

머신러닝 기법을 활용한 수입 수산물 통관검사결과 분류 모델

A Classification Model for Customs Clearance Inspection Results of Imported Aquatic Products Using Machine Learning Techniques

엄지성¹ · 이경희² · 조완섭^{2*}

환경부¹, 충북대학교 경영정보학과²

요약

수산물은 많은 국가의 주요한 단백질 공급원이며 소비가 증가하고 있다. 우리나라도 수산물 소비는 증가하는 반면 자급률은 낮아지고 있으며, 수산물의 수입량이 증가함에 따라 안전관리의 중요성이 높아지고 있다. 국내로 수입되는 수산물은 110여 개 국가로부터 수백 종에 이르며, 수입 수산물의 안전관리를 검사관의 경험에만 의존하는 것은 한계가 있다. 데이터를 기반으로 수입 수산물 통관검사 결과를 예측할 수 있는 모델을 개발하여 수입신고서가 제출되었을 때 수산물의 부적합 가능성을 판단하는 머신러닝 분류 모델을 생성한다.

수입수산물 통관검사 결과 부적합율은 1% 미만으로 매우 낮은 불균형 데이터이므로 이러한 특성을 보완할 수 있는 샘플링 방법을 비교 연구하였고, 분류결과를 해석할 수 있는 전처리 방법을 적용하였다. 여러 가지 머신러닝 기반의 분류모델 중에서 랜덤포레스트와 XGBoost가 좋은 성능을 보였다. 통관검사 결과 적합과 부적합을 모두 잘 예측하는 모델은 ADASYN과 원-핫 인코딩을 적용한 랜덤포레스트 기본 모델이며 정확도 99.88%, 정밀도 99.87%, 재현율 99.89%, AUC 99.88%이다. XGBoost는 오버샘플링과 인코딩 종류에 상관없이 모든 지표가 90%를 넘겨 가장 안정적인 모델이다.

■ 중심어 : 수입수산물, 통관검사, 분류모델, 머신러닝

Abstract

Seafood is a major source of protein in many countries and its consumption is increasing. In Korea, consumption of seafood is increasing, but self-sufficiency rate is decreasing, and the importance of safety management is increasing as the amount of imported seafood increases.

There are hundreds of species of aquatic products imported into Korea from over 110 countries, and there is a limit to relying only on the experience of inspectors for safety management of imported aquatic products. Based on the data, a model that can predict the customs inspection results of imported aquatic products is developed, and a machine learning classification model that determines the non-conformity of aquatic products when an import declaration is submitted is created.

As a result of customs inspection of imported marine products, the nonconformity rate is less than 1%, which is very low imbalanced data. Therefore, a sampling method that can complement these characteristics was comparatively studied, and a preprocessing method that can interpret the classification result was applied. Among various machine learning-based classification models, Random Forest and XGBoost showed good performance. The model that predicts both compliance and non-conformance well as a result of the clearance inspection is the basic random forest model to which ADASYN

and one-hot encoding are applied, and has an accuracy of 99.88%, precision of 99.87%, recall of 99.89%, and AUC of 99.88%. XGBoost is the most stable model with all indicators exceeding 90% regardless of oversampling and encoding type.

■ Keyword : imported aquatic products, Clearance test, classification model, machine learning

I. 서론

수산물은 많은 국가의 식단에서 공급되는 단백질 영양소로 우리 일상의 영양과 균형 잡힌 식단에 중요한 역할을 하고 있다(FAO 2021). 한국 농촌경제연구원의 보고서에 따르면 연간 ‘1인당 수산물 소비량과 자급율’ 변화를 보면 2001년 52.8kg에서 2020년 68.4kg으로 15.6kg 증가하였다[1]. 반면에 수산물 자급률은 2012년 76.5%에서 2019년 68.4%로 감소하였다[1]. 이로 인해 수입수산물의 수요는 증가했고, 수입수산물 안전관리의 중요성이 더욱 높아졌다.

식약처 자료에 의하면 2022년 기준 110여 개 국가에서 수산물을 수입하고 있다. 우리나라의 수입 수산물 신고건수는 2018년부터 2021년까지 389,389건으로 연 평균 약 97만 건 이상이 수입되고 있다[2]. 수입 신고된 모든 수산물을 검사하기에는 어려운 실정으로 통관검사결과가 부적합인 수산물의 특성(품종, 품명, 수입시기, 유통방식 등) 데이터를 기반으로 한 분류모델 구축이 필요하다. 통관검사 결과 부적합 확률이 높은 경우 통관전단계 서류를 보완하거나 통관단계 검사 방식에 정밀한 식품검사로 지정하는 등 선제적인 안전관리 방법 적용이 가능할 것이다.

수입식품 수요가 증가함에 따라 2016년 2월부터 『수입식품안전관리특별법』의 시행으로 수산물, 축산물, 농산물, 가공식품, 기구용기포장재 등 모든 수입식품 안전관리를 식품의약품안전처에서 담당하고 있다. 식품은 사람의 건강과 생명에 직접적인 영향을 미치기 때문에 수입식품 통

관전단계, 통관단계, 유통단계 별로 각종 검사를 두어 수입식품안전관리를 시행하고 있다.

본 연구는 수입 식품 중 수산물에 한정하여 식품의약품안전처의 수입식품 신고데이터와 통관 검사 결과 데이터를 다각도로 분석한다. 다빈도 수입국과 다소비 수입 수산물의 통관검사 결과 분류모델을 구축하고 통관검사 결과가 부적합으로 분류되는 수입 수산물의 부적합 영향 요인을 파악하여 수입수산물 안전관리 정책 결정에 활용하고자 한다. 특히, 수입식품 통관검사 결과는 적합과 부적합 빈도는 99.94%:0.06%로 매우 불균형하기 때문에 불균형이 심한 데이터의 분류모델을 개발방법을 연구하였다. 불균형데이터를 그대로 활용한 분류모델은 극소수 샘플을 가진 데이터 분류가 안 되기 때문에 균형데이터로 샘플링 후 모델을 구축해야 한다. 샘플링 방법은 랜덤언더샘플링, 랜덤오버샘플링, SMOTE, ADASYN 4가지로 샘플링 후 변수를 그대로 사용한 기본 모델과 하이퍼파라미터 튜닝 모델 2가지를 적용하고 분류 알고리즘 3종을 적용한 전체 24가지 모델을 구축하고 성능을 비교하였다. 심한 불균형 데이터를 활용한 모델이기 때문에 모델 성능은 정확도, 재현율, 정밀도, AUC를 서로 비교하였다. 이 중 성능이 좋은 모델 4개를 선택한 다음 혼동행렬로 FPV (False Positive Value), FNV (False Negative Value)를 비교한 결과 데이터 샘플링은 랜덤오버샘플링방법, 변수는 하이퍼파라미터 튜닝을 하지 않고 그대로 사용한 랜덤포레스트 모델의 성능이 가장 좋았다.

논문의 구성은 다음과 같다. 2장에서는 수산물 안전관련 영향 요인 연구와 머신러닝기반 분류

모델에 대해 기술하였다. 3장에서는 수입 수산물 분류모델을 위한 전처리 방법, 모델 구축 방법, 모델의 성능평가 결과를 기술하였다. 4장에서는 연구의 결과와 시사점을 정리하였다.

II. 관련 연구

2.1 수산물 안전관련 영향 요인 연구

수산물은 해양 또는 강에서 얻을 수 있는 식품으로 오염물질의 최종목적지가 되는 해양생태계의 오염에 직접적으로 영향을 받는다. 일반적으로 미량의 오염물질, 약물 잔류물, 중금속은 지방 및 단백질과 결합하여 해양 생물에 축적될 수 있고, 이러한 수산물의 고유한 특성은 통관검사 결과 적합·부적합에 영향을 줄 수 있다.

참고문헌[3]은 우리나라 수입식품 신고데이터를 정제하고 8가지 분류 알고리즘을 적용한 부적합 예측 모형을 개발하고 성능을 비교한 결과 Gaussian Naive Bayes 모형이 가장 좋은 성능을 보인다는 결론을 제시하였다.

참고문헌[4]에서 수산물은 단백질, EPA 및 DHA와 같은 장쇄 오메가-3 다가불포화 지방산, 다수의 비타민 및 미네랄을 비롯한 다수의 영양소를 제공한다. 그러나 높은 수분 활성, 중성 pH, 낮은 결합 조직 함량 및 바람직하지 않은 냄새와 풍미의 급속한 발달을 유발하는 자가분해 효소가 존재하기 때문에 매우 부패하기 쉬운 식품이다. 이런 이유로 수산물의 품질을 유지하고 식품 안전 위해를 최대한 줄이기 위해 유통관리가 중요하다고 서술하였다.

참고문헌[5,6]에서는 해양 생태계의 화학오염 및 다양한 오염원의 영향을 받기 때문에 이를 고려한 수산물 안전관리의 필요성을 언급하였다.

수입신고 데이터의 수입국가, 유통방식, 품종, 검사유형을 활용하여 통관검사 결과 분류 모델의 적합한 변수 전처리 방법을 연구해야 한다.

2.2 머신러닝기반 분류모델

본 연구는 과거의 4년간의 수입식품 수입신고 및 통관검사 결과 데이터로 수입신고의 통관검사 결과를 분류하는 머신러닝 모델을 개발하기 위해 지도학습 기반 모델을 구축하고자 한다. 지도학습 중 GBM, 랜덤포레스트(Random Forest, RF), XGBoost 알고리즘을 사용한다. 이와 같은 알고리즘은 앙상블을 기반으로 배깅(bagging)과 부스팅(boosting)을 사용한다[7].

앙상블(Ensemble) 방법은 지도학습으로 여러 개별 분류기(Classifiers)에 예측에 대한 가중치를 두어 결합해 모든 분류기의 능력을 증가하는 분류기를 얻는 방법이다. 강한 분류기 하나 보다 약하지만 약간 더 정확한 분류기를 생성하여 올바른 패턴을 적용할 수 있다[7].

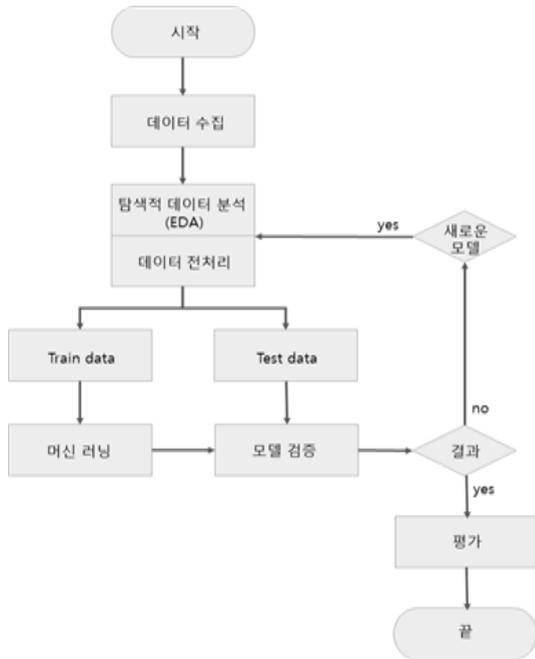
앙상블의 장점은 과적합을 회피하는 것과 클래스 불균형 문제를 완화하는 방식으로 적용될 수 있다는 것이다[8]. 배깅은 여러 개 모델을 병렬 형태로 훈련하는 기법이며, 모델은 각각 독립적으로 훈련하며 최종 결과는 모든 모델 결과를 종합하여 얻는 방식이다[9]. RF 알고리즘이 대표적인 배깅 알고리즘이다[7].

부스팅은 모델을 훈련한 후 이전 모델에서 잘못 분류한 데이터에 높은 가중치를 주어 다음 모델 훈련에 사용한다. 이러한 방식으로 오류를 줄여 나가는 부스팅의 대표적인 알고리즘은 Adaboost, GBM, XGBoost이다[10].

III. 연구 방법

본 장에서는 연구 절차, 전처리 방법 그리고 분류모델과 성능평가 결과를 설명한다.

본 연구의 절차는 그림 1과 같이 2018년에서 2021년까지 수입신고데이터를 탐색적 분석 후에 전처리를 수행한다. 데이터를 훈련 데이터(train data)와 테스트 데이터(test data)로 나누고 통관



〈그림 1〉 연구 절차

검사 결과 분류 모델 구축하고 모델검증 후 결과에 따라 모델을 재구축하고 검증결과에 따라 통관검사 분류모델을 결정한다.

3.1 활용 데이터 및 전처리 방법

논문에서 활용한 데이터는 2018년부터 2021년까지 4년간의 식품의약품안전처 수입신고 데이터이며, 데이터의 구조는 표 1과 같다. 수입식품 정보마루[2]의 수입식품 수입신고 데이터를 활용하여 연구에 필요한 데이터 세트를 재생성하였다. 수입신고 데이터의 접수일자, 수출국, 품명, 용도, 검사종류, 수입건수 변수를 통관검사결과 부적합 분류모델 구축에 적합하게 가공하고 통관검사 분류모델을 구축하고 검증한다.

분석을 위하여 기존 변수를 그대로 사용하거나 일부 변수는 파생변수를 생성하였다. 예를 들면, 품명 변수는 연어(냉장, 필렛), 흰다리새우(냉동, 살)과 같은 값을 저장한다. 품명 변수가 수산물 이름(품명)과 유통방식 및 형태를 포함하고 있다. 이러한 이유로 품명 변수는 유통방식과 품명으로 나누어 파생변수를 생성하였다. 또한, 품종을 국립수산물과학원에서 제공하는 생물종정보를 기초로 계통분류의 강과 목을 이용한 상위 분류를 표기하였으며, 표준 수산물성

〈표 1〉 변수 설명

기존변수	파생변수	변수 설명
접수일자	year(접수년도)	신고접수일자의 연도
	month(접수월)	신고접수일자의 월
	day(접수일)	신고접수일자의 날짜
	week(접수주차)	신고접수주차(1~52주차)
	day_of_week(접수요일)	신고접수요일(월~일)
품명	distribution method(유통방식)	활, 냉장, 냉동, 염장, 건조
	product(품명)	품명의 상위 분류(새우, 방어, 맨명어, 새치, 게, 가다랑어, 연어, 고등어, 오징어, 등 104개의 수산 생물명)
-	main_class(품종)	품명의 최상위 분류(어류, 패류, 갑각류, 두족류, 해조류, 기타)
수입국	-	수산물 수입국가
용도	-	수입수산물의 용도(판매용, 합작, 자사제품제조용, 외화획득용 원료, 외화획득용 제품, 연구조사용, 반송품(국내기준규격), 반송품(외화획득용원료), 반송품(기타))
검사종류	-	통관검사 종류(서류, 현장, 정밀, 무작위표본)
처리결과(종속변수)	-	통관검사 처리 결과(적합, 자진취하, 부적합, 반려)

분표 2018을 참조하여 어류, 패류, 두족류, 갑각류, 해조류, 기타로 분류하였다.

접수일자를 접수년도, 접수월, 접수일, 접수주차, 접수요일의 5가지 변수로 세분화한 파생변수로 분리한 이유는 수산물 등 식품은 계절(season), 설·추석의 명절이나 연말 등 특정 기간 등 시간적 속성에 따라 수요가 급증하여 수입물량이나 수입건수가 증가함에 따라 통관집중 기획검사비율이 상승하며 부적합 건수가 증가하기 때문이다.

수입신고한 수산물 품명데이터 설명변수를 이용하여 검사종류를 분류하는 본 연구에서는 종속변수가 처리결과이다. 지난 4년간의 수입 수산물 신고데이터 389,389건 중 처리결과 적합이 388,593건으로 99%이며, 자진 취하는 386건, 부적합은 241건, 반려는 169건으로 각각 1% 미만이다. 또한, 통관검사 4종류인 서류검사, 현장검사, 정밀검사, 무작위검사의 2021년 기준 부적합률은 0.0002%, 0.06% 0.29%, 1.11%로 평균 부적합률 0.06%에 비하여 정밀검사와 무작위검사의 부적합률이 높은 편이다.

수산물의 특성상 검사에 시간을 많이 소비할 경우 품질저하가 우려되기 때문에 모든 신고건을 검사할 수 없고, 부적합 확률이 높은 식품을 분류하여 검사를 진행하는 것이 필요하다.

통관검사 분류 모델은 블랙박스 모델보다는 설명가능한 인공지능 모델로 구축해야 수입신고 관리 정책에 활용할 수 있고 수입식품 신고 접수한 수입업자에게 검사가 필요한 이유를 설명할 수 있기 때문에 각 변수를 원-핫 인코딩하였다.

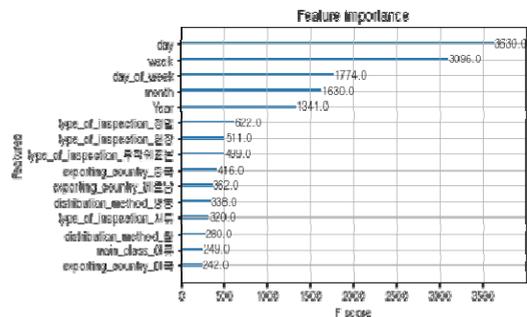
3.2 분류 모델 구축

본 연구에서 활용하는 데이터의 종속변수인 통관검사결과가 적합 99% 이상 부적합 1% 미만으로 매우 불균형한 데이터를 활용하여 분류모

델을 구축해야하기 때문에 데이터를 샘플링 기법을 적용하여 균형데이터로 샘플링 후 분류모델을 구축해야 한다. 랜덤언더샘플링, 랜덤오버샘플링, SMOTE, ADASYN 4가지 샘플링을 적용하고, 분류 알고리즘도 GBM, 랜덤포레스트(RF), XGBoost(XGB)의 3가지로 구축하여 성능을 비교하였다.

또한, 통관검사 분류 기본 모델과 성능을 높이기 위하여 하이퍼파라미터 튜닝을 적용한 튜닝 모델과 튜닝을 적용하지 않은 기본 모델로 총 24가지 모델의 성능비교를 통해 가장 좋은 성능을 내는 모델을 선택하고자 한다.

통관검사 결과 분류를 위한 튜닝 모델의 변수 중요도는 그림 2와 같다. 수산물 통관검사 결과를 적합 부적합으로 분류할 때 접수일자와 관련된 변수가 중요한 것으로 분석되었다. 그리고 검사 종류인 정밀, 현장, 무작위표본 변수가 주요했으며, 수입국은 중국, 베트남, 미국, 유통방식은 냉동과, 활이 주요 변수로 도출되었다.



〈그림 2〉 XGBoost 모델의 변수 중요도

변수 중요도에서 접수일자(day, week, day_of_week month, year)이 높게 나온 이유는 수입식품 수요량이 많은 시기인 명절 또는 특정 기간에 집중 기획검사 기간 있고, 이 시기에 부적합 식품 적발도 많기 때문인 것으로 보인다. 그리고 검사유형이 정밀, 현장인 경우는 수입식품 검사관들의 업무경험에 비추어 검사관들이 검사유

형을 정하기 때문에 검사유형 변수가 부적합을 분류하는데 주요 변수로 작용하는 것으로 판단된다. 그리고 수입국은 중국, 베트남 미국은 우리나라에서 수산물을 가장 많이 수입하는 중국, 4위 베트남, 7위 미국의 수산물이 부적합 분류도가 높은 것으로 보인다. 유통방식을 확인해보면 냉동이나 활어 유통수산물이 부적합확률이 높으며, 종류는 어류가 다른 품종에 비하여 부적합이 높은 것으로 보인다. 수입수산물 품종은 어류가 전체의 50% 이상을 차지하기 때문에 부적합 건수도 많아 높은 영향요인으로 도출되었을 것으로 보인다.

3.3 분석 결과

본 연구에서 생성한 24개의 통관검사결과 분류모델의 성능지표를 산출하였다.

표 2의 모델 성능평가 비교표와 같이 랜덤포레스트의 언더샘플링을 제외한 나머지 샘플링에서 모델의 성능은 모든 항목에서 90%가 넘는다. 특히 본 연구에서 다루는 데이터는 매우 심한 불균형 데이터이기 때문에 정확도만을 가지고는 모델의 성능을 판단할 수 없고, 매우 낮은 확률의 부적합을 잘 분류하는 것을 목적으로 하기 때문에 정밀도가 중요하다. 세 가지 알고리즘 모두 데이터 샘플링방식은 랜덤오버샘플링이 정밀도 100%의 좋은 성능을 보였다. GBM 튜닝 모델의 랜덤오버샘플링은 정확도 99.12%, 정밀도 100%, 재현율 98.24%, AUC 99.12%이고, 랜덤포레스트 기본 모델의 랜덤오버샘플링은 정확도 99.78%, 정밀도 100%, 재현율 98.57%, AUC 99.78%이며, XGBoost 튜닝 모델의 랜덤오버샘플링은 정확도 99.72%, 정밀도 100%, 재현율 98.45%, AUC 99.72%이다.

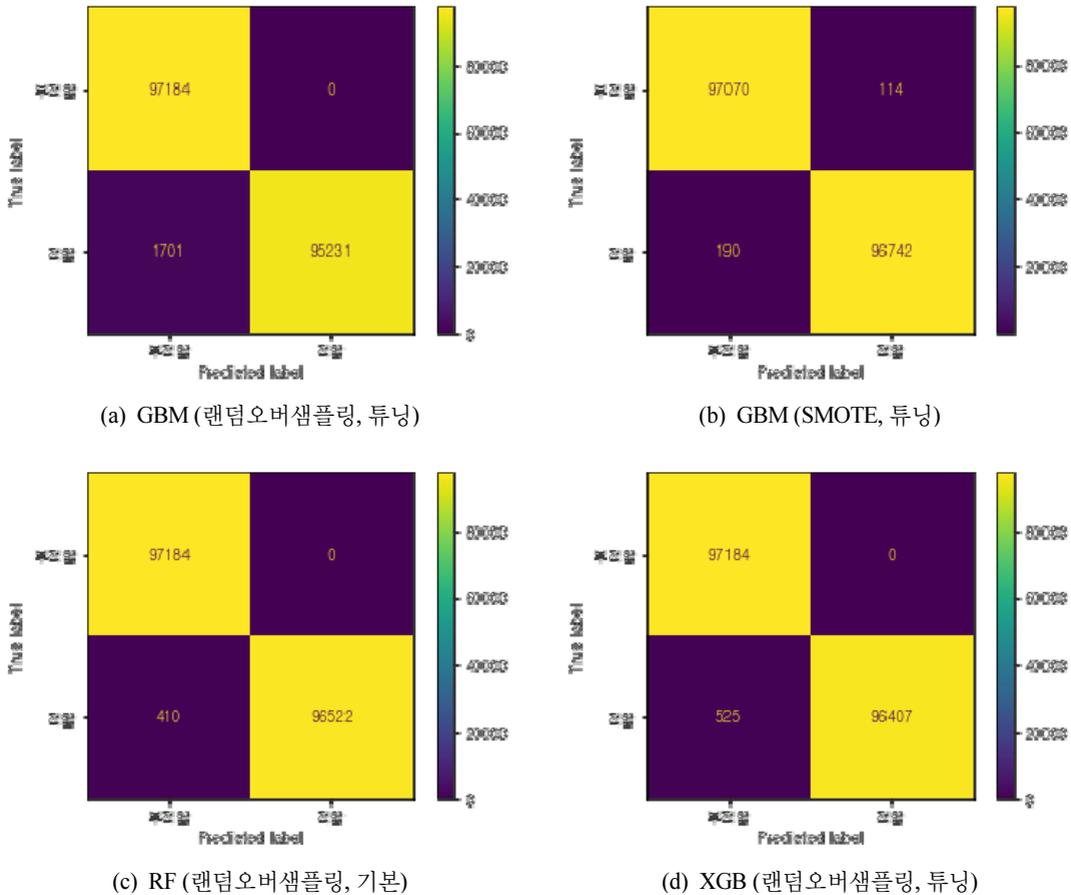
각 모델 정확도를 그림 3의 혼동행렬로 확인하였다. (a)는 GBM 튜닝 모델 중 랜덤오버샘플링을 나타낸 것이고, (b) SMOTE를 나타낸 것이

〈표 2〉 모델 성능평가 비교

알고리즘	데이터 샘플링	모델	정확도	정밀도	재현율	AUC
GBM	랜덤언더 샘플링	기본	81.8	79.6	79.0	81.5
		튜닝	81.8	77.6	82.6	81.9
	랜덤오버 샘플링	기본	82.4	87.0	76.1	82.4
		튜닝	99.1	100	98.2	99.1
	SMOTE	기본	95.1	94.2	96.1	95.1
		튜닝	99.8	99.9	99.8	99.8
	ADASYN	기본	95.3	94.1	96.7	95.3
		튜닝	99.8	99.9	99.8	99.8
RF	랜덤언더 샘플링	기본	79.2	74.2	81.2	79.4
		튜닝	79.2	75.5	78.3	79.1
	랜덤오버 샘플링	기본	99.7	100	99.6	99.8
		튜닝	84.6	89.5	78.8	84.6
	SMOTE	기본	99.9	99.9	99.9	99.9
		튜닝	95.0	97.9	92.8	95.0
	ADASYN	기본	99.9	99.9	99.9	99.9
		튜닝	95.4	96.7	94.0	95.4
XGB	랜덤언더 샘플링	기본	82.4	79.0	81.9	82.4
		튜닝	79.2	76.3	76.8	78.9
	랜덤오버 샘플링	기본	96.3	99.3	93.3	96.3
		튜닝	99.7	100	99.5	99.7
	SMOTE	기본	99.6	99.8	99.4	99.6
		튜닝	99.9	99.9	99.8	99.9
	ADASYN	기본	99.6	99.8	99.4	99.6
		튜닝	99.9	99.9	99.9	99.9

다. (c)는 랜덤포레스트 기본 모델의 랜덤오버샘플링을 나타낸 것이고, (d)는 XGBoost 튜닝 모델의 랜덤오버샘플링을 나타낸 것이다. 오류가 가장 적은 것은 (c)의 랜덤오버샘플링 해서 구축한 랜덤포레스트 기본 모델이다.

(c)RF(랜덤오버샘플링,기본)은 랜덤포레스트 모델에 오버샘플링과 하이퍼파라미터 튜닝을 적용하지 않은 기본모델이 혼동행렬에서도 FPV(0건), FNV(410건)으로 다른 모델에 비하여 우수한 결과를 얻었다.



〈그림 3〉 성능이 좋은 4가지 분류 모델의 혼동행렬 비교

IV. 결론 및 제언

본 연구에서는 수입 수산물의 통관검사 결과 부적합에 미치는 영향요인을 분석하였다. 활용 데이터는 식품의약품안전처의 수입식품정보마루에서 다운로드한 2018년부터 2021년까지의 데이터이다. 탐색적 데이터 분석을 통해 변수의 특성을 파악하고 독립 변수를 선택하여 이상치 제거, 파생변수 생성, 변수 결합 등의 전처리 과정을 거쳤다. 통관검사 결과 분류 모델은 머신러닝 기법의 GBM, 랜덤포레스트, XGBoost 알고리즘을 이용하였다. 이러한 알고리즘은 데이터 인코딩이 필요하며, 데이터의 종속변수인 통관검사

결과가 적합 99% 이상 부적합 1% 미만으로 매우 불균형하여 균형데이터로 샘플링 하였다. 데이터 샘플링은 랜덤언더샘플링, 랜덤오버샘플링, SMOTE, ADASYN의 4가지를 모두 적용하였고, 변수를 그대로 사용한 기본 모델과 하이퍼파라미터 튜닝 모델로 나누어 실험하고 비교하였다. 그 결과 랜덤오버샘플링과 원-핫 인코딩을 적용한 랜덤포레스트 기본 모델이 가장 좋은 성능을 보였다. 성능지표는 정확도 99.78%, 정밀도 100%, 재현율 98.57%, AUC 99.78%이다. 적합과 부적합을 모두 잘 예측하는 모델은 ADASYN과 원-핫 인코딩을 적용한 랜덤포레스트 기본 모델이고, 정확도 99.88%, 정밀도 99.87%, 재현율 99.89%,

AUC 99.88%이다. XGBoost는 오버샘플링과 인코딩 종류에 상관없이 모든 지표가 90%를 넘겨 가장 안정적인 모델이다.

구축된 수입수산물 통관검사결과 분류 예측 모델은 4년간 수입신고 데이터를 기반으로 한 것이므로, 식품의 수입 트렌드가 해마다 바뀌고 새로운 품목이나 새로운 수입국이 등장하기 때문에 이러한 특성을 반영한 통관검사결과 분류 모델 개발이 필요하다. 또한, 본 연구에서는 활용한 GBM, 랜덤포레스트, XGBoost 알고리즘에 추가하여 ANN, LiteGBM과 같은 다른 분류 모델의 비교 분석도 필요하다.

수산물 안전관리와 관련된 선행 연구에서 동물용 의약품, 사료, 미세 플라스틱, 잔류 농약, 항생제 사용 등의 해양 오염 원인이 수산물 안전에 영향을 주고 있음이 알려져 있다. 수입 수산물도 국제 협약을 통해 수산물 안전관리에 필요한 데이터를 실시간 공유하고 국내 수산물 안전관리와 같이 원산지 증명과 같은 수산물 이력제 제도를 보완하고, 수입 수산물의 양식장 또는 어획한 위치 정보를 등록받아 관리할 필요가 있다. 그리고 수입 수산물 환경오염 노출 여부를 측정할 수 있는 수산물 양식장 및 해양의 수온, 화학적 산소요구량, 용존물질, 용존무기질소량, 수중 미생물, 병원성 미생물 존재 유무 등 데이터를 유기적으로 연계하여 활용할 필요가 있다.

사 사

본 연구는 2023년도 식품의약품안전처의 연구 개발비(21163MFDS517-1)로 수행되었으며, 이에 감사드립니다. 그리고, 2023학년도 충북대학교 석사학위논문(저자 엄지성)의 내용을 바탕으로 재구성하였습니다.

참 고 문 헌

- [1] e-나라지표 수산물 소비량(연간 1인당)과 자급률 https://www.index.go.kr/unity/potal/main/EachDtlPageDetail.do?idx_cd=1317
- [2] 식약처 수입식품정보마루, <https://impfood.mfds.go.kr/>
- [3] 조상구 and 최경현. (2018). 수입식품 빅데이터를 이용한 부적합식품 탐지 시스템에 관한 연구. 한국빅데이터학회 학회지, 3(2), 19-33.
- [4] A Hassoun., R Karoui(2017), "Quality evaluation of fish and other seafood by traditional and non-destructive instrumental methods: Advantages and limitations," *Critical Reviews in Food Science and Nutrition*, 57(9), 1976-1998.
- [5] Vandermeersch, G., Lourenço, H. M., Alvarez-Muñoz, D., Cunha, S., Diogène, J., Cano-Sancho, G., Robbens, J(2015), "Environmental contaminants of emerging concern in seafood - European database on contaminant levels," *Environmental Research*, 143, 29-45.
- [6] Kumar, V., Kumar, S(2021), "ANN-Based Integrated Risk Ranking Approach: A Case Study of Contaminants of Emerging Concern of Fish and Seafood in Europe," *International Journal of Environmental Research and Public Health*, 18(4), 1598.
- [7] Rokach, L(2010), "Ensemble-based classifiers," *Artificial intelligence review*, 33(1), 1-39.
- [8] Sagi, O., Rokach, L(2018), "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1249.
- [9] Breiman, L(1996), "Bagging predictors," *Machine learning*, 24(2), 123-140.
- [10] Gómez-Ríos, A., Luengo, J., Herrera, F(2017), "A study on the noise label influence in boosting algorithms: AdaBoost, GBM and XGBoost," In

International Conference on Hybrid Artificial Intelligence Systems, 268-280.

저자 소개



엄지성 (Ji Seong Eom)

- 2021.03-2023.02 충북대학교 대학원
- 빅데이터학과 공학석사
- 2023.03 환경부
- 관심분야: 빅데이터, 머신러닝



이경희 (Lee Kyung Hee)

- 2022.01-충북대학교 경영정보학과 초빙교수
- 2020.03-2021.12 ㈜힐링소프트 데이터사업부 이사
- 2016.04-2020.02 충북대학교 빅데이터학과 초빙교수
- 관심분야: 빅데이터, 데이터분석, 알고리즘



조완섭 (Wan-Sup Cho)

- 1987년 : 한국과학기술원 컴퓨터과학과 (박사)
- 1996년~현재 : 충북대학교 경영정보학과(교수)
- 관심분야 : 빅데이터, 비즈니스 인텔리전스, 머신러닝