머신러닝을 이용한 식중독 발생 예측모형에 관한 연구

A Study on the Prediction Model of Food Poisoning Occurrence Using Machine Learning Algorithm

저자 (Authors)
엄선현, 이우창, 배재권
Sun Hyun Um, Woo Chang Lee, Jae Kwon Bae

출처 (Source)
로고스경영연구 19(2), 2021.6, 63-76 (14 pages)
Logos Management Review 19(2), 2021.6, 63-76 (14 pages)

발행처 (Publisher)
한국로고스경영학회
The Korean Association Of Logos Management

URL
http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE10576394

APA Style
엄선현, 이우창, 배재권 (2021). 머신러닝을 이용한 식중독 발생 예측모형에 관한 연구. 로고스경영연구, 19(2), 63-76.

로고스경영연구
제19권 제2호 (2021.06) pp. 63-76
한국로고스경영학회

Logos Management Review
DOI: 10.22724/LMR.2021.19.2.63

머신러닝을 이용한 식중독 발생 예측모형에 관한 연구

엄 선 현* · 이 우 창** · 배 재 권***

〈요 약〉

식중독(food poisoning)이란 식품의 섭취로 인해 인체에 유해한 미생물 또는 유독물질에 의하여 발생하였거나 발생한 것으로 판단되는 감염성 또는 독소형 질환을 말한다. 여름철(6~8월)은 폭염과 고온다습한 날씨로 식중독 발생 위험성이 높아지며 특히 감염에 취약한 영유아의 감염성 질환과 건강취약 계층인 어린이들의 집단 식중독 사고가 발생한다. 식중독 발생사고 및 식중독 발생 위험을 낮추기 위한 식중독 지수 개발 및 발생예측에 관한 연구가 필요한 시점이다. 기존 선행연구는 주로 통계방법을 이용하여 식중독 발생확률을 계량화하려는 시도가 있었으나 빅데이터 분석 및 머신러닝(machine learning)을 활용한 연구는 거의 없는 실정이다. 따라서 본 연구는 식중독 발생을 사전에 예측하기 위해 식중독 발생 데이터와 머신러닝 알고리즘을 이용하여 식중독 발생 예측모형을 구축하는 것이다. 본 연구는 국내 *A*시의 2015년부터 2019년까지 총 5개년 79건의 식중독 발생 현황 데이터와 기상청 데이터를 이용하여 식중독 발생 현황에 따른 식중독 발생 현황 기초통계분석과 XGBoost 머신러닝 알고리즘을 이용한 식중독 발생 예측모형을 구축하였다. 예측모형의 성능 측정 결과, Accuracy 92%, F1-Score 95%로 준수한 성능으로 식중독 발생 예측이 가능하다. 또한 '일조시간', '일강수량', '상대습도', '평균풍속', '평균운량', '최고기온'과 같은 기상요인이 식중독 발생여부에 있어 중요한 요인임을 확인하였다.

주제어 : 식중독(food poisoning), 식중독 지수, 머신러닝(machine learning), 식중독 발생 예측모형, XGBoost.

* 계명대학교 일반대학원 경영정보학전공 석사과정, djatjsgus123@naver.com
** 다이텍연구원 테스트베드연구센터 연구원, wcl0104@dyetec.or.kr
*** 교신저자, 계명대학교 경영대학 경영정보학전공 부교수, jkbae99@kmu.ac.kr
논문접수: 2021.06.09, 1차수정: 2021.06.16, 게재확정: 2021.06.28

Ⅰ. 서 론

식중독(food poisoning)이란 식품의 섭취로 인해 인체에 유해한 미생물 또는 유독물질에 의하여 발생하였거나 발생한 것으로 판단되는 감염성 또는 독소형 질환을 말한다. 식중독은 일반적으로 고열, 복통, 설사, 구토, 두통 등이 대표적인 증상으로 때로는 호흡곤란, 탈수증상 등을 일으켜 생명을 위험하게 할 수 있다. 식품의약품안전처의 식품안전나라(https://www.foodsafetykorea.go.kr)에서 제공하는 최근 3년간의 국내 식중독 발생 현황을 보면 2019년 4,075명, 2020년 2,747명, 2021년 2,262명으로 지속적으로 식중독이 발생하는 것을 알 수 있다. 특히 여름철(6~8월)은 폭염과 고온다습한 날씨로 식중독 발생 위험성이 높아지며 감염에 취약한 영유아의 감염성 질환과 건강취약 계층인 어린이들의 대규모 식중독 발생 위험이 높아진다. 여름철 식중독 발생의 주된 원인은 장염 비브리오 식중독이며, 이는 바닷물에서 존재하는 식중독균으로 해수의 온도가 15℃이상일 때 증식을 시작하고, 수온이 높을수록 빠르게 증식하여 여름에 환자가 집중적으로 발생한다. 식품의약품안전처 자료에 따르면 최근 5년간 병원성대장균 식중독 발생 현황은 총 195건(8,881명)의 환자 중 여름철에만 114건(58%), 6천357명(72%)의 환자가 발생한 바 있다. 집단발생 장소는 주로 어린이집이 가장 많고, 그 이외에 음식점, 산후조리원, 요양원 등 집단시설 관련 장소이다. 질병관리청에 따르면 2021년 5월 한 달간 총 52건의 집단발생이 보고되었고, 628명이 의료기관에서 치료를 받는 중이다. 이는 2017~2019년 평균 62건에 비해 적으나, 코로나19 펜데믹(pandemic)으로 인한 사회적 거리 두기 시행 후 감소하였던 집단발생이 무더위가 본격화되는 6월 이후 증가하여 주의가 요구된다. 정부와 지자체, 그리고 식품위생감시원, 소비자식품위생감시원은 식중독 사고 발생을 예방하기 위해 집단 식중독 발생 주의 요망과 식품안전사고 교육 캠페인을 지속적으로 펼치고 있다.

이처럼 다양한 식중독 발생 예방대책을 마련하기 위해 교육 이외에도 식중독 발생을 사전에 예방할 수 있는 예측모형의 필요성이 대두되고 있다. 기존 선행연구는 주로 전통적인 통계방법으로 식중독 발생을 예측하였으나 빅데이터와 머신러닝(machine learning) 알고리즘을 활용한 연구는 거의 없는 실정이다. 따라서 본 연구의 목표는 식중독 발생을 사전에 예측하기 위해 국내 *A*시의 식중독 발생 데이터와 XGBoost 머신러닝 알고리즘을 이용하여 식중독 발생 예측모형을 구축하는 것이다. 빅데이터(BigData)란 기존 데이터베이스의 데이터 저장·관리·분석 능력을 초과하는 다양한 형식을 가진 대량의 데이터를 의미한다. 다양한 분야에서 빅데이터가 생성, 분석, 활용되고 있는데 특히 보건의료 및 바이오 분야에서 빅데이터 분석은 사회경제적으로 큰 영향력을 발휘할 수 있기 때문에 크게 주목 받고 있다.

본 연구는 *A*시의 2015년부터 2019년까지 총 5개년 동안 79건의 식중독 발생 현황 데이터와 기상청 데이터를 이용하여 *A*시의 식중독 발생 현황에 따른 식중독 발생 현황 기초통계분석과 XGBoost 머신러닝 알고리즘을 이용한 식중독 발생 예측모형을 구축하고자 한다.

본 연구의 구성은 다음과 같다. 제Ⅱ장에서는 식중독 정의와 원인, 식중독 지수에 대한 이론적 배경을 서술하고, 식중독 발생예측 관련 선행연구를 살펴본다. 제Ⅲ장과 제Ⅳ장은 본 연구에서 진행한 *A*시 식중독 발생 현황 기초통계분석, 식중독 발생 예측모형 구축 및 결과에 대해 기술한다. 마지막으로 제Ⅴ장에서는 본 연구의 결론을 기술하고, 시사점 및 향후 연구계획에 대하여 기술한다.

Ⅱ. 이론적 배경

2.1 식중독 정의와 식중독 지수

‘식품위생법’ 제2조 제14항에 의하면 “식중독”이란 식품 섭취로 인해 인체에 유해한 미생물 또는 유독물질에 의하여 발생하였거나 발생한 것으로 판단되는 감염성 질환 또는 독소형 질환을 의미한다. 식중독의 종류는 〈표 1〉과 같이 총 3가지로 미생물 식중독, 자연독 식중독, 화학적 식중독으로 분류할 수 있다.

〈표 1〉 식중독의 분류

분류	종류	원인균 및 물질	
미생물 식중독 (30종)	세균성(18종)	감염형	살모넬라, 장염비브리오, 콜레라, 비브리오 불니피쿠스, 리스테리아 모노사이토제네스, 병원성대장균(EPEC, EHEC, EIEC, ETEC, EAEC), 바실러스세레우스(설사형), 쉬겔라, 여시니아 엔테로콜리티카, 캠필로박터 제주니, 캠필로박터콜리
		독소형	황색포도상구균, 클로스트리디움 퍼프린젠스, 클로스트리디움 보툴리눔, 바실러스세레우스(구토형)
	바이러스성(7종)	-	노로, 로타, 아스트로, 장관아데노, A형간염, E형간염, 사포 바이러스
	원충성(5종)	-	이질아메바, 람블편모충, 작은와포자충, 원포자충, 쿠도아
자연독 식중독		동물성	복어독, 시가테라독
		식물성	감자독, 원추리, 여로 등
		곰팡이	황변미독, 맥각독, 아플라톡신 등
화학적 식중독		고의 또는 오용으로 첨가되는 유해물질	식품첨가물
		본의 아니게 잔류, 혼입되는 유해물질	잔류농약, 유해성 금속화합물

	제조·가공·저장 중에 생성되는 유해물질	지질의 산화생성물, 니트로아민
	기타물질에 의한 중독	메탄올 등
	조리기구·포장에 의한 중독	녹청(구리), 납, 비소 등

우리나라 식중독 발생 건수는 해마다 차이는 있으나, 전반적으로 증가추세를 나타내고 있다. 급식, 외식문화, 배달음식 등 요식업 시장이 확대되면서 대형 식중독에 노출될 확률이 점점 높아지고 있다. 식중독 발생의 대부분은 집단 급식소에서 주로 발생하는데, 지하수, 정수기 등의 물과 김밥·비빔밥 등 복합조리식품, 육류요리, 샐러드 등 가열공정이 없는 신선식품 등이 주요 원인이다. 최근 들어 우리나라 식중독은 노로바이러스(Norovirus), 병원성대장균, 황색포도상구균, 살모넬라균이 주요 원인균으로 알려져 식품뿐만 아니라 환경위생도 지속적으로 점검해야 한다(이현아 외 2019).

식중독을 일으키는 세균이 번식하기 좋은 환경은 4℃~60℃ 사이 온도에서 증식하고, 대부분 35℃~36℃ 내외에서 번식 속도가 빠르기 때문에 여름철에 세균성 원인균에 의한 식중독 발병 위험이 가장 높다. 최근 위생관념이 발달하고 생활이 윤택해지면서 부패한 음식에 의한 세균성 식중독에 비해 바이러스성 식중독이 증가하고 있다. 노로바이러스는 물을 통해 전염되고 2차 감염이 흔하기 때문에 집단적인 발병양상을 보인다. 로타바이러스(Rotavirus)는 주로 영유아나 아동급성 설사병의 가장 흔한 원인으로 일교차가 크고 건조한 11월부터 환자 발생이 늘기 시작해 1월~3월경에 많이 발생하여, 주로 호흡기와 손으로 전염된다(박지애 외, 2016).

고수일(2017)은 본 연구의 분석대상인 *A*시를 분석하면서 최고기온이 33℃ 이상인 날의 횟수가 평균 32일이며, 민감도 부문의 대응변수 중 인구 10만 명당 식중독 진료환자수는 *A*시가 779명으로 가장 높게 나타났다. 민감도가 비슷한 *B*시와 *C*시보다 식중독 취약계층인 어린이와 노인 인구 비율이 낮음에도 불구하고, 식중독 진료환자 수가 높게 나타났다. 식중독이 가장 많이 발생하는 원인시설은 음식점으로 음식점의 위생수준을 파악할 수 있는 식품접객위반율은 *A*시가 8.53%로 전국에서 2위를 차지한 바 있다.

다음으로 식중독지수의 정의는 다음과 같다. 최근 5년 동안의 세균성, 바이러스성 식중독 발생 유무를 기반으로 기상에 따른 식중독 발생 확률을 예측한다. 일 2회(6시, 18시)에 생산되며, 식중독지수는 기상청, 식약처, 국민건강보험공단, 국립환경과학원이 공동으로 개발한 지수이다. 산출방법은 기상과 환경에 따른 식중독 발생 위험도를 세균성, 바이러스성, 식중독으로 나누어 계산하고 이를 과거 월별 세균성, 바이러스성 발생 비율에 따라 계산한 후 확률로 변환하여 최종 식중독 지수를 산출한다.

식중독 지수는 지수 범위에 따라 관심, 주의, 경고, 위험으로 구분되며 관심의 지수범위의 경우 55미만, 주의는 55이상 71미만, 경고는 71이상 86미만, 위험의 경우는 86이상이며 세부내용은 〈표 2〉와 같다.

〈표 2〉 식중독 지수와 단계별 대응요령

단계	지수범위	대응요령
위험	86 이상	- 식중독 발생가능성이 매우 높으므로 식중독 예방에 각별한 경계가 요망됨 - 설사, 구토 등 식중독 의심 증상이 있으면 의료기관을 방문하여 의사 지시에 따름 - 식중독 의심 환자는 식품 조리 참여에 즉시 중단하여야함
경고	71 이상 86 미만	- 식중독 발생가능성이 높으므로 식중독 예방에 경계가 요망됨 - 조리도구는 세척, 소독 등을 거쳐 세균오염을 방지하고 유통기한, 보관방법 등을 확인하여 음식물 조리, 보관에 각별이 주의하여야함
주의	55 이상 71 미만	- 식중독 발생가능성이 중간 단계이므로 식중독예방에 주의가 요망됨 - 조리음식은 중심부까지 75℃(어패류 85℃)로 1분 이상 완전히 익히고 외부로 운반할 때에는 가급적 아이스박스 등을 이용하여 10℃이하에서 보관 및 운반
관심	55 미만	- 식중독 발생가능성은 낮으나 식중독 예방에 지속적인 관심이 요망됨 - 화장실 사용 후, 귀가 후, 조리전에 손 씻기를 생활화

출처 : 기상청 날씨누리(https://www.weather.go.kr/plus/life/jisudaymap_A01_2.jsp)

2.2 식중독 및 식중독 발생 예측 관련 연구

우리나라의 식중독 발생률은 1960년대 이후 개인위생 및 식품위생 수준이 향상됨에 따라 감소하다가 1990년대 이후 증가 추세를 보이고 있는데, 여러 요인들 중에도 집단급식의 증가와 노로바이러스와 같은 신종 균의 출현을 지적할 수 있다. 또한 국내에서 식중독이 5월과 6월에 많은 사례가 발생하고, 12월에는 새로운 사례가 발생하고 있으며 최근 식중독의 새로운 특징은 노로바이러스와 장출혈성대장균감염증(EHEC)와 같은 병원체에 점점 더 많이 기인한다(권준욱 외, 2007). 바이러스성 식중독은 장염바이러스 중 노로바이러스가 가장 주요 병원체로 작용하며 이외에도 아스트로바이러스(Astroviruses)나 로타바이러스에 의한 집단 설사 사례가 국내에서 보고된 바 있다. 노로바이러스는 오염된 식수와 굴 등 어패류의 생식을 통한 감염 사례가 많이 보고되어 있으나 사람 간 전파도 자주 일어나는 전염력이 매우 높은 바이러스다. 2000년 이후 질병관리본부는 바이러스성 설사의 국내 발생 현황을 파악하기 위하여 전국의 17개 시도보건환경연구원과 노로바이러스를 포함한 4종의 바이러스성 장염 원인 병원체에 대한 전국적인 실험실 감시체계를 운영한 결과 바이러스성 병원체가 확인된 사례의 약 18%에서 노로바이러스가 검출되었고, 집단설사 사례는 대부분 노로바이러스가 원인병원체로 확인된 바 있다(지영미, 2006). 향후 지구 온난화에 비례해서 전염병 질환 등 각종 질병이 확산될 것이라는 세계보건기구의 경고처럼, 건강 및 의료분야는 이상고온, 가뭄, 홍수, 폭염 등의 이상기후로 인해 소화기 질환을 유발하는 세균, 바이러스 및 기생충에 의한 매개성 및 수인성 식중독 관련 전염성 질환의 다발이 예상된다. 이는 지역별 차이는 있겠으나 2030년에는 설사병을 앓는 환자들이 기후변화가 없는 환경조건과 대비하여 10%정도 증가할 것으로 예측하

고 있다(김정선, 2010). 정명섭 외(2009)는 일본의 니카타, 도쿄 및 오사카와 미국의 버지니아, 노스캐롤라이나 및 사우스캐롤라이나 지역의 식중독 발생 현황을 비교·검토한 결과 우리나라의 향후 발생양상의 변화를 줄 수 있는 식중독 병원체로는 살모넬라, 병원성 대장균, 캠필로박터, 비브리오가 주목되며, 식품원재료로부터 이들 미생물의 관리가 필요한 것으로 파악되었다. 국내 식중독 발생 통계자료를 이용한 식중독 발생 예측을 일반화 선형모델(GLM-Poisson Regression Model)로 분석한 결과, 살모넬라, 장염비브리오 및 황색포도상구균의 식중독 발생건수에 가장 큰 영향을 주는 기후인자는 '월 평균기온'이며, '월 평균기온'이 1℃ 상승시 식중독 발생건수는 살모넬라의 경우 47.8% 증가하였다. 노로바이러스의 경우 '월 평균온도', '습도', '최소습도'가 발생에 가장 많은 영향을 주지만 온도 상승에 따른 감소(16.4%/℃)를 나타내었으며, 병원성대장균은 '강수량'이 가장 큰 기후인자로 '월 평균강수량' 증가에 따라 감소(15.1%/mm)를 보였다. 노로바이러스와 병원성대장균에 대해서는 주기성에 대한 고려가 가능한 계절 자기회귀 누적이동평균기법(SARIMA)을 활용하였으나 일반화 선형모델에 비해 모델 적합도가 떨어지는 것으로 나타났다. 신호성 외(2009)는 기후변화와 식중독 발생과의 관계를 조명하고 미래 기후변화 시나리오에 입각하여 식중독 발생 예측을 연구하였다. 이들에 따르면 '상대습도'는 식중독 발생 환자 수에 통계적으로 유의한 영향을 미치며, 상대습도가 1% 증가할수록 환자 수는 1.7% 감소하는 것으로 예측되었다. 우리나라 식중독 발생은 시차를 두고 기후 변수에 영향을 많이 받고 있으나 포아송 회귀모형을 이용한 식중독 발생 예측은 이들 변수보다 이전 시점의 식중독 발생 건수에 더 많이 영향을 받는 것으로 나타났다(여인권, 2012). 세균성 원인균에 의한 식중독 발병률에는 '평균기온', '일조량편차', '기온편차'가 유의미한 영향을 미치고, 바이러스성 원인균에 의한 식중독 발병률에 영향을 미치는 기상요인은 '최소증기압', '일조량편차', '기온편차'로 나타났다(박지애 외, 2016). 육현준 외(2018)는 정형 및 비정형 자료를 활용한 전국 식중독 일별 예측모형을 제시하였다. 이들은 식중독의 발생건수와 발생확률을 추정하는 모형을 제시하였고, 식중독을 유발하는 원인에 대해 여름철에 발생하는 박테리아성과 겨울철에 발생하는 바이러스성으로 구분하여 모형을 구축하였다. 연구결과, 박테리아성 식중독은 영과잉 음이항 회귀모형, 바이러스성 식중독은 음이항 회귀모형이 가장 적합한 모형이라고 주장하였다. 김종규(2020)는 우리나라에서 병원성 대장균 식중독 발생과 기후요소의 영향요인을 분석하였다. 이들은 병원성 대장균 식중독의 발생이 주요 기상변수와 관련이 있고, 특히 최저 및 최고기온과 강수량이 유의한 영향력을 보여 이는 병원성 대장균 식중독 발생이 기후변화에 의해 영향을 받았음을 주장한 바 있다.

본 연구는 기존 선행연구에서 시도하지 않았던 XGBoost 머신러닝 알고리즘을 이용하여 식중독 발생 예측모형을 제시하고자 한다. 또한 선행연구를 통해 밝혀진 기상요인이 식중독에 큰 영향을 미친다는 연구결과를 바탕으로 기상청 데이터와 식중독 발생 데이터

를 통합하여 분석에 활용하였다.

Ⅲ. 연구 설계 및 분석 방법

3.1 A시 식중독 발생 현황 기초통계분석

본 연구는 식중독 발생 현황 기초통계분석을 위해 *A*시의 2015년부터 2019년 총 5개년 식중독 통계데이터, 식중독 조사결과보고서, 발생조사결과데이터, 그리고 기상청의 기상데이터 등 총 79건의 데이터를 활용하였다. 기초통계분석을 위해 4개의 데이터를 통합하는 전처리 과정을 진행하였다. 전처리된 데이터셋의 변수는 〈표 3〉과 같다.

〈표 3〉 통합 데이터셋 변수

변인	변수명	변인	변수명
X1	구	X26	원인균
X2	gu	X27	평균기온
X3	날짜	X28	최고기온
X4	년	X29	최저기온
X5	월	X30	평균운량
X6	계절	X31	일강수량
X7	신업종구분	X32	일조시간
X8	어패류식당	X33	상대습도
X9	뷔페	X34	평균풍속
X10	기타음식점	X35	업종구분_가정집
X11	유아교육시설	X36	업종구분_기타
X12	학교	X37	업종구분_음식점
X13	기타	X38	업종구분_학교
X14	연령 1~4	X39	업종구분_학교외집단급식소
X15	연령 5~9	X40	노로바이러스
X16	연령 10~14	X41	바실러스세레우스
X17	연령 15~19	X42	병원성대장균
X18	연령 20~29	X43	불명
X19	연령 30~39	X44	살모넬라
X20	연령 40~49	X45	원충
X21	연령 50~59	X46	캠필로박터제주니
X22	연령 60~69	X47	퍼프린젠스
X23	연령 70이상	X48~59	1월~12월
X24	업종구분	X60	평균기온(범주)
X25	환자수	X61	평균상대습도

*A*시의 2015년~2019년 원인시설, 원인물질에 따른 식중독 발생 현황에 대한 기초통계 분석 결과는 〈표 4〉와 같다. 원인 시설별 식중독 발생 전체 건수는 음식점이 79점 중 49건으로 가장 많이 발생되었다. 또한 원인 시설별 식중독 환자 수는 학교가 1,023명 중 564명으로 나타나 전체 환자 중 절반 이상이 학교에서 나타났다. 발생 건수가 가장 많았던 음식점은 발생 건당 평균 환자 수가 약 4명인 것에 비해 학교는 발생 건당 평균 환자 수가 51명으로 나타나 학교급식과 같은 집단급식의 식중독 위험이 더 크다는 것을 보여주고 있다. 학교 외 집단급식으로 인해 나타나는 발생 건당 평균 환자 수는 26명으로 나타났다.

〈표 4〉 원인시설별 식중독 발생 현황 (단위 : 건, 명)

		학교	학교외 집단급식	음식점	가정집	기타	합계
2019	발생 건수	2(0.18)	1(0.09)	4(0.36)	0(0.00)	4(0.36)	11
	환자 수	96(0.61)	26(0.17)	10(0.06)	0(0.00)	25(0.16)	157
2018	발생 건수	3(0.30)	0(0.00)	5(0.50)	0(0.00)	2(0.20)	10
	환자 수	221(0.90)	0(0.00)	17(0.07)	0(0.00)	7(0.03)	245
2017	발생 건수	1(0.06)	1(0.06)	15(0.83)	0(0.00)	1(0.06)	18
	환자 수	20(0.22)	11(0.12)	52(0.58)	0(0.00)	6(0.07)	89
2016	발생 건수	4(0.17)	4(0.17)	11(0.48)	0(0.00)	4(0.17)	23
	환자 수	192(0.50)	119(0.31)	43(0.11)	0(0.00)	29(0.08)	383
2015	발생 건수	1(0.06)	0(0.00)	14(0.82)	1(0.06)	1(0.06)	17
	환자 수	35(0.23)	0(0.00)	60(0.40)	3(0.02)	51(0.34)	149
합계	발생 건수	11(0.14)	6(0.08)	49(0.62)	1(0.01)	12(0.15)	79
	환자 수	564(0.55)	156(0.15)	182(0.18)	3(0.00)	118(0.12)	1,023

〈표 5〉는 *A*시의 2015년~2019년 원인물질별 식중독 발생 현황을 조사한 결과이다. 원인 물질별 식중독 전체 발생 건수는 원인불명이 32건(41%)으로 가장 많고, 원인불명을 제외한 식중독 전체 발생 건수는 노로바이러스가 16건(20%), 병원성 대장균의 환자 수는 272명(27%)으로 나타났다. 특히 식중독 발생 건수와 환자 수가 가장 많았던 2016년에는 노로바이러스와 병원성 대장균으로 인한 식중독 발생 건수 및 환자 수가 가장 높은 비율을 차지하였다. 또한 발생 건당 평균 환자 수는 캠필로박터제주니가 약 100명으로 가장 많이 발생하였다. 다음으로 살모넬라가 약 51명, 클로스트리디움퍼프린첸스가 약 35명, 병원성 대장균이 약 23명, 노로바이러스가 약 15명으로 나타났다.

〈표 5〉 원인물질별 식중독 발생 현황 (단위: 건, 명)

		병원성 대장균	살모넬라	캠필로박터 제주니	클로스트리디움 퍼프린젠스	바실러스 세레우스	노로 바이러스	원충	불명
2019	발생 건수	2(0.18)	0(0.00)	0(0.00)	2(0.18)	0(0.00)	1(0.09)	1(0.09)	5(0.45)
	환자 수	68(0.43)	0(0.00)	0(0.00)	69(0.44)	0(0.00)	6(0.04)	2(0.01)	12(0.08)
2018	발생 건수	0(0.00)	0(0.00)	2(0.20)	0(0.00)	2(0.20)	0(0.00)	2(0.20)	4(0.40)
	환자 수	0(0.00)	0(0.00)	201(0.82)	0(0.00)	8(0.03)	0(0.00)	5(0.02)	31(0.13)
2017	발생 건수	4(0.22)	0(0.00)	0(0.00)	0(0.00)	2(0.11)	3(0.17)	2(0.11)	7(0.39)
	환자 수	19(0.21)	0(0.00)	0(0.00)	0(0.00)	4(0.04)	19(0.21)	7(0.08)	40(0.45)
2016	발생 건수	4(0.17)	0(0.00)	0(0.00)	0(0.00)	0(0.00)	8(0.35)	5(0.22)	6(0.26)
	환자 수	166(0.43)	0(0.00)	0(0.00)	0(0.00)	0(0.00)	176(0.46)	20(0.05)	21(0.05)
2015	발생 건수	2(0.12)	1(0.06)	0(0.00)	0(0.00)	0(0.00)	4(0.24)	0(0.00)	10(0.59)
	환자 수	19(0.13)	51(0.34)	0(0.00)	0(0.00)	0(0.00)	41(0.28)	0(0.00)	38(0.26)
합계	발생 건수	12(0.15)	1(0.01)	2(0.03)	2(0.03)	4(0.05)	16(0.20)	10(0.13)	32(0.41)
	환자 수	272(0.27)	51(0.05)	201(0.20)	69(0.07)	12(0.01)	242(0.24)	34(0.03)	142(0.14)

3.2 식중독 발생 예측모형 구축

본 연구는 식중독 발생 예측모형 구축을 위한 데이터셋으로 기초통계분석에서 전처리된 데이터셋을 머신러닝 알고리즘을 이용한 식중독 발생 예측모형 구축에 적합하도록 전처리를 수행하였다(〈표 6〉 참조). 전처리된 데이터셋은 *A*시의 5개년 동안의 월별 식중독 발생여부에 대한 데이터로 개수는 60개다. 데이터 수가 작은 것을 고려하여 학습용 데이터셋 80%(48개), 검증용 데이터셋 20%(12개)로 설정하였으며, XGBoost 머신러닝 알고리즘을 이용하여 식중독 발생 예측모형을 구축하였다. 검증용 데이터셋은 가장 최근 년도의 데이터를 사용하였으며, 학습용 데이터는 과거의 데이터로 사용하였다. XGBoost 머신러닝 알고리즘은 의사결정나무기반의 앙상블 기법으로 Greedy 알고리즘을 통한 가지치기로 과적합을 규제하여 분류와 회귀영역에서 뛰어난 예측 성능을 보여준다. 또한 모형에 사용된 변수의 중요도를 알 수 있어 분석 결과 해석에 용이한 기법이다. 본 연구는 빅데이터 분석과 머신러닝 알고리즘 개발에 유용한 *Python* 프로그램을 활용하였다.

〈표 6〉 식중독 전체 발생 여부 예측 데이터 변수

변인	변수명	변인	변수명
X1	년	X10	평균풍속
X2	월	X11~22	1월~12월
X3	평균기온	X23~30	구군(동구 외 7개)

X4	최고기온	X31	봄
X5	최저기온	X32	여름
X6	평균운량	X33	가을
X7	일강수량	X34	겨울
X8	일조시간	Y	식중독 발생여부(종속변수)
X9	상대습도		

Ⅳ. 연구결과

전치리된 학습용 및 검증용 데이터셋를 이용하여 XGBoost 머신러닝 알고리즘을 적용해 결과를 도출하였으며, 파라미터는 Learning rate 0.1, Max_depth 3, N estimators 400개로 설정하였다. 예측모형의 성능은 〈표 7〉과 같이 혼동행렬을 이용하여 Accuracy와 F1-Score을 평가지표로 사용하여 측정하였다.

〈표 7〉 식중독 발생 여부 혼동행렬

		예측	
		발생 X	발생 O
실제	발생 X	TN	FP
	발생 O	FN	TP

〈표 7〉에서 TN은 True Negative를 의미하며 실제로 식중독이 발생하지 않았고 예측모형도 식중독이 발생하지 않았다고 예측한 것이다. FN은 False Negative를 의미하며 실제로 식중독이 발생하지 않았는데 예측모형은 식중독이 발생했다고 예측한 것이다. FP는 False Positive를 의미하며 실제로 식중독이 발생하지 않았는데 예측모형은 식중독이 발생했다고 예측한 것이다. TP는 True Positive를 의미하며 실제로 식중독이 발생하였고 예측모형도 식중독이 발생했다고 예측한 것이다. 식 (1)~(4)는 혼동행렬을 바탕으로 도출할 수 있는 예측모형의 성과지표다.

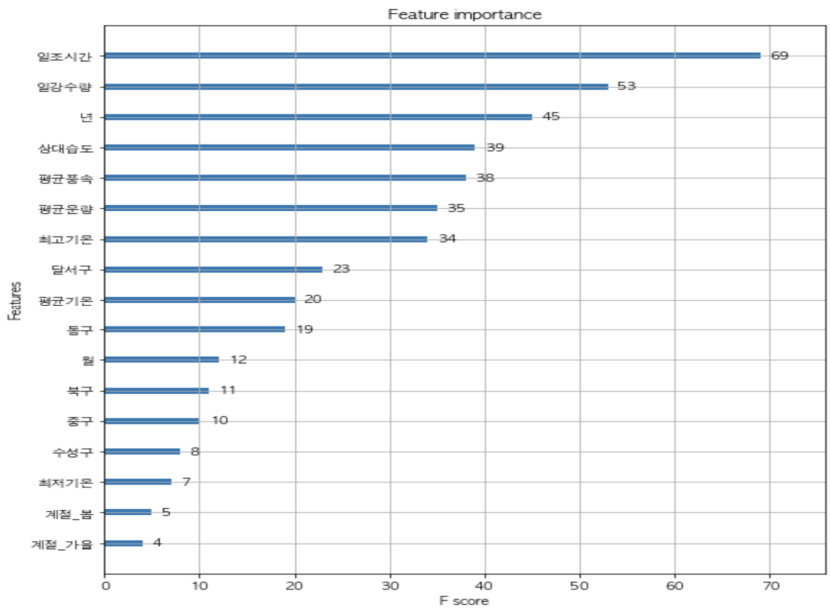
$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (1)$$

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

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

$$F1\,score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

식중독 발생 예측모형의 성능 측정 결과 Accuracy 92%, Precision 100%, Recall 90%, F1-Score 95%로 식중독 발생 예측에 있어서 준수한 성능을 보이는 것을 확인할 수 있었다. 〈그림 2〉는 XGBoost 머신러닝 알고리즘를 이용하여 F1-Score를 기반으로 각 변수의 중요도를 나타낸 것이다. 상위에 있는 변수들을 살펴보면 '일조시간', '일강수량', '상대습도', '평균풍속', '평균운량', '최고기온'으로 기상요인이 식중독 발생여부에 있어 중요한 요인임을 확인할 수 있다.



〈그림 2〉 XGBoost 머신러닝 알고리즘을 이용한 변수의 중요도 도출

V. 결론 및 시사점

본 연구는 식중독 발생 데이터와 기상청의 기상데이터를 이용하여 XGBoost 머신러닝 알고리즘을 적용해 *A*시의 식중독 발생 예측모형을 구축하였다. 예측모형의 성능 측정 결과 Accuracy 92%, F1-Score 95%로 준수한 성능으로 *A*시의 식중독 발생을 예측할 수 있음을 확인하였다. 이는 *A*시뿐만 아니라 전국의 각 지자체가 수집한 식중독 발생 데이터와 기상 데이터를 활용한다면 전국 또는 지자체별 식중독 발생 예측모형을 구축할 수

있음을 시사한다. 연구변수의 중요도 분석결과, '일조시간', '일강수량', '상대습도', '평균풍속', '평균운량', '최고기온'과 같이 기상요인이 식중독 발생에 있어 중요한 요인임을 확인할 수 있었다. 이는 기상요인들이 식중독균의 생육환경에 큰 영향을 미치며, 기후변화에 따라 식중독 발생률이 높아질 수 있음을 의미한다. 또한 변수의 중요도를 알 수 있다는 점을 이용해 지자체별 식중독 발생에 영향을 주는 변수를 분석하여 지자체별 식중독 발생 예방대책 마련에 활용할 수 있을 것으로 기대한다.

본 연구는 식중독 발생 데이터와 기상데이터를 머신러닝을 이용하여 식중독 발생 예측모형을 구축했다는 것에 학술적 의의가 있다. 식중독 발생 예측에 관한 기존 연구는 대부분 전통적인 통계방법을 활용하였다. 그러나 본 연구는 예측모형의 정확도 향상을 위해 앙상블 기법의 XGBoost 머신러닝 알고리즘을 이용하여 식중독 발생 예측모형 구축에 활용함으로써 기존 연구와 차별성을 갖는다.

본 연구는 다음과 같은 한계점을 가진다. 본 연구에서 사용한 식중독 발생 데이터의 수가 적다는 것이 한계점이며, 향후 분석기간을 늘려 더 많은 식중독 발생 데이터를 확보한다면 예측모형의 성능을 향상시킬 수 있을 것이다. 또한 소셜데이터(social data) 및 기후관련 센서 데이터(sensor) 등 반정형, 비정형 데이터를 활용하고, 다양한 인공지능기법을 적용한다면 식중독 발생예측 모형의 고도화에 큰 기여를 할 것이다.

참고문헌

김정선(2010), 국내외 식중독 관리 및 정책 동향, 식품문화 한맛한얼, 3(2), 174-180.

김종규(2020), 우리나라에서 병원성 대장균 식중독 발생과 기후요소의 영향, 한국환경보건학회지, 46(3), 353-358.

권준욱, 이철헌(2007), 최근 우리나라 식중독 발생 현황 고찰, 대한의사협회지, 50(7), 573-581

고수일(2018), 기후변화에 따른 식중독 취약성 평가, 국내박사학위논문, 충남대학교 대학원.

박지애, 김장묵, 이호성, 이해진(2016), 기상요인과 식중독 발병의 연관성에 대한 빅 데이터 분석, 디지털융복합연구, 14(3), 319-327.

신호성, 정기혜, 윤시몬, 이수형(2009), 기후변화와 식중독 발생 예측, 보건사회연구, 29(1), 143-16.

여인권(2012), 식중독 발생 예측모형, 한국데이터정보과학회지, 23(6), 1117-1125.

여인권(2013), 원인균별 식중독 발생 건수 예측, 응용통계연구, 26(6), 923-932.

우영춘, 이성엽, 최완, 안창원, 백옥기(2019), 디지털 헬스케어 데이터 분석을 위한 머신러닝 기술 활용 동향.

육현준, 황은지, 나종화(2018), 정형 및 비정형 자료를 활용한 전국 식중독 일별 예측모형, 한국데이터정보과학회지, 29(6), 1491-1503.

이지혜, 제미정, 조명지, 손현석(2014), 보건의료 분야의 빅데이터 활용 동향, Information and Communication Magazine, 32(1), 63-75.

이현아, 최지혜, 박성민, 남해성, 최진화, 박준혁(2019), 2019년 충남지역 고등학교에서 발생한 다병원체에 의한 집단식중독의 역학적 분석, 한국환경보건학회지, 45(5), 434-442.

지영미(2006), 노로바이러스 식중독의 국내 발생 및 실험실 감시 현황, 보건복지포럼, 2006(8), 26-34

정명섭, 오상석(2009), 기후변화에 따른 식중독 발생 영향분석 및 관리 체계 연구, 식품의약품안전청 최종보고서 정책-식품-2009-09, 서울.

A Study on the Prediction Model of Food Poisoning Occurrence Using Machine Learning Algorithm

Sun Hyun Um* · Woo Chang Lee** · Jae Kwon Bae***

〈Abstract〉

Food poisoning is an illness caused by eating contaminated food. In most cases of food poisoning, the food is contaminated by bacteria, such as salmonella or norovirus. In summer, due to the heat and humid weather, the risk of infectious diseases in infants and children and large-scale food poisoning in children increases. It is necessary to study food poisoning occurrence prediction to lower the risk of food poisoning. Previous studies mainly predicted the occurrence of food poisoning using statistical methods, but there are few studies using big data analysis and machine learning algorithm. Therefore, this study is to construct a food poisoning occurrence prediction model using food poisoning occurrence data and machine learning algorithms to predict food poisoning in advance. In this study, food poisoning occurrence prediction model using XGBoost machine learning algorithm was constructed using 79 cases of food poisoning occurrence data for 5 years from 2015 to 2019 in city A and data from the Korea Meteorological Administration. As a result of measuring the performance of the predictive model, it was confirmed that the occurrence of food poisoning in A city could be predicted with performance that was satisfactory with Accuracy 92% and F1-Score 95%. Also, it was confirmed that meteorological factors such as 'sunshine time', 'daily precipitation', 'relative humidity', 'average wind speed', 'average cloud amount', and 'maximum temperature' were important factors in the occurrence of food poisoning.

Key words : Food Poisoning, Food Poisoning Index, Machine Learning, Food Poisoning Occurrence Prediction Model, XGBoost.

* Graduate School of MIS Master's Program, Keimyung University, djatjsgus123@naver.com
** Dyetec Research Institute Testbed Research Center Researcher, wcl0104@dyetec.or.kr
*** Corresponding Author, Dept. of Management Information Systems, Keimyung University, jkbae99@kmu.ac.kr
Received: 2021.06.09, 1st Revised: 2021.06.16, Accepted: 2021.06.28