4가지 머신러닝 모델의 성능을 비교하였다. 특히 범죄 데이터는 특정 클래스에 분포가 치중되어 있는 불균형 데이터이다. 따라서 데이터의 불균형을 해결하기 위한 리샘플링 기법의 적용 전후 모델의 예측력을 Confusion Matrix를 통해 비교 분석하였다. 분석 결과 리샘플링이 이루어진 데이터를 활용한 모델의 예측 정확도가 상대적으로 높게 나왔으며, 4가지 모델 중 Random Forest 모델의 성능이 가장 우수한 것을 확인할 수 있었다. Kim et al.(2020)은 한국조세재정연구원의 재정패널데이터에서 생존분석, OLS 기법 등을 활용하여 주택 보유기간에 영향을 미치는 결정요인을 선별한 후 다양한 머신러닝 모델들의 예측력을 비교하였다. 사용한 모델은 SVM, Decision Tree, Random Forest, Gradient Boosting, XGBoost, LightGBM 등 총 6가지이며, RMSE(Root Mean Squared Error, 평균 제곱근 오차)를 기준으로 Random Forest가 가장 높은 예측력을 가지는 것을 알 수 있었다. Shin and Kwahk(2018)은 부실화 가능성이 높은 기업들에 대한 경보체계를 의미하는 관리종목 지정 예측 모델을 개발 및 검증하였다. 코스닥(KOSDAQ) 상장 기업들 중 관리종목으로 지정된 기업과 그렇지 않은 기업들을 표본으로 Logistic Regression과 Decision Tree 기법을 활용하여 관리종목 탐지 모델을 개발하였다. Confusion Matrix를 바탕으로 예측 정확도를 비교 분석한 결과 두 모델의 예측률은 거의 유사하였으나 Decision Tree 기반의 모델이 조금 더 나은 성능을 보이는 것을 확인할 수 있었다. Eom et al.(2020)은 기업의 부도위험을 예측하기 위해 Random Forest, Multiple Layers Perceptron, Convolution Neural Network 등 세 분석기법을 조합한 Stacking Ensemble 모델을 활용하였다. MAE, MSE, RMSE 세 평가지표를 기준으로 예측 성능을 비교해본 결과 Stacking Ensemble 모델이 단일 분석모델 대비 오차가 가장 낮은 것을 알 수 있었다. Lee(2021)는 서울 시내 5성급 호텔 20곳을 대상으로 머신러닝 기법을 활용한 호텔 부실화 예측 모델을 개발하고자 하였다. 선행연구들을 참고하여 선정한 14개의 지표 중 t-test를 통해 건전호텔과 부실호텔을 구분할 수 있는 요인을 분별하였다. 최종 10개 변수를 토대로 MLP, SVM, Decision Tree 기법을 활용한 예측 모델을 생성하고 정확도를 비교하였다. 정확도 지표 중 하나인 AUC(area under ROC curve)를 기준으로 세 모델의 예측률을 살펴본 결과 인공신경망 기법 중 하나인 MLP 기반의 모델이 아주 높은 예측 정확도를 가지는 것을 확인할 수 있었다. Lee and Lee(2021)는 약 3년간 서울시에서 발생한 보행자 교통사고 자료 및 머신러닝 기법들을 활용하여 보행자 교통사고 심각도 예측모델을 개발하고자 하였다. 이 과정에서 추가로 보행자 교통사고 수준의 심각도별 사고 수의 불균형 문제를 해결하기 위해 리샘플링 기법 중 하나인 오버샘플링을 진행하였다. 모델의 성능 평가 지표로는 Kubat et al.(1997)

이 제안한 불균형 데이터에서 사용할 수 있는 G-mean을 사용하였으며 Logistic Regression, SVM, Random Forest, Gradient Boosting, MLP 등 총 5가지 모델의 성능을 비교 분석하였다. 분석 결과 오버샘플링 기법을 적용한 경우 소수 표본 유형의 예측 정확도가 상대적으로 증가하였지만 전반적으로 모든 모델의 정확도가 낮게 도출되었다. 그나마 5개의 모델 중 MLP가 가장 좋은 성능을 보이고 있음을 확인하였다. 이러한 선행연구들을 바탕으로 본 연구에서는 약수터 수질 예측 모델을 개발하고자 주요 머신러닝 기법들을 활용한 예측 모델들의 정확도를 비교, 분석하여 가장 우수한 성능을 가진 모델이 무엇인지 살펴보고자 한다.

3. 연구방법

3.1. 표본 약수터 선정

본 연구는 약수터의 수질 검사 데이터를 활용하여 약수터 별 수질 예측 모델을 구축하고자 한다. 기간은 각 지역별 수질 검사 결과 데이터가 어느정도 갖춰져 있는 2015년부터 2020년까지로 정했으며, 지역은 데이터 수집의 시간적 한계 상 수도권 즉 서울과 경기도(전체 31개 시 중 약수터 수질 검사 결과 데이터가 있는 18개 시 선별)로 한정했다. 서울시와 경기도는 각각 '우리동네 약수터'와 '먹는물공동시설' 서비스를 통해 사람들에게 수질 검사 현황을 제공하고 있다. 수도권에 위치하는 약수터는 약 600여개이지만 이 중 제대로 된 측정이 이루어지지 않은 곳들을 제외한 총 314개의 약수터를 분석에 사용할 표본으로 선정 후 데이터를 수집하였다.

3.2. 변수 선정 및 데이터 수집

2021년 2월 서울시 보건환경연구원은 근 3년간 수행해온 약수터(먹는물공동시설) 수질검사 결과 수질 부적합률이 높게 나오는 원인으로 기후온난화, 약수터 샘의 얕은 깊이로 인한 빗물의 유입, 이용객 증가 및 조류/야생동물/애완견의 분뇨, 대기오염 물질을 제시하였다. 이와 더불어 앞서 살펴본 선행연구들을 참고하여 약수터 수질에 영향을 미칠 수 있는 변수들을 살펴본 결과 크게 기상, 토양, 계절, 부적합 판정기록, 환경 및 주변시설 등의 요인들로 정리할 수 있었다.

보다 구체적으로 살펴보면 기상 요인으로는 기온, 강수량, 미세먼지, 초미세먼지 변인을 선별하였으며, 토양 요인으로는 토양의 산성도, 유기물 함량, 토양 내 물이 빠져나가는 정도를 나타내는 토양 배수 변인을 선별하였다. 특히 여름이라는 특정 시기에 여러 위생 문제가 많이 발생하는 점을 고려하여 계절과 분기 변인 또한 예측변인군에 포함시켰으며, 직전 검사에서의 수질 부적합 판정 기록 또한 영향을 미칠 것으로 여겨 선별하였다. 마지막으로 환경 및 주변시설 요인으로는 약수터의 고도, 이용객 수, 애완동물 출입 가능 여부, 주변에 가축 축사 또는 공장지대가 존재하는지 유무 그리고 해당 약수터가 그린벨트 즉 개발제한구역에 속하는지 유무를 고려하였다.

1차적으로 선별된 약 20가지의 예측변인 중 데이터의 양과 질을 고려한 가용성, 수집 가능성 등을 고려하여 다시 한 번 예측변인을 선별하였다. 국가통계포털, 공공데이터포털 및 각종 기관에서 제공하는 데이터 허브 등을 살펴보았음에도 불구하고 데이터 자체가 존재하지 않아서 수집이 불가능한 변인, 중간중간 측정이 제대로 이루어지지 않은 변인, 서울과 경기도 두 곳 중 한

〈Table 2〉 Data collection by variable

Variable		Source
Weather	Temperature, Precipitation	Korea Meteorological Administration
	Fine dust, Fine particulate matter	Open Data Plaza, Gyeonggi Data Dream, Airkorea
Environment	Greenbelt	Korea National Spatial Data Infrastructure Portal
	Altitude	GoogleEarth
	Visitors	Each District office(Seoul), Soil Groundwater Information System(Gyeonggi)
Season	Season, Semester	
Fail record		

곳만 데이터가 존재하는 변인 등은 제외하였다. 2차적으로 선별된 예측변인은 기온, 강수량, 미세먼지, 초미세먼지, 계절과 분기, 과거 수질 부적합 판정 기록, 약수터 고도, 이용객 수, 약수터의 그린벨트 소속 유무 등 총 10가지이다. 약수터 관련 데이터는 서울시 각 구청 홈페이지, 경기토양지하수종합정보시스템에서 수집하였으며 그 외 변인들의 데이터 수집 경로는 <Table 2>에 정리하였다.

3.3. 데이터 전처리

본 연구에서 수집한 데이터셋에서는 별도의 결측치 및 이상치는 없었으며, 수질 검사일 기준 직전 3일의 평균 기온, 평균 강수량, 평균 미세먼지, 평균 초미세먼지 수치 또한 검사 당일 수질 검사 결과에 영향을 미칠 것으로 판단되어 추가하였다. 계절과 분기 및 과거 수질 부적합 판정 기록은 각 지자체 홈페이지에서 수집한 수질 검사 결과 파일의 검사일을 기준으로 엑셀의 함수를 활용해 변수화 시켰다. 계절, 분기 요인은 수도권 전반적으로 수질 검사 결과가 1분기, 2분기, 7월, 8월, 9월, 4분기로 나누어져 진행되어 왔기에 계절 구분이 명확하지 않아 둘 중 분기 변인을 활용하는 것에 초점을 맞췄다. 다만 3분기는 강우량이 집중되는 시기여서 그런지 월별로 수질 검사가 진행되었기에 본 연구에서도 6개로 구분하였으며 원-핫 인코딩(one-hot encoding)을 통해 6개의 변수로 더미화 시켰다. 분석에 사용한 예측변인들은 <Table 3>에 정리하였다.

3.4. SHAP

일반적인 회귀분석을 수행할 경우 표준화 회귀계수를 통해 종속변수에 미치는 독립변수들의 영향 정도를 비교할 수 있다. 마찬가지로 머신러닝에서도 분석에 사용된 변인들의 중요도를 파악할 수 있는 방법들이 존재한다. 본 연구에서는 약수터의 수질을 예측하는 데 있어 각 요인들이 어느 정도 영향을 미치는지, 중요한 요인들이 무엇인지 확인하기 위하여 SHAP(SHapley Additive exPlanations) 기법을 활용하였다.

SHAP 기법은 '설명 가능한 인공지능'이란 의미를 가진 XAI(eXplainable Artificial Intelligence) 기법 중 하나로 전체 성과의 창출에 있어 각 변인의 기여도를 수치로 표현해주는 Shapley Value

〈Table 3〉 Input Variables

Type	Variables	Information
Continuous Variable	temp	Temperature on the day of the test
	temp_m	The average temperature for 3 days right before the test day
	precip	Precipitation on the day of the test
	precip_m	The average precipitation for 3 days right before the test day
	fd	Fine dust on the day of the test
	fd_m	The average fine dust for 3 days right before the test day
	fpm	Fine particulate matter on the day of the test
	fpm_m	The average fine particulate matter for 3 days right before the test day
	visitor	Monthly average visitor
	altitude	Altitude of the mineral spring
Binary Variable	1q	Test day belongs to the first quarter
	2q	Test day belongs to the Second quarter
	3q_7	Test day belongs to the July
	3q_8	Test day belongs to the August
	3q_9	Test day belongs to the September
	4q	Test day belongs to the fourth quarter
	greenbelt	Whether the mineral spring belongs to the green belt or not
	fail_p1	Whether the result of right before the test is fail or not
	fail_p2	Whether the result of the test was fail twice in a row in the past or not

를 이용한다. 여기서 특정 변인의 기여도는 모든 변인이 포함되었을 경우 도출되는 성과에서 각 변인이 제외된 경우 도출되는 성과의 차이를 모두 계산 후 가중평균을 적용시킨 수치이다(Lundberg et al., 2018). 달리 표현하면 예측 모델의 출력 결과를 분석에 사용된 변인들의 기여도로 분해가 가능하며 나아가 각 변인들 간 영향을 주고 받을 가능성을 의미하는 의존도까지 고려하여 종속변수 예측에 미치는 평균 영향 수치를 제공함으로 다른 기법들에 비해 보다 정확한 영향력을 제시해 준다(Ahn, 2020).

3.5. 분석 모델 설계

전처리 과정을 거친 데이터셋의 70%를 학습용으로, 30%를 시험용으로 활용하기 위한 분류 과정을 거친 후 모델화 단계를 수행하였다. 최적의 분류 예측 모델을 개발하기 위해 본 연구에서는 Logistic Regression, SVM, Random Forest, XGBoost, LightGBM 등을 포함한 약 20개의 회귀 및 분류 모델의 성능 비교 결과를 제시해주는 Python의 오픈소스 라이브러리인 Pycaret을 활용하였다. Pycaret은 최근 주목받고 있는 자동화 머신

〈Table 4〉 Top5 Model performance offered by Pycaret

Machine learning models		Training set		Test set	
		AUC	Accuracy	AUC	Accuracy
catboost	CatBoost Classifier	0.8126	74.97%	**0.8174**	**75.26%**
lightgbm	Light Gradient Boosting Machine	**0.8111**	**75.63%**	0.8143	74.91%
xgboost	Extreme Gradient Boosting	0.8058	74.79%	0.8114	74.74%
gbc	Gradient Boosting Classifier	0.7994	75.08%	0.8114	75.09%
rf	Random Forest	0.7989	74.13%	0.7919	74.23%

러닝 기술인 AutoML(automated machine learning) 라이브러리 중 하나이며, AutoML은 2017년부터 Google이 추진해온 프로그램으로 시간 소모적이고 반복적인 머신러닝 모델의 개발 작업을 자동화하는 프로세스이다(Lee et al., 2021).

본 연구에서 수집한 수질 검사 결과 데이터에서 수질 부적합 판정을 받은 데이터의 비율은 약 32%에 해당한다. 일반적으로 현실에서 얻을 수 있는 데이터셋은 대체로 클래스 불균형 상태이다(Kim and Kwahk, 2022). 본 데이터셋은 특정 범주에 과도하게 편향되지는 않았지만 그럼에도 데이터는 불균형 상태라고 볼 수 있다. 이를 고려하여 모델의 성능은 최적의 분류모델을 선정하는 일반적 기준인 AUC(area under the curve)를 평가지표로 비교하였다. AUC란 ROC curve의 아래 면적을 나타내며 1에 가까울수록 완벽한 모델로 판단할 수 있다.

4. 실증분석 결과

4.1. 분석 결과

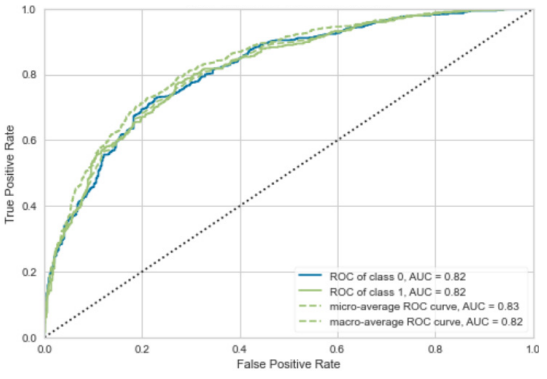
본 연구는 데이터 수집 과정에서 최종 선정된 10가지 변인이 약 300여개의 서울 및 경기지역 약수터 수질에 미치는 영향을 머신러닝 기법을 활용하여 살펴보았다. 먼저, 학습용 데이터셋을 대상으로 Pycaret이 제공하는 각 모델들의 성능을 비교해보았다. 현존하는 다양한 머신러닝 분석기법들을 일일이 비교하기에는 시간적으로 비효율적이다. 하지만 Pycaret 라이브러리는 20여 가지의 머신러닝 기법들 중 예측 성능을 기준으로 5개의 상위 모델을 정리해서 제시해준다. 이를 바탕으로 시험용 데이터셋을 사용한 5가지 분류 예측 모델의 성능을 <Table 4>에 정리하였다.

구체적으로 살펴보면 학습용 데이터셋의 경우 LightGBM이 AUC가 0.8111, 75.63%의 예측율로 가장 높은 성능을 보였으며 Random Forest 모델은 AUC가 0.7989, 예측율은 74.13%로 상위 5개의 모델 중 상대적으로 성능이 가장 낮은 것을 확인하였다. 시험용 데이터셋의 경우 학습용 데이터셋의 결과와 달리 CatBoost 모델이 AUC가 0.8174, 정확도가 75.26%로 가장 좋은 성능을 보이는 것을 알 수 있다. 학습용, 시험용 데이터셋 모두 Boosting 계열의 모델이 예측 성능 기준 상위 5가지 모델 중 다수를 차지하고 있는 것을 <Table 4>에서 확인할 수 있다. AUC가 1에 가깝게 높을수록 수질 부적합을 부적합으로, 적합을 적합으로 예측이 잘한다는 것을 의미한다. <Table 4>를

보면 전반적으로 예측 모델들의 AUC가 0.8 이상으로 클래스 분류 능력이 뛰어나다고 볼 수 있다. 시험용 데이터셋에서 가장 성능이 좋은 CatBoost 모델의 혼동행렬과 ROC curve를 <Table 5>와 <Figure 1>에 제시하였다.

〈Table 5〉 Confusion Matrix of the CatBoost Classifier Model

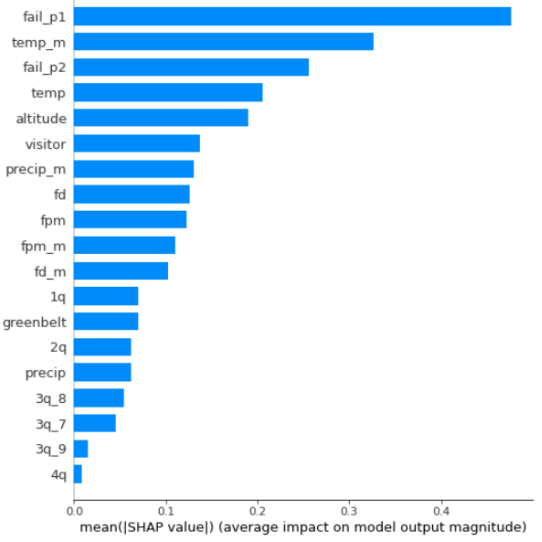
Observation \ Prediction	Test Set	
	0 (Non-fail)	1 (Fail)
0 (Non-fail)	617	115
1 (Fail)	176	268



〈Figure 1〉 ROC Curves for CatBoost

4.2. SHAP Summary Plot

분석에 사용된 변인들이 예측에 미치는 영향을 살펴보기 위해 SHAP 기법을 활용하였다. <Figure 2>는 앞서 가장 성능이 좋았던 CatBoost 분류 예측 모델에서 도출해낸 SHAP 값을 시각화한 막대 플롯이다. 이를 통해 각 변인들이 예측에 미치는 절대 영향도를 파악할 수 있다. 영향요인의 중요도 순으로 결과를 살펴보면 직전 수질 검사에서 부적합 판정을 받았는지 여부가 가장 영향력이 컸으며 다음으로 검사일 직전 3일 간 평균 기온, 과거 연속 2번 수질 부적합 판정 기록 유무, 수질 검사 당일 기온, 고도가 상위 5가지 요인임을 알 수 있다. 다음으로 약수터 방문객, 검사일 직전 3일 간 평균 강우량, 미세먼지 및 초미세먼지 등도 상위 5개 요인만큼은 아니지만 결과에 어느정도 영향을 미치고 있음을 확인할 수 있었다. 이 외에 약수터가 위치한 장소가 그린벨트에 속하는지 유무와 더불어 6개의 분기 요인들 또한 약수터 수질에 영향을 미치는 것을 알 수 있었다.



〈Figure 2〉 Bar plot of SHAP value for CatBoost (average impact on model output)

5. 결론 및 시사점

본 연구는 2015년부터 2020년까지 수도권에 위치하는 약수터들을 대상으로 수질 검사 결과

에 대한 예측 모델을 개발하고 성능을 확인해보았다. 본 연구에서는 분류 예측에 사용되는 다양한 머신러닝 기법들을 한 번에 비교 분석해주는 Python의 Pycaret 라이브러리를 활용하여 최적의 예측 모델을 선정하였다.

AUC 평가지표 기준 상위 5가지 모델로 CatBoost, LightGBM, XGBoost, GBC, Random Forest가 도출되었으며 가장 좋은 성능을 가진 모델은 75.26%의 예측 정확도를 가진 CatBoost 분류기임을 확인할 수 있었다. 또한 CatBoost 모델을 사용하여 분류 예측을 수행한 경우 각 요인들이 결과에 어느정도 수준의 영향력을 미치는지 확인하기 위하여 SHAP 기법을 수행하였다. 분석 결과 수질 예측에 가장 큰 영향력을 미치는 요인은 직전 검사에서의 부적합 판정 유무이며, 기온과 고도 방문객 등 또한 적지 않은 영향을 미치고 있음을 확인하였다. 하지만 선행연구들을 통해 수질 예측에 영향을 미칠 것으로 고려했던 요인들 중 분기 요인은 상대적으로 적은 영향을 미치고 있음을 알 수 있었다. 기존 문헌에서 알려진 바와 같이 기상 요인들과 약수터 방문객은 약수터 수질에 영향을 미치는 중요 요인으로 확인되었다. 하지만 기온, 강수량 등 단순 기상 요인이 아니라 특히 검사일 직전 3일 간의 평균 기온이 수질 검사 당일의 기온 대비 약수터 수질에 미치는 영향이 더 큰 것을 알 수 있으며, 평균 강수량도 당일 강수량 대비 마찬가지 결과를 보이는 것을 알 수 있었다. 또한 상대적으로 낮은 영향력을 가지고 있긴 하지만 약수터가 위치한 지역이 그린벨트 유무인지에 대한 부분도 약수터 수질에 영향을 미치는 요인임을 확인하였다.

본 연구의 시사점은 다음과 같다. 본 연구는 2017년 구글이 처음 제시한 이후 점점 발전함과 동시에 주목받고 있는 자동화 머신러닝 기법인 AutoML을 약수터 수질 예측 분류에 처음으로 접목시켰다. 또한 약수터 수질은 약수터의 입지 요건 및 주변 환경에 영향을 받는 다는 기존 문헌들의 내용을 기반으로 여러 기상 요인, 탐방객 수 등 기존에 잘 알려진 인자들 외에 약수터 고도, 그린벨트 유무, 애완동물 출입가능 여부, 강수 등의 기상요인이 수질에 영향을 미칠 시간을 고려한 검사일 기준 직전 3일의 평균 수치 등을 새로이 분석에 활용한 부분 또한 기존 연구 대비 발전된 부분이라고 할 수 있다. 선행연구들이 밝혀낸 요인과 함께 다양한 파생변수들을 활용한 본 연구의 예측 모델은 약 75%의 정확도를 보이고 있다. 향후 데이터 관리가 원활히 이루어져서 서울과 경기지역 약수터 모두가 동일하게 가지고 있는 데이터가 아니라 탈락한 변인들까지 고려한다면 더욱 향상된 성능을 얻을 가능성이 높다.

실무적 관점에서의 시사는 다음과 같다. 첫째, 공공안전의 관점에서 본 연구는 효과적인 기여를 하였다. 약수터는 '약수'터란 이름에 걸맞게 깨끗하고 몸에 좋은 물을 마실 수 있다는 생각에 여전히 찾는 이들이 많다. 하지만 주요 식수원으로 쓰이는 약수터가 산업활동과 폐기물 처리로 오염된 사례는 최근까지 매해 지속되어 왔다. 2019년 파주시의 한 약수터는 인근 반도체 공장 폐수로 오염되어 수질이 저하되고 이를 이용하는 지역 주민들의 건강에 위험이 제기되었으며, 2020년 천안시의 한 약수터는 인근 제지공장 폐수로 오염되어 지역 주민과 야생동물 건강을 위협하였고, 2021년 아산시 약수터 또한 인근 화학공장 폐수의 영향을 받아 식수를 약수에 의존하는 주민들의 건강에 위험이 제기되었다. 대규모의 피해를 입힌 구체적인 사례는 아직 확인된 바 없지만 십여 년이 넘도록 인근 주민의 건강을 위협할 수 있는 신호가 지속되고 있다는 것은 결코

가볍게 넘어갈 부분이 아니다. 사건 발생 전 위험요소를 사전에 방지할 수 있다면 통제의 사각지대에서 발생하는 크고 작은 문제들 또한 예방할 수 있을 것이다. 둘째, 범용성 측면에서도 본 연구는 기여를 하고 있다. 본 연구는 약수터의 수질 문제에 집중하여 분석을 진행하였지만 비단 약수터 외에도 강이나 호수 등 다른 곳의 수질 관리에도 활용될 수 있다. 본 연구에서 약수터 수질에 영향을 미칠 수 있는 요인으로 선정된 변인들을 기반으로 분석을 원하는 환경, 지역 등의 요소 등을 고려한 추가 요인들을 고려한다면 보다 수월한 분석이 이루어질 수 있을 것이다. 마지막으로 본 연구의 실시간 수질 예측 모델 개발 시도는 일정한 시간 간격을 두고 이루어지는 약수터 수질 검사일 사이에 약수를 음용하는 시민들에게 실시간 약수터 수질 정보를 제공할 수 있는 서비스의 첫 걸음이 될 수 있다. 연구를 진행하는 과정에서 수질에 중요한 영향을 미칠 것으로 판단되었지만 데이터가 부재하거나 꾸준한 관리가 이루어지지 않아서 분석에 활용할 수 없었던 데이터 관리 측면에서의 부족한 부분이 존재하였다. 또한 직전 약수터 수질검사 결과가 약수터 수질 예측에 영향을 미치는 가장 중요한 요인 중 한 가지로 나온 부분도 연구의 한계이다. 서울과 경기 지역의 약수터가 타지역과 비교하여 상대적으로 주기적인 관리가 이루어지고 있지만 지역구마다 관심 및 관리의 차이가 존재하기 때문에 약수터마다 수질검사 시행 주기 및 결과 제공 기간에서의 차이가 존재한다. 하지만 약수터에 대한 수요가 꾸준히 이어져 왔으며 약수에 대한 위험 경보 또한 지속적으로 있다는 것은 큰 피해가 발생하기 전에 관심의 사각지대에 놓인 약수터에 구조적인 부분에서의 관리가 필요하다는 것을 의미한다. 주기적인 수질 검사와 함께 미흡했던 데이터 관리 측면이 개선된다면 수질 예측에 유의한 영향을 미칠 수 있는 다양한 변인들을 활용하여 보다 정확한 예측 결과를 얻을 수 있을 것이다. 나아가 정확한 정보를 제공받지 못해 안전하지 않은 약수를 음용함으로 발생할 수 있는 시민들의 건강상 문제 등의 사회적 비용도 점차 감소할 수 있을 것으로 기대된다.

참고문헌(References)

Ahmed, A. N., F. B. Othman, H. A. Afan, R. K. Ibrahim, C. M. Fai, M. S. Hossain, ... and A. Elshafie, “Machine learning methods for better water quality prediction,” *Journal of Hydrology*, (2019), 578, 124084.

Ahn, J. H. *XAI Explainable AI, dissect artificial intelligence*, Wikibooks, 2020.

Choi, P., P. Heo, K. Lee, D. Cho, C. Kim, and T. Kim, “Study on Water Quality Improvement in Public Drinking Water Facilities in Gyeonggi-do,” *Journal of the Korean Society for Environmental Analysis*, Vol.21, No.3(2018), 148~153.

Costa, D. D., A. A. Gomes, M. Fernandes, R. L. da Costa Bortoluzzi, M. D. L. B. Magalhães, and E. Skoronski, “Using natural biomass microorganisms for drinking water denitrification,” Journal of Environmental Management, Vol.217, (2018), 520~530.

Eom, H. N., J. S. Kim, and S. O. Choi, “Machine learning-based corporate default risk prediction model verification and policy recommendation: Focusing on improvement through stacking ensemble model,” *Journal of Intelligence and Information Systems*, Vol.26, No.2(2020), 105~129.

Faruk, D. Ö., "A hybrid neural network and ARIMA model for water quality time series prediction," *Engineering applications of artificial intelligence*, Vol.23, No.4(2010), 586~594.

Fram, M. S. and K. Belitz, "Occurrence and concentrations of pharmaceutical compounds in groundwater used for public drinking-water supply in California," *Science of the Total Environment*, Vol.409, No.18(2011), 3409~3417.

Gibson, R., E. Becerril-Bravo, V. Silva-Castro, and B. Jimenez, "Determination of acidic pharmaceuticals and potential endocrine disrupting compounds in wastewaters and spring waters by selective elution and analysis by gas chromatography-mass spectrometry," *Journal of chromatography*, Vol.1169, No.1~2 (2007), 31~39.

Han, W. W. and D. H. Park, "A Study on the Characteristics of Ground Water Quality in Taejon (I)," *DaeJeon University the Institute of Environmental Studies*, Vol.2, (1997), 17~28.

Herrero-Hernandez, E., M. S. Andrades, A. Alvarez-Martin, E. Pose-Juan, M. S. Rodriguez-Cruz, and M. J. Sanchez-Martin, "Occurrence of pesticides and some of their degradation products in waters in a Spanish wine region," *Journal of hydrology*, Vol.486, (2013), 234~245.

Hyun, G. T., "Studies on the contamination properties of soil and groudwater in a densely populated livestock area in Jeju island," *Doctoral Dissertation*, Jeju national university, 2011.

Khan, A., A. Khan, F. A. Khan, L. A. Shah, A. U. Rauf, Y. I. Badrashi, W. Khan, and J. Khan, "Assessment of the Impacts of Terrestrial Determinants on Surface Water Quality at Multiple Spatial Scales," *Polish Journal of Environmental Studies*, Vol.30, No.3(2021), 2137~2147.

Kim, C. S., N. K. Kim, and K. Y. Kwahk, "Research Trends Analysis of Machine Learning and Deep Learning: Focused on the Topic Modeling," *Journal of the Korea Society of Digital Industry and Information Management*, Vol.15, No.2(2019), 19~28.

Kim, D. Y. and S. W. Jung, "Comparison of Crime Forecasting Models based on Spatio-Temporal Data and Machine Learning," *Journal of the Architectural Institute of Korea*, Vol.37, No.1 (2021), 135~143.

Kim, E. M., S. B. Kim, and E. S. Cho, "Using Mechanical Learning Analysis of Determinants of Housing Sales and Establishment of Forecasting Model," *Journal of Cadastre & Land InformatiX*, Vol.50, No.1(2020), 181~200.

Kim, I., H. Ha, W. Seo, J. Bae, H. Moon, C. Park, E. Oh, S. Kim, and M. Kim, "A Study of Water Quality Characteristic of Natural Mineral Water - In Chonnam Area -," *Korean Journal of Environmental Health*, Vol.24, No.1(1998), 87~97.

Kim, J. H. and K-. Y. Kwahk, "Class Imbalance Resolution Method and Classification Algorithm Suggesting Based on Dataset Type Segmentation," *Journal of Intelligence and Information Systems*, Vol.28, No.3(2022), 23~43.

Kim, K., B. Lee, O. Kim, M. Hur, K. Kim, J. Ro, C. Choe, J. Go, and Y. Kim, "A Study on pollution of spring in Incheon Area," *Korean Journal of Sanitation*, Vol.22, No.3(2007), 35~50.

Kim, K., H. Gil, H. Kim, B. Roh, J. Hong, J. Lee, J. Kim, M. Lee, S. Eom, and J. Lee, "Study on Water Quality of Spring Water in Seoul," *Journal of soil and groundwater environment*,

Vol.15, No.6(2010), 99~106.

Kim, K., H. Gil, M. Lee, S. Eom, and J. Lee, "Survey of Citizens Public Opinion for Natural Spring Water in Seoul," *Journal of soil and groundwater environment*, Vol.16, No.2(2011), 1~5.

Kubat, M., R. Holte, and S. Matwin, "Learning when negative examples abound," *In European conference on machine learning*, Springer, Berlin, Heidelberg, (1997), 146~153.

Lee, D. J., J. Kang, and K. Chung, "Data Processing of AutoML-based Classification Models for improving Performance in Unbalanced Classes," *Journal of Convergence for Information Technology*, Vol.11, No.6(2021), 49~54.

Lee, G. T., "A study on the development of a predictive model of hotel financial distress by machine learning algorithm," *International Journal of Tourism and Hospitality Research*, Vol.35, No.1(2021), 59~71.

Lee, H. H., "Arsenic distribution characteristics of surface water and groundwater in southern Hwasun region," *Doctoral Dissertation*, Chonnam national university, 2002.

Lee, H. J. and S. G. Lee, "Comparative Analysis of Machine Learning Models for the Prediction of Pedestrian Crash Severity: Focused on Balancing Pedestrian Crash Dataset," *Journal of Korean Society for Geospatial Information Science*, Vol.29, No.2(2021), 3~15.

Lee, S., H. Song, C. Cho, Y. Lee, S. Lee, H. Jeon, D. Jung, and W. Jang, "The Characterization of the Rainfall Effects on the Chemical and Microbiological Mineral Water Quality (in Daegu Area)," *Journal of Korean Society of Environmental Engineers*, Vol. 24, No.12(2002), 2213~2225.

Lee, Y., O. Park, S. An, Y. Kim, J. Kim, S. Bae, K. Paik, and Y. Moon, "Quality of Spring Water Influenced by Rainfall in Mudeung Mountain," *Journal of the Korea Society for Environmental Analysis*, Vol.14, No.3(2011), 146~157.

Lundberg, S. M., G. G. Erion, and S. I. Lee, "Consistent individualized feature attribution for tree ensembles," University of Washington, 2018.

Moon, H. and K. H. Park, "Mineral Characteristics of Spring Water ill Chonnam," *Korean journal of food science and technology*, Vol.30, No.2(1998), 253~259.

Muniruzzaman, M., and D. Pedretti, "Mechanistic models supporting uncertainty quantification of water quality predictions in heterogeneous mining waste rocks: a review," *Stochastic Environmental Research and Risk Assessment*, Vol.35, No.5(2021), 985~1001.

Nam, S. W. and K. D. Zoh, "A Study on Characteristics of Contamination and Target compounds for Water Quality in Public Spring Waters, Korea," *The Korean journal of public health*, Vol.51, No.1(2014), 55~66.

Ok, Y., S. Kim, K. Kim, S. Lee, D. Moon, K. Lim, J. Sung, S. Hur, and J. Yang, "Monitoring of selected veterinary antibiotics in environmental compartments near a composting facility in Gangwon Province, Korea" *Environmental monitoring and assessment*, (2011), 693-701.

Park, J., S. Kim, Y. Lee, N. Kim, Y. Kang, S. Bae, and J. Kim, "Evaluation of Characteristics of Microorganisms Isolated from Public Drinking Water Facilities in Gwangju City," *Journal of Environmental Health Sciences*, Vol.47, No.2 (2021), 182~191.

Ryu, D. K., “Water quality and human health risk assessment on springs in Seoul,” *Master thesis*, University of Seoul, 2005.

Sattari, M. T., M. Abbasgoli Naebzad, and R. Mirabbasi Najafabadi, “Surface water quality prediction using decision tree method,” *Irrigation and Water Engineering*, Vol.4, No.3(2014), 76-88.

Shin, D. I. and K-. Y. Kwahk, “Development of a Detection Model for the Companies Designated as Administrative Issue in KOSDAQ Market,” *Journal of Intelligence and Information Systems*, Vol.24, No.3(2018), 157~176.

Shin, S. K., “Measures to improve the quality of mineral springs in Busan,” *BDI Policy Focus*, No.281(2015), 1~12.

Song, H., H. Lim, G. Park, H. Park, H. Lee, M. Jo, Y. Kim, and J. Oh, “Mineral Components of Water Supply Plants and Spring Waters in Northern Gyeonggi Area,” *Journal of environmental health sciences*, Vol.45, No.3(2019), 238~246.

Song, H., N. Kim, D. Jeong, Y. Lee, H. Jeon, Y. Kim, U. Jang, and J. Kim, “Water Quality and Influencing Factors at Dalbi Spring in Daegu,” *Journal of Korean Society of Environmental Engineers*, Vol.25, No.12(2003), 1570~1577.

Stidson, R. T., C. A. Gray, and C. D. McPhail, “Development and use of modelling techniques for real-time bathing water quality predictions,” *Water and Environment Journal*, Vol.26, No.1 (2012), 7~18.

Woo, J. S., “A study on the water quality of springs well in Cheoan city,” *Master thesis*, Hanbat national university, 2008.

Yang, S., J. Bae, H. Lim, E. Oh, B. Park, and N. Heo, “The Management and Water Quality of the Public Mineral Water in Jeonnam Area,” *Joint spring conference*, (2006), 93~100.

Yoon, T., H. Lee, G. Choi, S. Lee, M. Lee, and S. Eo, “Occurrence of Indicator Bacteria and Identification of Total Coliforms Using 16S rRNA Gene in Drinking Spring Water in Seoul,” *Journal of environmental health sciences*, Vol.39, No.6(2013), 513~521.

Abstract

Development of a water quality prediction model for mineral springs in the metropolitan area using machine learning

Yeong-Woo Lim* · Ji-Yeon Eom** · Kee-Young Kwahk***

Due to the prolonged COVID-19 pandemic, the frequency of people who are tired of living indoors visiting nearby mountains and national parks to relieve depression and lethargy has exploded. There is a place where thousands of people who came out of nature stop walking and breathe and rest, that is the mineral spring. Even in mountains or national parks, there are about 600 mineral springs that can be found occasionally in neighboring parks or trails in the metropolitan area. However, due to irregular and manual water quality tests, people drink mineral water without knowing the test results in real time. Therefore, in this study, we intend to develop a model that can predict the quality of the spring water in real time by exploring the factors affecting the quality of the spring water and collecting data scattered in various places. After limiting the regions to Seoul and Gyeonggi-do due to the limitations of data collection, we obtained data on water quality tests from 2015 to 2020 for about 300 mineral springs in 18 cities where data management is well performed. A total of 10 factors were finally selected after two rounds of review among various factors that are considered to affect the suitability of the mineral spring water quality. Using AutoML, an automated machine learning technology that has recently been attracting attention, we derived the top 5 models based on prediction performance among about 20 machine learning methods. Among them, the catboost model has the highest performance with a prediction classification accuracy of 75.26%. In addition, as a result of examining the absolute influence of the variables used in the analysis through the SHAP method on the prediction, the most important factor was whether or not a water quality test was judged nonconforming in the previous water quality test. It was confirmed that the temperature on the day

* Graduate School of Business IT, Kookmin University
** Graduate School of Business IT, Kookmin University
*** Corresponding Author: Kee-Young Kwahk
College of Business Administration/Graduate School of Business IT, Kookmin University
[02707] 77, Jeongneung-ro, Seongbuk-gu, Seoul, Korea
Tel: +82-2-910-4738, Fax: +82-2-910-4017, E-mail: kykwahk@kookmin.ac.kr

of the inspection and the altitude of the mineral spring had an influence on whether the water quality was unsuitable.

Key Words : mineral spring, water quality, prediction model development, autoML, SHAP

Received : November 15, 2022 Revised : February 17, 2023 Accepted : March 11, 2023
Corresponding Author : Kee-Young Kwahk

저 자 소 개

임영우
현재 국민대학교 비즈니스IT전문대학원 박사과정에 재학 중이며, 서울대학교 생활과학대학 소비자학 석사 학위를 취득하였다. 주요 관심분야는 Data analytics, Data mining, Deep Learning, Social network analysis, Knowledge management 등이다.

엄지연
현재 국민대학교 비즈니스IT전문대학원 석사과정에 재학 중이며, 주요 관심분야는 등이다. 주요 관심분야는 Data analytics, Data mining, Deep Learning, machine learning, Knowledge management 등이다.

곽기영
현재 국민대학교 경영대학과 비즈니스IT전문대학원 교수로 재직 중이다. 서울대학교 경영대학을 졸업하고 KAIST 경영과학과와 테크노경영대학원에서 석사 및 박사학위를 취득하였다. 주요 연구관심분야는 Social network analysis and its application, Data analytics, Users' behavior in social media, Social communication ecology, Knowledge management 등이다.