ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2022 September: 28(3): 139~159 http://dx.doi.org/10.13088/jiis.2022.28.3.139

효율적 수입식품 검사를 위한 머신러닝 기반 부적합 건강기능식품 탐지 방법*

이경수
경희대학교 빅데이터응용학과
(nowksu@khu.ac.kr)

박예린
경희대학교 빅데이터응용학과
(yyam1020@khu.ac.kr)

신윤종
경희대학교 경영학과
(yoonjong12@khu.ac.kr)

손권상
경희대학교 경영학과
(miroo1215@khu.ac.kr)

권오병
경희대학교 경영학과
(obkwon@khu.ac.kr)

코로나19 이후 건강기능식품의 관심이 높아짐에 따라 수입 식품 안전성 검사의 중요성도 더욱 커지고 있다. 그러나 매년 증가하는 건강기능식품 수입량과 반대로 식품 검사에 필요한 예산과 인력은 한계점에 다다르고 있다. 따라서 본 연구의 목적은 수출입 식품 중 건강기능식품을 대상으로 데이터의 특성을 살펴보고, 판별의 정확성과 결과의 설명 가능성을 고려하여 효율적으로 부적합 식품을 탐지할 수 있는 기계학습 모델 기반 자동화 시스템 설계 방안을 제시하는 것이다. 이를 위해 첫째, 부적합 판정에 영향을 미치는 식품 검사 데이터로부터 부적합 판정에 유의한 파생변수를 생성하며, 둘째, 건강기능식품 수출입 검사 데이터에 대한 탐색적 분석을 통해 클래스 불균형과 비선형성 등을 고려하여 영향변수를 선정하며, 셋째, 다양한 머신러닝 기법을 적용하여 모델 별 성능과 해석가능성에 대해 비교를 수행하고자 한다. 성능 분석 결과, 앙상블 모델이 가장 우수하였으며, 본 연구에서 제안하는 파생변수 및 모델이 수출입 식품 검사에서 활용하고 있는 시스템에 도움이 될 수 있음을 확인하였다.

주제어 : 수입식품, 건강기능식품, 식품안전, 부적합 탐지, 머신러닝, 데이터 불균형

논문접수일 : 2022년 8월 22일 논문수정일 : 2022년 9월 8일 게재확정일 : 2022년 9월 15일
원고유형 : Regular Track 교신저자 : 권오병

1. 서론

식품에 대한 기호도의 다양화로 한국의 식품 수입 물량은 꾸준히 증가하고 있다. 동시에 경제 수준 향상으로 소비자들의 식품 안전에 대한 수준이 높아지면서, 수입 식품에 대한 안전성 검사의 중요성도 더욱 커지고 있다. 식품안전 검사는 세계적인 추세이며 식품 관련 기업(Food Business Operators (FBO))들은 Global Food Safety Initiative과 같은 프로그램 등 식품안전에 대한 각종 인증, 감사에 대비해야 한다 (Kleboth et al., 2022; GFSI, 2022). 한국의 경우에도 수입식품 안전성 검사를 위해 식품 위해도가 높을 것으로 예상되는 수입식품을 무작위로 선별하여 이화학검사 등을 통해 살펴보는 무작위정밀검사를 실시하고 있다. 그러나 매년 수입식품이 증가하고 있고 코

* This research was supported by a grant (21163MFDS516) from Ministry of Food and Drug Safety in 2022.

로나 이후 건강기능 식품에 대한 관심이 높아짐에 따라 부적합 식품 검사 시간과 비용 등의 자원적 한계가 존재한다. 따라서 부적합 수입식품의 사전 자동 예측과 같은 관련 행정업무의 자동화를 통해 효율적으로 보완할 필요가 있다.

따라서, 부적합 탐지의 효율성과 신뢰성 제고를 위해 기계학습 방법을 적용하려는 시도가 진행되어 왔다. 식품의 부적합 여부 파악을 위해 사물인터넷과 센서네트워크 등으로부터 획득된 데이터로 기계학습(Sharif et al., 2021) 또는 딥러닝(Wu et al., 2019)을 통해 부적합 또는 감염여부를 판정할 수 있다. 그러나 본 연구에서 관심을 가지는 문제의 경우에는 수입 단계에서 부적합 가능성이 있어 보이는 식품을 실제 검사 전에 판정하는 경우이므로 센서네트워크를 통한 부적합 판정은 적합하지 않다. 한편 식품의 안전성 문제는 그 판정의 위중함 때문에 설명 가능하지 않은 블랙박스유형의 판별알고리즘을 활용하기에 적합하지 않다. 그래서 혹여 예측 성능이 떨어지더라도 설명 가능한 알고리즘을 사용하고 있다.

더욱이 수입식품 부적합 판정은 수입 사례 건수에 비해 극소수이므로 데이터 불균형이 심하다는 특징이 있다. 따라서 알고리즘 외에 데이터 불균형 해소를 위한 전처리와 특성공학이 필요하다. 그러므로 부적합식품 탐지 관련된 기존 연구에서 예측모형에 적합한 변수를 생성하고 불균형 데이터를 정제하는 등 다양한 실험을 수행한 바 있다. 그러나 다양한 분류모형을 사용하지 않은 점과 예측모형의 성과를 제고할 수 있는 파생변수를 충분히 고려하지 않은 점 등의 한계가 있다 (조상구, 최경현, 2018). 더욱이 수입식품에 대한 데이터 분석 및 기계학습 모형 활용에 대한 연구는 부족한 실정이다.

따라서 본 연구의 목적은 수출입 식품 중 건강기능식품을 대상으로 데이터의 특성을 살펴보고, 판별의 정확성과 결과의 설명 가능성을 고려하여 효율적으로 부적합 식품을 탐지할 수 있는 기계학습 모델 기반 자동화 시스템 설계 방안을 제시하는 것이다. 특히 부적합식품 탐지 문제는 전형적인 데이터 불균형이어서 예측 모형의 성능을 저하시킬 수 있으므로, 기존의 선행 연구의 하이브리드 접근법을 기반으로 건강기능식품 데이터셋의 불균형 데이터 분류 문제를 해결하여 최적의 건강기능식품의 부적합 탐지 모델을 구축하였다. 이를 위해 수입식품 판정과 관련한 2016년부터 2021년까지 총 5년여간의 실제 데이터셋을 대상으로 최적의 기계학습 방법을 개발하였다. 이때 모형의 일반화와 강건성을 보장하기 위해 최근 데이터는 학습에 사용하지 않고 오직 검증에만 사용하였다. 특히, 확보된 수입식품 관련 데이터셋 중에서 변수 선정 및 기존 변수로부터 새로운 변수를 확보하는 파생변수 생성 과정을 거쳤으며, 일부 변수의 경우 값의 정규화 및 구간화나 확률값으로의 치환 작업을 거침으로써 부적합 식품 탐지율을 제고하였다.

2. 관련 연구

2.1. 부적합 식품 탐지

우리나라에 수입되는 식품의 증가 추세와 함께 부적합 식품 적발을 위한 수입 식품 검사 단계에서의 위험분석 기반의 사전적 안전관리의 필요성이 부각되고 있다(조상구, 최경현, 2018). 사전적 안전관리, 즉, 예방적 위험관리 체제는 최종 소비자에게 집중되던 식품위험과 부적합

〈Table 1〉 Literature Review on Food Defection Detection

Reference	Method	Data	Comments
Cho, Choi. (2018)	Logistic Regression, KNN, DT, RF, Bagging, Gaussian Naïve Bayes, Gradient Boosting, MLP	Food import declaration data	Defective food detection prediction is performed by creating derived variables based on food import declaration data and applying defect prediction techniques based on machine learning.
Cho, Cho. (2020)	Gradient Boosting, Ada Boost, Logistic Regression, SVM, RF, KNN, Naïve Bays, DT	Integrated Food Safety Information Network Data	By applying the supervised learning prediction model to the characteristic variable derived through the characteristic variable extraction method, a food hygiene inspection enforcement screening system is prepared and efficiency is improved
Marvin et al. (2016)	Bayesian Network	Food Fraud Database	Establish a Bayesian Network model for predicting food fraud based on food fraud database to detect food fraud types and identify risk factors

식품의 사후적 처리과정에서 발생가능한 식품 생산자에 대한 신용위험 및 경영위험을 완화시키고 자원낭비와 같은 사회적 비용을 줄일 수 있다(장동식, 이상호, 2016). 이처럼 식품 안전관리의 중요성이 높아지며 식품 부적합 탐지를 위한 연구가 이어지고 있다.

그동안 식품 안전 연구는 <Table 1>과 같이 식품 안전 영역의 빅데이터 활용에 대한 검토(Marvin et al., 2017; Jin et al, 2020)와 식품 사기 예측과 관련된 연구(Marvin et al., 2016)가 주를 이루었다. Marvin et al. (2016)은 Bayesian Network 모델로 식품 사기 유형을 탐지하고 위험 요인을 식별함으로써 식품 안전 개선에 기여하였다. 그러나 해당 연구는 식품 사기 사건에 대한 데이터 만으로 모델을 구축하여 식품 데이터의 다양한 특성을 충분히 반영하지 못했다. Marvin et al. (2017)은 방대한 양의 정형 및 비정형 데이터로 식품 안전 영역의 빅데이터 활용과 적용에 대한 평가 및 추세를 살펴보고, 개발된 빅데이터 도구 중 일부만이 식품 안전에 적용될 수 있음을 밝혔다. 식품 안전 영역에서의 빅데이터 활용이 아직 초기 단계에 머물러 있는 실정이다(Jin et al., 2020). 이처럼 식품 안전관리에 대한 중요성이 증대되고 있음에도 불구하고, 수입식품에 관한 데이터 분석 및 머신러닝 활용에 대한 연구는 부족한 실정이다. 수입식품 데이터에 적합한 머신러닝 알고리즘으로 수입식품 부적합 예측 모델을 구축하여 식품 안전을 도모하고 수입식품 검사 업무의 효율성을 개선할 필요가 있다.

2.2. 이상탐지 기계학습

부적합 식품 판정은 이상탐지(anomaly detection)의 영역에 속한다. 이상탐지는 비디오 이상 탐지(Zhao et al., 2017), 네트워크 이상 탐지 (Cui, 2016; Eltanbouly et al., 2020), 의료 정보 이상 탐지(Pachauri, 2015), 재무적 이상 탐지(Ahmed et al., 2016) 등 다양한 문제 영역에서 매우 중요한 이슈로 제기되어 왔다. 이상탐지는 통계적 기법이나 스펙트럼검출방법 등의 방법도 있으나 대

체적으로 기계학습 방법을 사용한다(Nassif et al., 2021).

지도기반 학습을 통하여 이상탐지 정확도를 올리기 위해서는 충분한 학습 데이터를 확보해야 한다 (Omar et al., 2013). 학습데이터가 충분하다면 심층신경망(Deep Neural Networks), 서포트 벡터 머신(Support Vector Machines), K-최근접 이웃 알고리즘(K-Nearest Neighbors), 베이지안 네트워크(Bayesian Networks), 의사결정나무(Decision Tree) 등이 통상적으로 사용되고 있다. K-평균 알고리즘(K-Mean), 퍼지 군집(Fuzzy C-Means) 등 비지도기반 학습으로도 이상탐지가 이루어지고 있으나, 일반적으로 라벨링이 된 학습데이터가 충분한 경우에는 지도기반 학습보다는 판별 성능이 떨어지는 경향이 있다.

이상탐지 방법은 일반화된 방법 및 특정 문제에 특화된 방법이 모두 가능하다. 이중 본 연구와 관련한 부적합 식품 판정은 데이터셋의 각 필드가 전문적이고 데이터불균형이 높아서 일반화된 방법으로 접근하게 되면 정확도 등 탐지 성능이 떨어지게 마련이다. 이에 부적합 식품판정을 위해서는 특화된 방법을 제안하는 경우가 많다. 한 예로 부적합 사례는 드물 것으로 가정하고, 문제가 발생할 경우를 기준으로 학습하는 베이지언 방법이 부적합 식품 판정에 자주 사용된다(Kleboth et al., 2022). 그러나 이러한 방법들은 매우 많은 컴퓨팅을 요구하거나 이상치 판정 중에 인간의 판단이 개입되어야 하므로 효율적이지 않다는 문제점을 가지고 있어, 다른 방법으로의 개선이 필요하다.

2.3. 데이터 불균형 처리

데이터 불균형 문제는 종속변수의 분포가 상당한 차이를 보이는 것으로 예측 성능 저하의 원인이 된다(Kaur et al., 2019; Kim et al., 2020; Kang 2021). 따라서 데이터가 불균형한 경우 샘플링 방법 또는 오분류 조정 등의 방안을 활용하여 문제를 해결해야 한다(김은미, 홍태호, 2015). 불균형 데이터 분류 문제를 해결하기 위한 방법으로는 크게 전처리 방법(Pre-processed Method), 알고리즘 기반 접근법(Algorithmic Centered Approaches), 하이브리드 접근법(Hybrid Approaches)이 있다(Kaur et al., 2019).

첫째, 전처리 방법에는 오버 샘플링(Oversampling)과 언더 샘플링(Undersampling) 방법이 있다. 이중 언더 샘플링은 다수 범주 데이터의 정보 인스턴스를 잃을 수 있다는 한계가 있어(Nguyen et al., 2012; Kaur et al., 2019), 랜덤 오버 샘플링이나 SMOTE(Synthetic Minority Over-Sampling Technique)과 같은 오버 샘플링 방법을 선호한다. 단, 랜덤 오버 샘플링은 소수 범주의 데이터의 반복적 복제를 통해 데이터의 크기를 증가시키므로 과적합의 원인이 되기도 한다(Ganganwar, 2012). 이에 비해 SMOTE는 소수 범주 데이터를 기반으로 합성 샘플을 생성한다(Chawla et al., 2002). 이러한 방법들은 예측 모형의 성능 개선에 대체로 도움을 주지만(Yap et al., 2014), 그 중 대체로 SMOTE방법이 가장 적합하다고 알려져 있다(Bach et al., 2017).

둘째, 알고리즘 중심 접근법은 불균형 데이터 분류 문제를 해결하는 알고리즘을 별도로 만들거나 기존 알고리즘을 업그레이드하는 것이다(Kuar et al., 2019). 종속 변수의 예측을 위해 오분류 비용을 할당하는 방법(Singh, Purohit, 2015), 언더 샘플링의 한계를 보완하기 위한 클러스터 기반 언더 샘플링(Cluster-based Undersampling) 방법(Zhang et al., 2010), 예측의 분산과 편향을

〈Table 2〉 Literature Review on Algorithm Centered Approach

Reference	Method	Comments
Kamei et al. (2007)	Logistic Regression	The effects of four sampling methods (random oversampling, SMOTE and cross-section selection, etc.) were evaluated by learning the logistic regression
Cieslak et al. (2005)	Decision Tree	Decision tree algorithms were used for severe unbalanced data due to data bias and high sparsity
Burez et al. (2008)	Random Forest	Unbalanced data were learned through undersampling and random forest, and weights were applied to minority classes, especially when learning random forest
Chomboon et al. (2013)	MLP	Tried to solve the data imbalance problem by learning multi-layer artificial neural networks and applying SMOTE techniques
Liu et al. (2021)	LightGMB	LightGBM was learned and ADASYN oversampled for the efficiency and accuracy of training and detection time on unbalanced network intrusion detection data
Guo et al. (2004)	XGBoost	To alleviate the problem of data imbalance, this research proposes a model that combines XGBoost and data augmentation techniques
Hancock et al. (2020)	CatBoost	Unbalanced Medicare raid detection was performed using CatBoost
Ntekouli et al. (2022)	EBM	Model learning using EBM and pattern analysis between variables were performed in existing individual modeling with mental illness data

줄이기위한 배깅(Bagging) 및 부스팅(Boosting) 알고리즘 등이 여기에 속한다. 일례로 Tanha et al.(2020)은 다수의 불균형 데이터셋을 대상으로 이진 및 다중 클래스 분류 부스팅 알고리즘을 적용함으로써 불균형 데이터 분류 문제 해결을 위한 전반적인 부스팅 알고리즘 기술을 검토하고 CatBoost 및 LogitBoost 알고리즘이 다른 부스팅 알고리즘에 비해 우수함을 입증하였다. 그 외에도 랜덤 포레스트(Burez et al., 2008), 나이브 베이즈, 인공신경망(Chomboon et al, 2013), LightGBM(Liu et al., 2013), XGBoost(Guo et al., 2004), CatBoost(Hancock et al., 2020), EBM(Ntekouli et al, 2022) 등 다양한 알고리즘을 활용한 불균형 데이터 문제 해결 연구가 <Table 2>와 같이 수행되어 왔다.

셋째, 하이브리드 접근법은 불균형 데이터 분류 문제를 효율적으로 처리하기 위해 전처리 방법과 알고리즘 중심 접근법을 결합한 방법이다(Kuar et al., 2019). 예를 들어, Abouelenien et al.(2013)은 클러스터 기반 샘플링 및 앙상블 알고리즘을 결합한 방법을 제안하고, 해당 방법이 예측 모형의 성능을 향상시킬 수 있음을 보였다. 한편, Jeong et al.(2018)은 불균형한 형태의 교통사고 데이터를 대상으로 언더 샘플링과 오버 샘플링을 적용하고 5가지 머신러닝 알고리즘을 기반으로 분류 모델을 구축하여 배깅 및 보팅

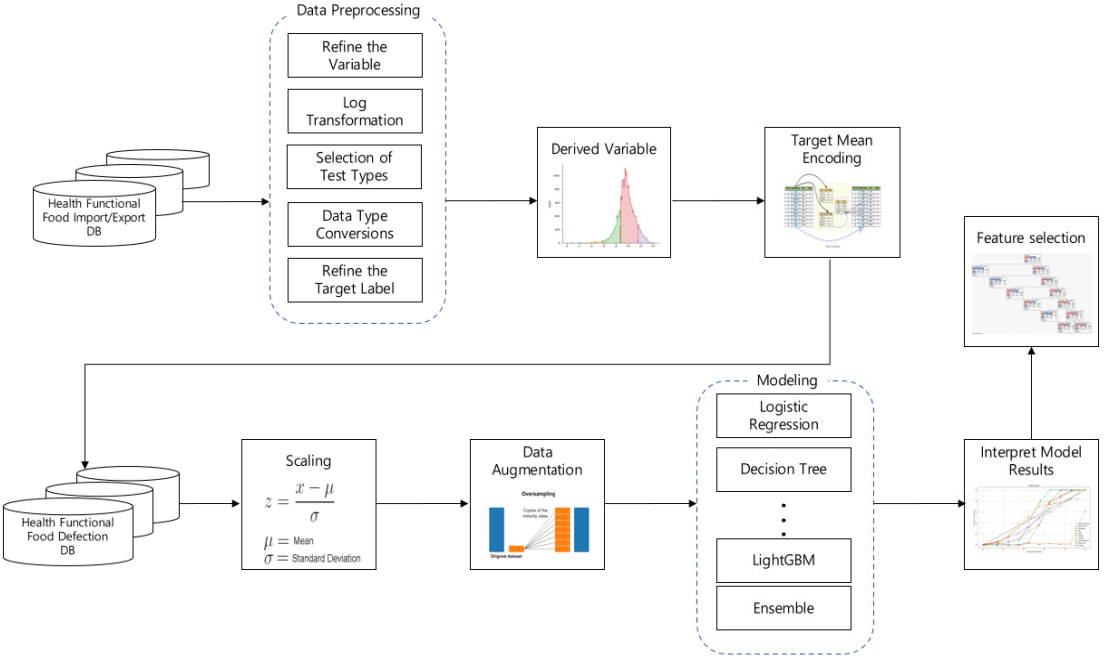
(Voting)을 통해 모델의 성능을 향상시키는 방안을 제안하였다. 그러나 해당 연구는 분류 모델의 성능을 향상시킬 수 있는 다양한 변수를 고려하지 않았다는 한계점을 지닌다.

3. 연구 방법

3.1. 분석절차

식품 수출입 검사 과정 중 부적합 건강기능식품을 탐지하기 위한 판정 방법을 <Figure 1>과 같이 제안하였다. 우선, 건강기능식품 수출입 데이터를 전처리하기 위하여 변수 정제, 로그 변환, 검사종류 선별, 데이터 형태 변환, 종속변수 정제 등을 수행하였다. 데이터 전처리가 끝난 후 모델의 성능을 올리기 위하여 파생변수(Derived Variable)를 생성한다. 파생변수 생성 과정에서는 기존에 존재하는 연속형 변수를 구간화 하거나 비율로 변환하고 범주형 변수를 상위 카테고리 범주형 변수로 치환하는 등 내부 파생변수를 만들 수 있다. 필요에 따라서는 공공데이터셋을 활용하여 외부 파생변수를 생성할 수도 있다.

본 연구에서는 타겟 평균 인코딩 방법(Target Mean Encoding)을 응용하여 건강기능식품 데이터 중 범주형 변수 값들을 부적합확률로 변환하였다. 타겟 평균 인코딩 방법은 범주형 변수의 카디널리티(Cardinality)가 높아 원핫 인코딩 등의 인코딩 방법이 제한될 때 사용하는 방법으로 알려져 있다(Pargent et al., 2022). 그러나 타겟 평균 인코딩 방법을 사용하는 것은 오버피팅을 유발할 수 있고 카테고리의 출현 빈도와 상관없이



〈Figure 1 〉 Analysis Process

평균으로만 계산되기 때문에 부적합률의 신뢰성에 문제가 생길 수 있다. 이러한 문제점을 해결하기 위해 부적합률의 치우쳐진 평균을 전체 평균으로 맞추는 Prior Probability for Regularization 규제 기법을 추가로 사용하여 오버피팅을 방지하고자 하였다. 또한 단순하게 타겟 평균 인코딩 방법만 활용할 경우 수입식품 10개 중 부적합 식품이 1개일 확률과 1000개 중 100개가 부적합 식품일 확률이 0.1로 동일하게 나오게 된다. 때문에 본 연구에서는 동일한 부적합 확률이더라도 수입 건수가 많을수록 가중치를 부여하기 위해 수입건수에 로그를 취한 후 부적합 확률과 곱셈을 하는 도식을 고안하였다. 해당 식은 아래와 같으며 $\tau(X)$, $\lambda(X)$는 각각 빈도와 부적합률을 의미한다.

$$x \rightarrow \Pr(x)$$
$$\Pr(x) = \tau(x) \times \log(\lambda(x))$$

타겟 평균 인코딩을 적용한 해당 데이터셋은 건강기능식품 부적합 탐지 데이터셋으로 재구축한 뒤 모델링을 위해 데이터 스케일링을 수행하여 각 변수들의 범위 혹은 분포를 동일하게 만들었다 (Ahsan et al., 2021). 본 연구에서는 대표적으로 값들의 최솟값을 0, 최댓값을 1로 스케일링을 하는 최소-최대 스케일링과(Min-Max Scaling) 중앙값을 0, 사분범위(IQR)가 1이 되도록 변환하며 이상치의 영향을 최소화할 수 있는 로버스트 스케일링(Robust Scaling), 평균을 0으로 잡고 표준편차를 1로 간주하여 정규화 하는 표준화 스케일링(Standard Scaling) 중 데이터셋의 특성에 맞추어 표준화 스케일링을 적용하였다.

또한 종속변수 분포의 불균형 해결을 위해 가장 우수한 성능을 보이는 SMOTE(Synthetic Minority Over-Sampling Technique) 방법을 적용한 후에 통계 모델과 머신러닝 모델을 개발하였다. 본 연구에서는 통계 모델과 머신러닝 모델의 알고리즘을 기반으로 성능을 입증한 기존의 연구(Kamei et al., 2007; Cieslak et al., 2005; Burez et al., 2008; Hulse et al., 2009; Chomboon et al, 2013; Liu et al., 2021; Guo et al., 2004; Hancock et al, 2020; Ntekouli et al., 2022)를 바탕으로 최적의 분석 모델을 선별하였으며, 그 결과 로지스틱 회귀, 일반화 선형 모델, 의사결정나무, 랜덤 포레스트, 나이브 베이즈, 인공신경망, LightGBM, XGBoost, CatBoost, EBM 등 총 10개의 분류 모델을 선정하였다. 그 외에 이들의 앙상블 모델도 개발하였다. 한편 K-교차검증을 통해 편향을 줄임으로써 과적합을 최대한 방지하였다 (Cawley, 2010).

3.2. 데이터 수집

본 연구에서 사용한 데이터는 2016년도부터 2020년까지의 수입신고 건을 대상으로 하였으며 식품 데이터 중 건강기능식품만을 대상으로 하였다. 활용한 건강기능식품 관련 수입신고 데이터는 2016년 23,463건, 2017년 18,213건, 2018년 23,616건, 2019년 26,187건, 2020년 47,963건, 2021년 46,732건 등 총 203,011건이다. 원본 데이터 셋의 변수 개수는 총 65개이며 (<Table 3> 참조), 이는 제품 기본정보(Product Basic Information), 제품 상세정보(Product Details), 검사정보(Inspection Information), 판정정보(Judgment Information), 선적정보(Shipping Information), 화주정보(Shipper Information) 등을 분류된다. 제품 기본정보에는 접수번호(Receipt Number), 제품명(Product Name), 품목명(Item Name) 등 제품에 관한 기본적인 정

〈Table 3〉 Classification of variables in the import declaration dataset

Classification	Variable Names
Product Basic Information	Receipt Number, Product Name, Item Name, etc
Product Details	Exam, Toddlers, Organic, etc
Inspection Information	Test, Test Type, Test Code, etc
Judgment Information	Judgment Result, Nonconformity Final Treatment Result Code, Nonconformity Action Plan, etc
Shipping Information	Port, Import Data, Exporter, etc
Shipper information	Import Shipper Name, Manufacturer, Manufacturer Region, etc

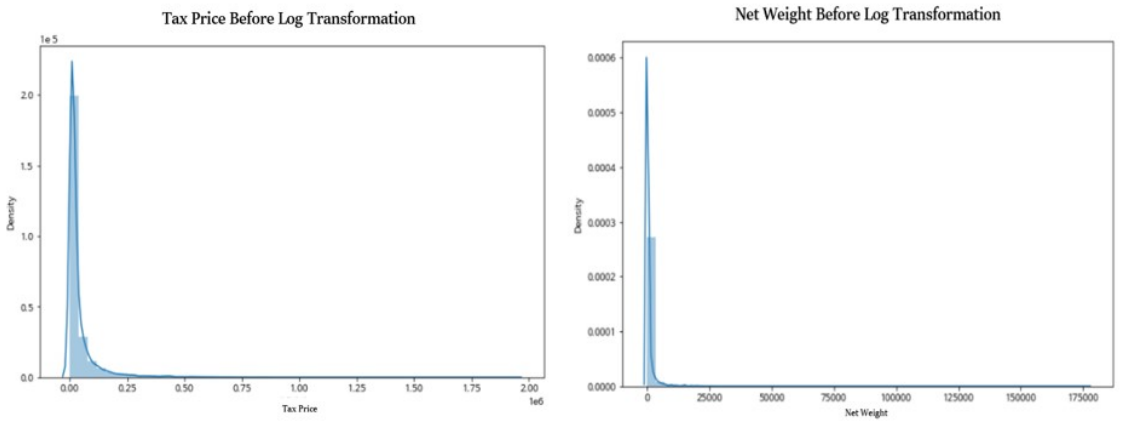
보를 나타내는 변수들로 구성된다. 제품 상세정보는 식품조사처리 여부(Exam), 영유아섭취대상(Toddlers), 유기식품여부(Organic) 등 제품에 관해 상세한 정보를 나타내는 변수들로 구성된다. 검사정보는 검사명령제도 해당여부(Test), 검사종류(Test Type), 검사코드(Test Code) 등 수출입 식품 검사를 하는데 있어 필요한 정보를 나타내는 변수들로 구성된다. 판정정보는 판정결과(Judgment Result), 부적합 최종처리 결과코드(Nonconformity Final Treatment Result Code), 부적합조치계획(Nonconformity Action Plan) 등 부적합을 판정하는데 있어 필요한 변수들로 구성된다. 이 중 판정결과는 부적합을 탐지하는 모델의 종속변수로 활용되는데 적합 201,198개(99.1%), 부적합 1,272개 (0.6%), 기타(0.3%)로 건강기능식품 데이터는 클래스 불균형이 상당히 심하다. 선적정보는 국내도착항명(Port), 반입일자(Import Date), 수출국(Exporter) 등 수출입 과정에서 발생하는 정보를 나타내는 변수들로 구성된다. 마지막 화주정보는 수입화주 상호(Import Shipper Name), 해외제조업소명(Manufacturer), 해외제조업소 지역(Manufacturer Region) 등 제품을 수출입 할 때 연관돼 있는 업소명에 관한 정보로 구성된다.

3.3. 데이터 전처리 및 파생변수 생성

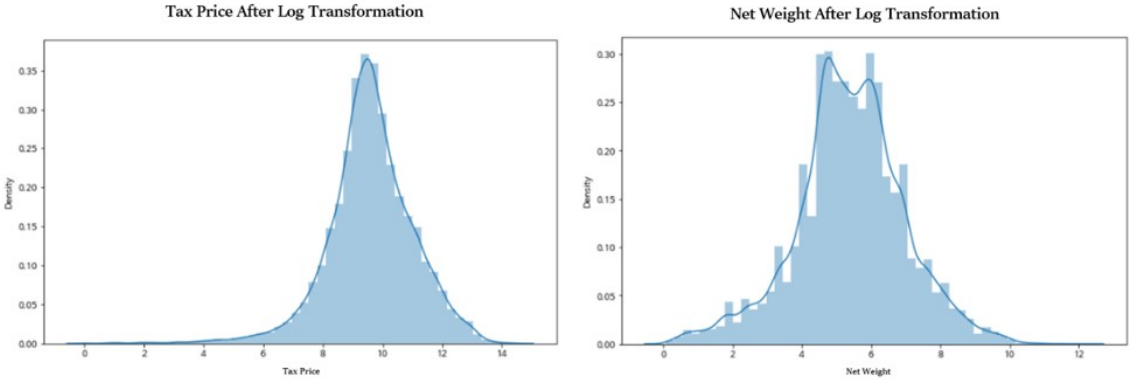
건강기능식품 데이터셋의 전처리로 변수 정제, 로그 변환, 종속변수 정제의 과정을 거쳤다. 첫째, 독립변인으로서 무의미하거나(예: 접수번호, 품목코드, 용도코드), 변수 간의 상관성이 높고(예: 수출국과 제조국), 변수의 결측치 비율이 높은 같은 변수(예: 부적합조치계획, 행정조치사유2, 등기번호 등)들을 제거했다. 이러한 과정을 거쳐 65개 변수 중 과세가격, 순중량, 수출국 등 19개 변수를 선정하였다.

둘째, 선정된 변수들 중 <Figure 2>와 같이 과세가격(Tax Price)과 순중량 (Net Weight) 변수가 로그 변환을 해야 되는 대상이다. 과세가격을 로그 변환할 경우 왜도가 -0.611, 첨도가 2.673이며, 순중량을 로그 변환할 경우 왜도가 -0.158, 첨도가 0.618로 모두 정규성 기준을 만족하였다. 로그 변환 후의 결과는 <Figure 3>과 같다.

마지막 전처리 과정은 종속변수 정제이다. 건강기능식품의 경우 종속변수가 적합, 부적합, 자진취하, 반려로 구성돼 있다. 자진취하는 수입신고를 개인사정에 의해 스스로 취소하는 경우이며 반려의 경우 수입검사원에 의해 반려된 경우이다. 이에 최종 종속변수는 적합과 부적합으로



〈Figure 2〉 Before Log Transformation



〈Figure 3 〉 After Log Transformation

만 선정하였으며, 적합과 부적합을 탐지하는 것은 이진분류 과업이기 때문에 0과1의 숫자형태로 인코딩 후 진행하였다.

전처리 완료 후 부적합 판정 성능을 올리기 위하여 <Table 4>와 같이 의미 있는 파생변수를 생성하였다. 생성한 파생변수로 원본 데이터셋으로 가공한 내부 파생변수와 공공데이터셋을 활용한 외부 파생변수가 있다. 첫째, 내부 파생변수로서 원본 데이터셋의 변수 중 과세가격과 순중량 등의 이산형 변수들을 구간화하여 과세가격 구간(저가, 보통, 고가), 순중량 구간(가벼움, 보통, 무거움) 등을 생성하였다. 또한 접수일 변수를 활용하여 시간대(새벽,아침,낮,저녁,밤)과 계절(봄, 여름, 가을, 겨울) 내부 파생변수를 생성하였으며, 수출국 변수를 사용하여 수출국을 대륙으로 재 분류하는 파생변수와 접수일과 선적일자의 차이를 수출국과의 거리로 나눠주는 거리 대비 소요일 파생변수도 생성하였다. 거리 대비 소요일 파생변수는 국가 간의 거리 대비 접수일이 상대적으로 늦어졌을 경우 음식의 변질이

〈Table 4〉 Derived Variable Description

Derived Variable	Comments
Price range	Binning the Tax Price Variable
Weight range	Binning the Net Weight Variable
Time Zone	Separating the Date of Receipt by Time Zone
Season	Separation of Reception Date by Season
Export Continent	Categorize the Exporting Country Variable by Continent
Required Days	Calculated as ((Receipt Date - Shipment Date) / Distance from Exporting Country)
Search Volume	Crawling Google Search Frequency for Defective Detection by Item

일어나 부적합 확률을 높일 수 있다는 것에 착안하여 생성하였다.

둘째, 외부 파생변수로 건강기능식품 데이터셋 내 품목명에 대한 구글 검색량을 선정했다. 이를 위해 건강기능식품의 품목 별로 '불량 적발' 키워드를 합성하여 구글 검색량을 도출하였다 (Durica & Svabova, 2015). 구글 검색 기능은 사회적으로 민감한 사건사고, 리뷰에 대한 양질의 데이터를 제공할 수 있어, 사회적으로 위협을 줄 수 있는 수입식품 위해 가능성을 탐지하는 것에 도움이 될 수 있다고 판단하였기 때문이다.

3.4. 모델링

일반화 선형 모형과 같은 통계 모델은 특성 가중치, 유의성 검사, 예측 구간 등의 신뢰 구간을 얻을 수 있고, 모델의 설명력이 높다는 장점이 있지만, 건강기능식품 데이터의 경우 이상값이 많고, 데이터의 비선형성이 의심되기 때문에 다른 기계학습도 함께 고려하였다. 우선 SMOTE로 데이터 증강을 한 후에 분류 문제에서 널리 활용되고 있는 로지스틱 회귀, 선형모형, 의사결정나무, 랜덤 포레스트, 나이브 베이즈, 인공신경망, LightGBM, XGBoost, CatBoost, EBM, 앙상블을 경쟁 모델로 선정하였다. 학습에는 2016~2019년 데이터, 검증에는 2020년 3분기까지의 데이터, 테스트에는 2020년 4분기 데이터를 사용하였다. 학습용 데이터에는 5-폴드 교차검증으로 성능을 최대한 일반화하였다.

4. 결과

4.1. 성능 비교

사용한 평가지표는 정확도, 재현율, 정밀도, F1-score 등이며 그 외에 학습 소요 시간도 고려하였다. 이중 부적합 탐지라는 본 연구의 목적에 따라 재현율이 가장 중요한 지표이다. 이외에도 클래스 불균형이 존재하는 분류 모델의 성능을 확인할 수 있는 ROC 커브, PR 커브, Calibration 플랏 등을 보조적으로 사용했다.

모델 성능 비교 결과 <Table 5>와 같은 혼동행렬과 이를 기반으로 <Table 6>을 얻을 수 있었다. 그 결과 재현율이 낮은 LightGBM, CatBoost와 재현율이 높지만 조화평균, 매튜 상관계수가 다른 모델에 비교해 현저히 낮고 부적합 식품으로 예측 가능한 기준을 큰 차이로 초과하는 나이브 베이즈의 경우 활용 가능한 모델의 요건에 부

〈Table 5〉 Confusion Matrix

BMT	PREDICTED NEGATIVE	PREDICTED POSITIVE
ACTUAL NEGATIVE	684	417
ACTUAL POSITIVE	20	9

GLM Model	PREDICTED NEGATIVE	PREDICTED POSITIVE
ACTUAL NEGATIVE	615	486
ACTUAL POSITIVE	2	27

Logistic Regression	PREDICTED NEGATIVE	PREDICTED POSITIVE
ACTUAL NEGATIVE	627	474
ACTUAL POSITIVE	1	28

Decision Tree	PREDICTED NEGATIVE	PREDICTED POSITIVE
ACTUAL NEGATIVE	881	220
ACTUAL POSITIVE	10	19

Random Forest	PREDICTED NEGATIVE	PREDICTED POSITIVE
ACTUAL NEGATIVE	910	191
ACTUAL POSITIVE	10	19

Naïve Bayes	PREDICTED NEGATIVE	PREDICTED POSITIVE
ACTUAL NEGATIVE	299	802
ACTUAL POSITIVE	1	28

MLP	PREDICTED NEGATIVE	PREDICTED POSITIVE
ACTUAL NEGATIVE	815	286
ACTUAL POSITIVE	8	21

LightGBM	PREDICTED NEGATIVE	PREDICTED POSITIVE
ACTUAL NEGATIVE	1021	80
ACTUAL POSITIVE	21	8

XGBoost	PREDICTED NEGATIVE	PREDICTED POSITIVE
ACTUAL NEGATIVE	780	321
ACTUAL POSITIVE	4	25

Cat Boost	PREDICTED NEGATIVE	PREDICTED POSITIVE
ACTUAL NEGATIVE	1044	57
ACTUAL POSITIVE	23	6

EBM	PREDICTED NEGATIVE	PREDICTED POSITIVE
ACTUAL NEGATIVE	919	182
ACTUAL POSITIVE	16	13

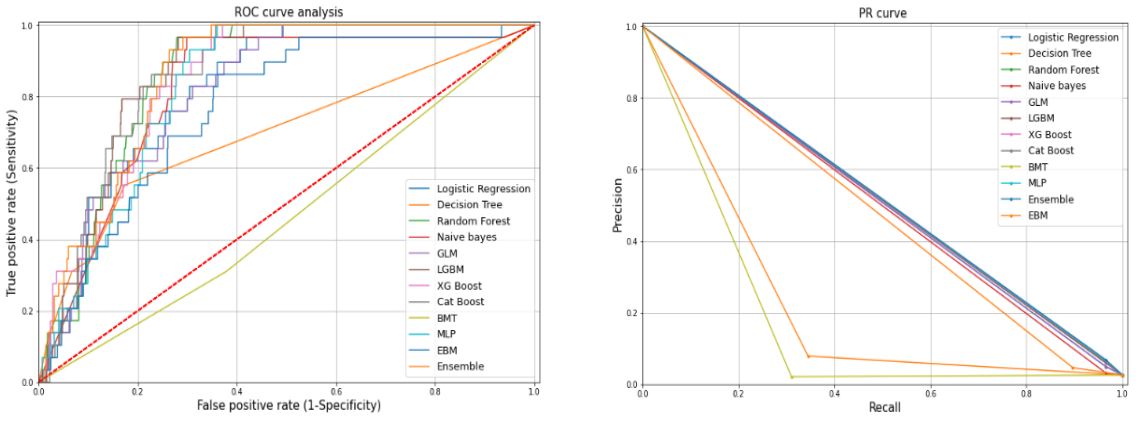
Ensemble	PREDICTED NEGATIVE	PREDICTED POSITIVE
ACTUAL NEGATIVE	683	418
ACTUAL POSITIVE	0	29

〈Table 6〉 Model performances

Model	AUC	Accuracy	Precision	Recall	F1-score	MCC	Elapsed Time (min)	Explainability
BMT	0.46	0.61	0.02	0.31	0.03	-0.02	-	High
LightGBM	0.86	0.91	0.09	0.27	0.13	0.12	1	Fair
Logistic regression	0.82	0.58	0.05	0.96	0.1	0.17	1	High
XG Boost	0.84	0.71	0.07	0.86	0.13	0.19	1	Fair
GLM	0.82	0.56	0.05	0.93	0.09	0.15	1	High
Random forest	0.85	0.82	0.09	0.65	0.15	0.19	1	High
Cat Boost	0.86	0.92	0.09	0.2	0.13	0.1	1	Fair
EBM	0.76	0.82	0.06	0.44	0.11	0.11	15	High
MLP	0.83	0.73	0.06	0.72	0.12	0.16	5	Low
Naïve Bayes	0.82	0.28	0.03	0.96	0.06	0.08	1	Low
Decision tree	0.72	0.79	0.07	0.65	0.14	0.17	1	High
Ensemble	0.85	0.63	0.06	1	0.12	0.2	10	Fair

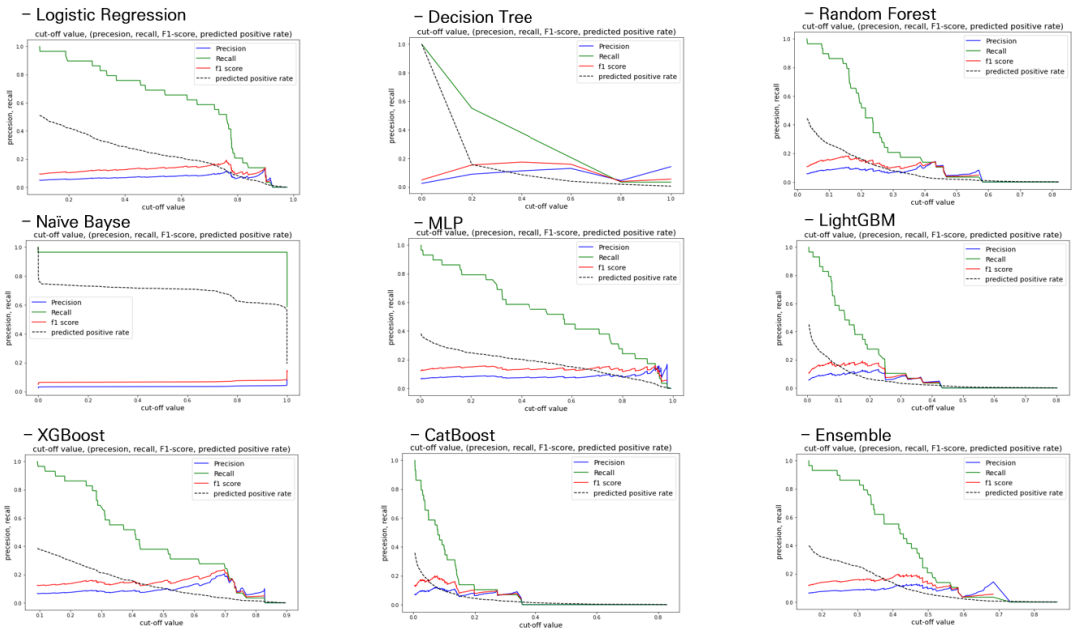
합하지 못하였다. 이와는 반대로 재현율이 높으며 예측 가능 개수가 기준에 충족하여 부적합 탐지 모델로 활용 가능한 것은 로지스틱 회귀, XGBoost, 일반화 선형 모형, 앙상블인 것으로 나타난다.

분류 성능의 시각화로 <Figure 4>와 같이 True Positive 비율과 False Positive 비율을 조정한 ROC 커브와 PR 커브를 보였다. 일반적으로 ROC 커브에서 우수한 모델은 곡선의 위치가 왼쪽 위 모서리에 가까워지며 PR커브의 경우 모델



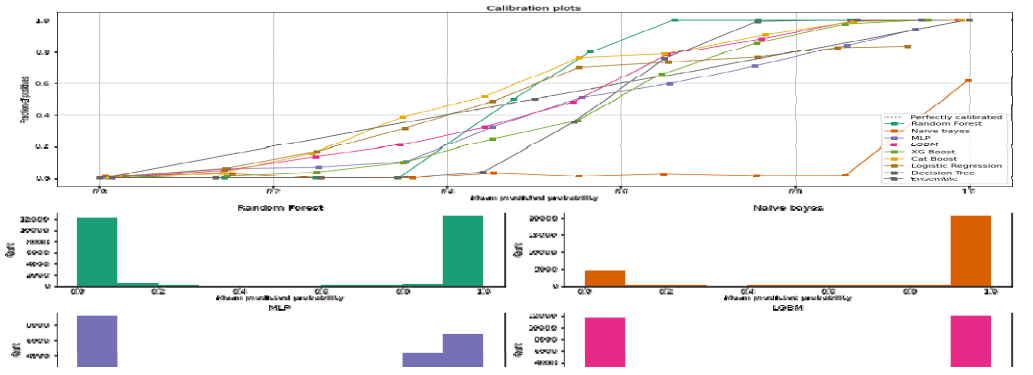
(a) ROC curve (b) PR curve

〈Figure 4〉 Modified Research Model



〈Figure 5〉 Cut-off value plots

의 정밀도와 재현율이 모두 우수하다면 곡선이 오른쪽 위 모서리에 가깝게 된다. 그러나 본 연구는 부적합 식품 예측 개수의 기준에 맞춤으로써 정밀도가 떨어지는 것을 감안하고 재현율을 최대화하는 모델이 바람직했기 때문에 PR 커브는 45도 형태의 일직선 곡선이 이상적이다. 때문에 PR커브 곡선이 일직선 형태가 아닌 BMT와 의사결정나무 모델은 다른 모델에 비해 성능이



〈Figure 6〉 Calibration Plot

좋지 않다고 판단하였다. 두 모델은 ROC 커브에서도 낮은 성능을 보여준다.

학습할 때 임계값(Cut-off Value)은 일반적으로 0.5로 초기화한다. 그러나 만약 재현율 또는 정밀도와 같은 특정 지표에 조금 더 가중치를 두려면 임계값을 상하로 조정할 수 있다. 일반적으로 조화평균이 가장 높은 지점 또는 재현율과 정밀도가 교차하는 지점이 평균적으로 모델 일반화에 도움이 되는 것으로 알려져 있다. 본 연구와 같이 재현율을 중시하는 경우, 임계값을 0으로 가져갈수록 검출대상으로 판단하는 기준이 낮아지면서 재현율이 높아지는 경향이 있다. 이에 임계값 조정에 따른 성능 지표 변화를 <Figure 5>와 같이 살펴보았다. 그 결과 나이브 베이즈의 경우 임계값 수준에 상관없이 재현율이 1로 나타났다. 즉, 부적합을 탐지하기 위해 최대한 많은 예측을 양성으로 판단한 것이다. 그러나 재현율만 고려했기 때문에 정밀도가 0에 가까워지면서 불필요한 양성 판단 결과가 증가했다. 따라서 연구 목적에 맞게 재현율과 동시에 조화평균과 매튜 상관계수와 같은 균형 잡힌 지표 또한 고려하여 모델을 선택했다.

예측 모형은 혼동행렬을 통한 분류 성능으로 평가할 수 있지만, 좋은 모델이란 정확해야 할 뿐만 아니라 잘 보정(Calibration)되어야 할 필요가 있다(Niculescu-Mizil et al., 2005). 보정을 평가하기 위해 <Figure 6>과 같은 Calibration 플랏은 예측된 확률과 실제 확률의 관계를 보여준다. Calibration 플랏의 경우 가장 성능이 좋은 형태는 선이 기울기가 1인 것이다. 즉, 실제 라벨의 비율과 마찬가지로 예측값의 비율이 똑같다는 뜻이다. 그 결과, 가장 재현율이 좋은 로지스틱 회귀와 앙상블 선분의 기울기가 1에 가장 가깝게 나타나 우수한 모델인 것으로 판정하였다.

4.2. 변수 선택 결과 분석

기계학습 모델의 성능을 높이기 위해서는 변수 선택이 중요하다(Tsamardinos et al., 2003). 본 연구에서 원천 데이터와 이를 활용하여 추가로 만들어낸 파생변수를 모두 포함하여 27가지의 변수들 중에서, 일반화선형모형(GLM)을 이용해 유의확률(p-value) 0.05를 기준으로 임계치보다 높은 변수를 제거하여 16개로 간소화하였다. 이때 베이스라인과 변수 선별 모델의 분류 성능은 <Table 7>과 같다. 정확도는 베이스라인이 앞서며 변수 선별 모델이 비교적 적합의 사례를 분류

〈Table 7〉 Effect of feature selection

Metrics	Baseline	Feature selection
Accuracy	**0.588**	0.575
Precision	0.055	0.055
Recall	0.931	**0.966**
F1-score	0.104	0.104
AUC	0.821	**0.831**
Cohen-Kappa score	0.058	**0.059**
Matthew's Correlation Coefficient	0.163	**0.169**

해내지 못했지만, 정밀도는 서로 동일하다. 특히, 본 연구에서 가장 무게를 둔 정밀도에서는 변수 선정 모델이 0.035만큼 더 높았다. 게다가, 정밀도와 재현율을 균형 있게 고려하는 코헨의 카파 계수(Cohen-Kappa score), 그리고 매튜 상관계수(Matthew's Correlation Coefficient)까지 변수 선별 모델이 더 높은 것으로 나타났다. 변수 선별 모델은 모델 복잡도와 학습 소요 시간이 줄어들었음에도 오히려 부적합의 사례를 더 많이 적발했다. 일반적으로 분류 문제에서는 정확도가 중요한 지표로 쓰이지만, 본 연구의 목적과 재현율과 정밀도와 같은 분류 지표를 모두 고려하는 관점에서 변수 선별 모델이 성능과 효율성이 더욱 높다.

5. 토의 및 결론

5.1. 시사점

본 연구에서 도출된 주요 결과는 다음과 같다. 첫째, 클래스 불균형이 심한 건강기능식품 데이터셋의 경우에도 부적합 식품을 탐지할 수 있는 기계학습 모델 기반 자동화 시스템 설계 방안을 제시하여 연구에서 제안하는 파생변수 및 모델이 수출입 식품 검사 과정에서 활용하고 있는 시스템에 도움이 될 수 있을 것이다. 이는 건강기능식품 데이터 특성에 맞는 파생 변수의 개발과 불균형 데이터셋에 특화된 학습 알고리즘 선택에 의한 것이었다. 특히 앙상블 모델의 경우 다른 경쟁 모델보다 성능이 우수하였다. 또한 변수 중요도를 통해 변수별로 부적합 판정에의 기여도를 파악함으로써 효율적이고 정확하게 부적합 식품을 탐지하였다.

둘째, 부적합률을 이용한 데이터 인코딩 방법은 성능 향상에 큰 역할을 하였다. 수입화주나 제조업소명과 같이 성능에 영향을 많이 끼치는 변수 중 카디널리티가 높아 일반적인 인코딩 방식으로는 수행할 수 없었던 문제점을 타겟 평균 인코딩 방법을 사용하여 해당 변수를 부적합률로 활용함으로써 성능 향상에 큰 기여를 하였다.

셋째, 구글 검색량과 같은 공공데이터를 크롤링을 통해 파생변수화 시킴으로써 예측 성능을 향상시킬 수 있었다. 검색량을 통해 생성한 변수는 변수 중요도를 통해 영향력이 높은 변수 중 하나임이 입증되었다. 내부 데이터셋과 외부의 공개 데이터셋을 조화할 때 성능이 제고되며, 특히 구글과 같은 뉴스 데이터에는 판정을 위한 지식이 내포되어 있어서 본 연구 문제와 같은 예측

성능에는 유의한 도움이 된다.

5.2. 한계점

본 연구의 한계점은 다음과 같다. 첫째, 본 연구에서는 아직 딥러닝 모델을 비교해보지 않았다. 특히 건강기능식품 데이터셋이 시계열 데이터인 만큼, LSTM(Long Short Term Memory) 계열 모델을 활용해 볼 수 있다. 둘째, 데이터 증강 방법 중 딥 오토인코더(Deep Autoencoder) 및 생성적 적대 신경망(Generative Adversarial Networks)을 활용하는 등 다양한 방법들이 최근에 많이 연구되고 있으므로 성능 향상을 위해 적용해 볼 수 있을 것이다. 셋째, 본 연구에서는 종속변수 카테고리 중 자진취하와 반려를 부적합으로 처리해야 할지 전문가와 논의가 아직 이루어지지 않았으며, 파생변수를 생성할 때도 전문가와 심도 깊은 논의 및 검증 없이 변수를 생성하였다. 이러한 문제점들을 보완한다면 추가적인 성능 향상은 물론이며 다른 식품 제품구분에 적용할 때에도 무난한 성능이 나오는 결과를 기대할 수 있을 것이다. 마지막으로, 본 모델의 개발은 부적합 식품 탐지 담당자들에게 유의한 탐지 지식을 생성하는 것이다. 추후에는 기계학습 모델에서 생성된 판정 지식과 인간이 담당자들이 보유하고 있는 경험 지식이 서로 상보하고 공진화하는 방법론을 개발할 예정이다.

참고문헌(References)

[국내 문헌]

김은미, & 홍태호. (2015). 불균형 데이터 환경에서 변수가중치를 적용한 사례기반추론 기반의 고객반응 예측. 지능정보연구, *21*(1), 29-45.

장동식, 이상호. (2016). 미국의 수입식품안전관리시스템 분석-가공식품을 중심으로. 국제상학, 31(4), 325-350.

조상구, 조승용. (2020). 기계학습을 이용한 식품 위생점검 체계의 효율성 개선 연구. 한국빅데이터학회지, 5(2), 53-67.

조상구, 최경현. (2018). 수입식품 빅데이터를 이용한 부적합식품 탐지 시스템에 관한 연구. 한국빅데이터학회지, 3(2), 19-33.

[국외 문헌]

Abouelenien, M., Yuan, X., Giritharan, B., Liu, J., & Tang, S. (2013). Cluster-based sampling and ensemble for bleeding detection in capsule endoscopy videos. *American Journal of Science and Engineering*, *2*(1), 24-32.

Ahmed, M., Mahmood, A. N., & Islam, M. R. (2016). A survey of anomaly detection techniques in financial domain. *Future Generation Computer Systems*, *55*, 278-288.

Ahsan, M. M., Mahmud, M. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. *Technologies*, *9*(3), 52.

Bach, M., Werner, A., Żywiec, J., & Pluskiewicz, W. (2017). The study of under-and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis. *Information Sciences*, *384*, 174-190.

Burez, J., & Van den Poel, D. (2009). Handling class imbalance in customer churn prediction.

Expert Systems with Applications, *36*(3), 4626-4636.

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. The Journal of Machine Learning Research, 11, 2079-2107.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Chomboon, K., Kerdprasop, K., & Kerdprasop, N. (2013). Rare class discovery techniques for highly imbalance data. In *Proc. International multi conference of engineers and computer scientists* (Vol. 1).

Cieslak, D. A., & Chawla, N. V. (2008, September). Learning decision trees for unbalanced data. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 241-256). Springer, Berlin, Heidelberg.

Cui, B., & He, S. (2016, July). Anomaly detection model based on hadoop platform and weka interface. In *2016 10th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS)* (pp. 84-89). IEEE.

Durica, M., & Svabova, L. (2015). Improvement of company marketing strategy based on Google search results analysis. *Procedia Economics and Finance*, *26*, 454-460.

Eltanbouly, S., Bashendy, M., AlNaimi, N., Chkirbene, Z., & Erbad, A. (2020, February). Machine learning techniques for network anomaly detection: A survey. In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)* (pp. 156-162). IEEE.

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, *2*(4), 42-47.

GFSI - what we do, https://mygfsi.com/what-we-do/harmonisation/, 2022.

Guo, H., & Viktor, H. L. (2004). Learning from imbalanced data sets with boosting and data generation: the databoost-im approach. *ACM Sigkdd Explorations Newsletter*, *6*(1), 30-39.

Hancock, J., & Khoshgoftaar, T. M. (2020, August). Medicare fraud detection using catboost. In *2020 IEEE 21st international conference on information reuse and integration for data science (IRI)* (pp. 97-103). IEEE.

Jeong, H., Jang, Y., Bowman, P. J., & Masoud, N. (2018). Classification of motor vehicle crash injury severity: A hybrid approach for imbalanced data. *Accident Analysis & Prevention*, *120*, 250-261.

Jin, C., Bouzembrak, Y., Zhou, J., Liang, Q., Van Den Bulk, L. M., Gavai, A., ... & Marvin, H. J. (2020). Big Data in food safety-A review. *Current Opinion in Food Science*, *36*, 24-32.

Kamei, Y., Monden, A., Matsumoto, S., Kakimoto, T., & Matsumoto, K. I. (2007, September). The effects of over and under sampling on fault-prone module detection. In *First international symposium on empirical software engineering and measurement (ESEM 2007)* (pp. 196-204). IEEE.

Kang, S., & Shin, K. S. (2021). Conditional generative adversarial network based collaborative filtering recommendation system. *Journal of Intelligence and Information Systems*, 27(3), 157-173.

Kaur, H., Pannu, H. S., & Malhi, A. K. (2019). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Computing Surveys (CSUR)*, *52*(4), 1-36.

Kim, J., Kim, M. Y., & Kwon, O. (2020). The effect of meta-features of multiclass datasets on the performance of classification algorithms. *Journal of Intelligence and Information Systems*, 26(1), 23-45.

Kleboth, J. A., Kosorus, H., Rechberger, T., & Luning, P. A. (2022). Using data mining as a tool for anomaly detection in food safety audit data. *Food Control*, *138*, 109004.

Liu, J., Gao, Y., & Hu, F. (2021). A fast network intrusion detection system using adaptive synthetic oversampling and LightGBM. *Computers & Security*, *106*, 102289.

Marvin, H. J., Bouzembrak, Y., Janssen, E. M., van der Fels-Klerx, H. V., van Asselt, E. D., & Kleter, G. A. (2016). A holistic approach to food safety risks: Food fraud as an example. *Food research international*, *89*, 463-470.

Marvin, H. J., Janssen, E. M., Bouzembrak, Y., Hendriksen, P. J., & Staats, M. (2017). Big data in food safety: An overview. *Critical reviews in food science and nutrition*, *57*(11), 2286-2295.

Nassif, A. B., Talib, M. A., Nasir, Q., & Dakalbab, F. M. (2021). Machine learning for anomaly detection: A systematic review. *Ieee Access*, *9*, 78658-78700.

Nguyen, H. M., Cooper, E. W., & Kamei, K. (2012, November). A comparative study on sampling techniques for handling class imbalance in streaming data. In *The 6th International Conference on Soft Computing and Intelligent Systems, and The 13th International Symposium on Advanced Intelligence Systems* (pp. 1762-1767). IEEE.

Niculescu-Mizil, A., & Caruana, R. (2005, August). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning* (pp. 625-632).

Ntekouli, M., Spanakis, G., Waldorp, L., & Roefs, A. (2022, April). Using Explainable Boosting Machine to Compare Idiographic and Nomothetic Approaches for Ecological Momentary Assessment Data. In *International Symposium on Intelligent Data Analysis* (pp. 199-211). Springer, Cham.

Omar, S., Ngadi, A., & Jebur, H. H. (2013). Machine learning techniques for anomaly detection: an overview. *International Journal of Computer Applications*, *79*(2).

Pachauri, G., & Sharma, S. (2015). Anomaly detection in medical wireless sensor networks using machine learning algorithms. *Procedia Computer Science*, *70*, 325-333.

Pargent, F., Pfisterer, F., Thomas, J., & Bischl, B. (2022). Regularized target encoding outperforms traditional methods in supervised machine learning with high cardinality features. *Computational Statistics*, 1-22.

Sharif, A., Abbasi, Q. H., Arshad, K., Ansari, S., Ali, M. Z., Kaur, J., ... & Imran, M. A. (2021). Machine learning enabled food contamination detection using RFID and internet of things system. *Journal of Sensor and Actuator Networks*, *10*(4), 63.

Singh, A., & Purohit, A. (2015). A survey on methods for solving data imbalance problem for classification. *International Journal of Computer Applications*, *127*(15), 37-41.

Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., & Asadpour, M. (2020). Boosting methods for multi-class imbalanced data classification: an experimental review. *Journal of Big Data*, *7*(1), 1-47.

Tsamardinos, I., & Aliferis, C. F. (2003, January). Towards principled feature selection: Relevancy, filters and wrappers. In *International Workshop on Artificial Intelligence and Statistics* (pp. 300-307). PMLR.

Wu, L., Liu, Z., Bera, T., Ding, H., Langley, D. A., Jenkins-Barnes, A., ... & Xu, J. (2019). A deep learning model to recognize food contaminating beetle species based on elytra fragments. *Computers and Electronics in Agriculture*, *166*, 105002.

Yap, B. W., Rani, K. A., Rahman, H. A. A., Fong, S., Khairudin, Z., & Abdullah, N. N. (2014). An application of oversampling, undersampling, bagging and boosting in handling imbalanced datasets. In *Proceedings of the first international conference on advanced data and information engineering (DaEng-2013)* (pp. 13-22). Springer, Singapore.

Zhang, Y. P., Zhang, L. N., & Wang, Y. C. (2010, September). Cluster-based majority under-sampling approaches for class imbalance learning. In *2010 2nd IEEE International Conference on Information and Financial Engineering* (pp. 400-404). IEEE.

Zhao, Y., Deng, B., Shen, C., Liu, Y., Lu, H., & Hua, X. S. (2017, October). Spatio-temporal autoencoder for video anomaly detection. In *Proceedings of the 25th ACM international conference on Multimedia* (pp. 1933-1941).

Abstract

A Method of Machine Learning-based Defective Health Functional Food Detection System for Efficient Inspection of Imported Food

Kyoungsu Lee*, Yerin Bak*, Yoonjong Shin**, Kwonsang Sohn**, Ohbyung Kwon***

As interest in health functional foods has increased since COVID-19, the importance of imported food safety inspections is growing. However, in contrast to the annual increase in imports of health functional foods, the budget and manpower required for inspections for import and export are reaching their limit. Hence, the purpose of this study is to propose a machine learning model that efficiently detects unsuitable food suitable for the characteristics of data possessed by government offices on imported food. First, the components of food import/export inspections data that affect the judgment of nonconformity were examined and derived variables were newly created. Second, in order to select features for the machine learning, class imbalance and nonlinearity were considered when performing exploratory analysis on imported food-related data. Third, we try to compare the performance and interpretability of each model by applying various machine learning techniques. In particular, the ensemble model was the best, and it was confirmed that the derived variables and models proposed in this study can be helpful to the system used in import/export inspections.

Key Words : Imported Food, Health Functional Food, Food Safety, Defective Detection, Machine Learning, Data Imbalance

Received : Augus 22, 2022 Revised : September 8, 2022 Accepted : September 15, 2022
Corresponding Author : Ohbyung Kwon

* Department of Big Data Analytics, Kyung Hee University
** School of Management, Kyung Hee University
*** Corresponding author: Ohbyung Kwon
School of Management, Kyung Hee University
26 Kyungheedae-ro, Dongdaemun-gu, Seoul, 02447, Korea
Tel: +82-2-961-2148, Fax: +82-2-961-0515, E-mail : obkwon@khu.ac.kr

저 자 소 개

이 경 수
현재 경희대학교 빅데이터 응용학과 석사과정에 재학중이다. 한남대학교에서 경영과 빅데이터전공 학사 학위를 취득하였고, 2019년 (주)코난테크놀로지 인공지능개발팀에 입사하여 1년 4개월간 연구원으로 재직하였다. 관심분야는 데이터마이닝, 텍스트마이닝 등이다.

박 예 린
현재 경희대학교 빅데이터 응용학과 석사과정에 재학중이다. 한성대학교에서 무역과 CRM · 디지털마케팅 학사 학위를 취득하였다. 관심분야는 빅데이터 애널리틱스, 비즈니스 애널리틱스, CRM, 개인정보보호 등이다.

신 윤 종
현재 경희대학교 경영학과 석사과정에 재학중이다. 한신대학교에서 IT경영학과 컴퓨터공학 학사 학위를 취득하였다. 관심분야는 자연어 처리, 그래프 마이닝, 추천 시스템, 등이다.

손 권 상
인하대학교에서 국제통상학 학사학위와 경영학 석사학위를 취득하였고, 경희대학교 경영학과에서 빅데이터경영 전공으로 박사학위를 취득하였다. 현재 한국능률협회컨설팅 데이터사업본부 컨설턴트 및 인하대학교 경영학과 강사로 재직 중이며, 관심분야는 AI 기반 의사결정지원, 비즈니스 애널리틱스, 디지털 트랜스포메이션 등이다.

권 오 병
현재 경희대학교 경영학과 및 빅데이터응용학과 교수로 재직 중이다. 서울대학교 경영학 학사학위와 한국과학기술원에서 석사 및 박사학위를 취득하였고, 카네기멜론대학 ISRI연구소에서 유비쿼터스 컴퓨팅 프로젝트를 수행한 바 있다. 관심분야는 AI비즈니스, 텍스트 분석, 휴먼로봇 인터페이스, 상황인식 서비스, 의사결정지원시스템 등이다.