한국빅데이터학회지
제8권 제1호, 2023, pp. 61-71 https://doi.org/10.36498/kbigdt.2023.8.1.61

식품 수입 절차에서의 효율적 의사결정을 위한 데이터 전처리 기술에 관한 연구

Research on Data Preprocessing Techniques for Efficient Decision-Making in Food Import Procedures

박재형[1] · 송용욱[2] · 강주영[3†]

아주대학교 비즈니스 애널리틱스학과[1], 연세대학교 미래캠퍼스 경영학부[2], 아주대학교 경영대학[3]

요 약

데이터 기반 의사결정 방법론, 고도화된 빅데이터 처리 기법의 발달로 데이터를 처리하는 방법에 대한 정보의 수요가 늘어나고 있다. 데이터를 활용하는 거의 모든 작업과 연구에서 데이터 전처리 과정이 포함되나, 이러한 과정은 주장하고자 하는 내용이나 결과물을 도출하기 위한 수단으로써 언급될 뿐 실질적인 과정에 대해서 자세하게 설명하고 있는 연구는 부족하였다. 실질적인 분석 기법을 활용하기 이전의 단계로 간단하게 언급되는 경우가 많아 데이터 처리에 대한 인사이트를 획득하기 어려운 경우가 많았다. 따라서 이 연구에서는, raw data에서부터 데이터를 처리하는 과정, 즉 데이터 처리 파이프라인에 대해서 자세하게 작성하고자 하였다. 특히 수입식품 수입 절차에 대한 설명을 구체화함으로써 해당 상황에서 데이터의 필드들이 어떻게 해석될 수 있고 어떠한 필드들을 왜 활용하게 되었는지에 대한 상황과 관련 도메인 지식을 공유하면서 흐름을 기술하고자 하였다.

■ 중심어 : 수입식품, 데이터 전처리, 데이터 기반 의사결정

Abstract

With the development of data-driven decision-making and sophisticated big data processing technique, there is a growing demand for information on how to process data. However, recent studies with data preprocessing mentioned only as a means to achieve a result. Therefore, in this study, we aimed to write in detail about the data processing pipeline, include preprocessing data. In particular, we shares the context and domain knowledge to aid fluent understand of the research.

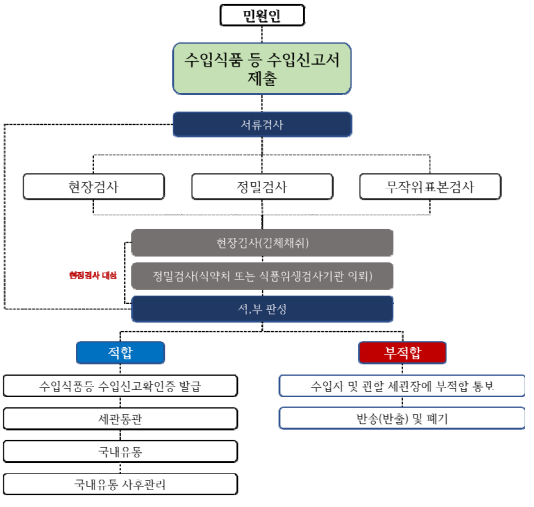
■ Keyword : Food import, Data preprocessing, Data-driven decision making

2023년 05월 30일 접수; 2023년 06월 16일 수정본 접수; 2023년 06월 16일 게재 확정.

† 교신저자 (jykang@ajou.ac.kr)

Ⅰ. 서 론

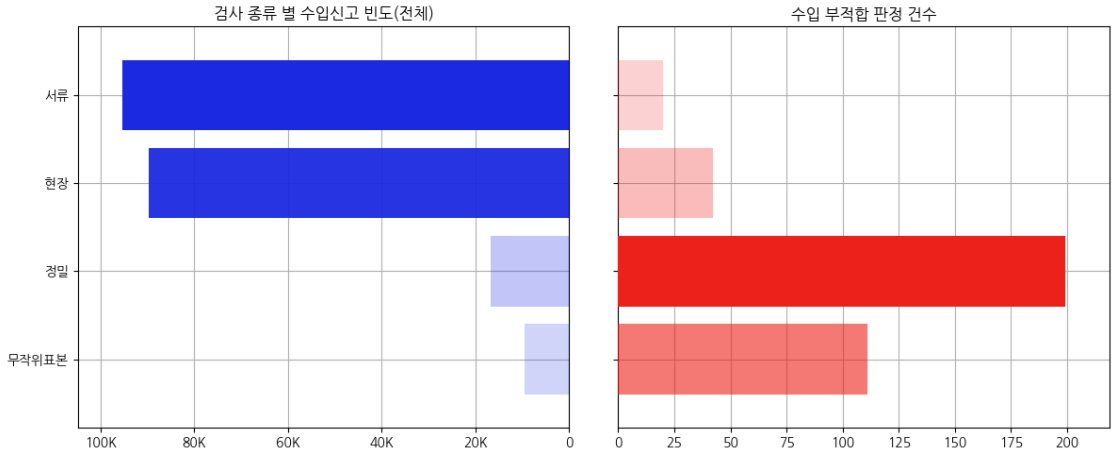
낮은 식품 자급률과 식품 기호의 다양화로 대한민국의 수입 식품의 비중은 더욱 늘어나는 추세이다[1-3]. 식품의 수입 과정에서 가장 중요한 부분은 수입 부적합 식품을 반입하는 오류를 피하는 것이다[4]. 따라서, 수입 식품 신고 프로세스를 구성할 때의 최우선적으로 고려되어야 하는 것은 식품의 수입 절차를 보수적으로 설정하더라도 부적합 식품을 국내로 들여오지 않는 것이다. 이러한 프로세스를 구축하기 위해서는 먼저 현행 국내 수입식품 통관 절차를 점검해볼 필요가 있다. <그림 1>에서 보는 것과 같이, 모든 수입 제품은 서류검사, 현장검사, 정밀검사, 무작위 표본검사 중 한 가지 이상의 검사를 받아야 한다. 「수입식품안전관리특별법」에서 요구하는 사항을 충실히 신고하였는지 서류검사, 최초로 수입되는 제품이나 부적합 이력이 있는 제품, 국내외 위해정보 이슈가 있는 제품 등은 정밀검사 대상이며, 현품의 성상, 맛, 냄새, 색깔, 표시, 포장 상태, 정밀검사 이력 확인이 필요하다고 판단되는 경우 현장검사를, 정밀검사 대상을 제외한



〈그림 1〉 현행 수입식품 통관 절차 (출처: '23년도 수입식품 등 검사연보'의 그림을 재구성)

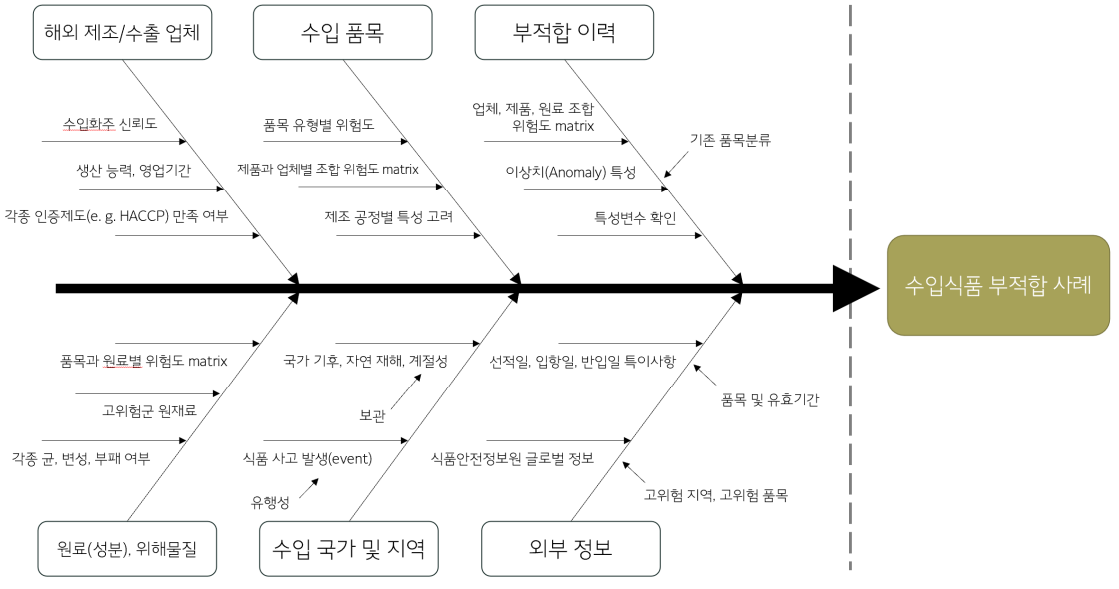
식품 등은 식품의약품안전처장의 표본추출계획에 의해 무작위 표본검사 대상으로 선정되어 물리적, 화학적 또는 미생물학적 방법에 따라 그 적부를 판단하는 검사를 진행한다. 이 표본추출계획은 정책회의와 위해도 예측 모델을 활용하여 위해 우려가 있는 식품 등을 선정하고 있다. 현행 수입식품 검사 제도에서 발생하는 부적합 데이터의 분포에서는 서류와 현장검사에서 대부분의 검사가 이루어지고 있으나, 실제 부적합 판정 건수의 대부분은 검사 빈도가 앞선 두 검사 종류에 비해 빈도가 훨씬 낮은 정밀검사와 무작위 표본검사에서 발견되고 있음을 확인할 수 있다<그림 2>. 이는 서류 검토 및 정밀검사와 무작위 표본검사 대상 품목 선별단계에서 위해도가 높은 식품을 적절하게 구분해내고 있음을 알 수 있다. 다만, 이러한 상황에도 여전히 분류 모델의 고도화는 필요하다. 첫 번째 이유로는 수입 식품의 신고 건수는 증가 추세이지만 활용할 수 있는 자원은 한정적이므로 보다 정밀한 분류가 요구되고 있고, 둘째는 여전히 검사 수에 비해 낮은 검출 빈도다. 수입 부적합 적발 건수가 가장 많은 정밀검사에서도 전체 약 2만 건의 검사 가운데 약 200건의 부적합 데이터를 적발하였다. 다른 검사 유형에 비하여 부적합 식품 검출 비율이 높기는 하지만, 여전히 개선할 여지가 있는 수치로 확인된다. 또한, 서류와 현장검사에서 발생되는 부적합 데이터 또한 반영할 수 있다.

하지만, 전체 수입식품 신고 접수 데이터에서 부적합 데이터가 차지하는 비율은 1% 미만으로, 정답(labeled) 데이터가 현저히 작은 불균형 데이터의 분포를 보인다. 이러한 불균형 데이터에서 데이터 기반 의사결정을 적용하기 위해서는, 수입신고 시 제품에 대한 데이터를 변수화하여 특성공학 적용, 이상치 발견 등을 통해 정밀검사 대상을 세심하게 분류하는 작업이 필요하다. 따라서 적절한 수입식품 위해도 분류 시스템을 구축하기 위해서는 수입식품 신고접수 데이터에



〈그림 2〉 검사 종류별 전체 빈도와 부적합 빈도의 분포

대한 명확한 이해를 기반으로 한 데이터 전처리 과정을 포함하고 있어야 한다. 앞서 확인한 바와 같이, 수입식품 신고접수 데이터는 데이터의 특성상 전체에서 부적합 판정 레이블 데이터가 차지하는 비율이 낮은 불균형 데이터 분포를 나타낸다[1]. 불균형 데이터 처리에 관한 특성 공학 방법론들은 SMOTE (Synthetic Minority Over-sampling Technique)[5,6]와 같은 샘플링 기법들과, 생성형 적대 신경망(Generative Adversarial Network: GAN) 및 오토인코더와 같은 딥러닝 기법들을 활용하여 이상치를 탐지하는 등의 연구들이 이루어지고 있으나[7,8] 이를 적용하기 위한 전처리 과정과 데이터 처리 파이프라인 구축에 대한 과정을 상세하게 수록한 연구는 많지 않다. 이러한 내용을 연구에 구체적으로 수록하기에는 여러 가지 사항이 존재하기도 하고, 실질적



〈그림 3〉 수입 식품 부적합 사례의 Fishbone Diagram

으로 연구에서 밝히고자 하는 내용이나 주장과는 직접적인 관련이 없는 과정과 내용이기 때문에 전략적인 생략이 이루어지는 것으로 해석하였다. 하지만, 데이터에 의존한 연구 및 과업이 늘어나고 있는 지금은 이러한 과정 또한 논의할 가치가 있고, 학술적인 기여와 인사이트를 제시할 수 있다고 판단하였다.

따라서 이 연구에서는 머신러닝 기반의 분류 알고리즘들을 적용하기 위한 데이터 전처리 과정 및 파생변수의 생성과 같은 특성 공학적 요소를 작성하였다. 이러한 작업을 위해 수입식품 신고접수 데이터의 특성과 관련 연구들을 참고하여 수입식품 부적합 사례의 피시본 다이어그램(fishbone diagram)을 작성하여 추후 파생변수나 변수 선택 등에 활용하고자 하였다<그림 3>. 이 다이어그램에서 확인할 수 있는 내용은, 수입식품 부적합 사례에서 발견되는 특성이 상당히 고차원적이고 다양하다는 것이다. 실제 수입신고 접수 데이터를 구성하는 컬럼의 수는 총 64개로, 다양한 특성들의 조합으로 이루어져 있다. 이 연구에서는 이러한 특성 가운데 불필요한 부분을 제거하고(data reduction and data integration), 필요한 부분을 제작(feature engineering)하는 형태로 분석에 필요한 변수 테이블을 구성하였고, 이 변수 테이블을 제작하는 과정과 실제 만들어진 변수 테이블을 연구의 방법론으로 수록하였다.

본 연구에서는 분석 방법론 정리 위주의 기존의 전처리 연구 동향에서, 실용적인 부분을 보완하여 상세하게 설명할 수 있는 과정을 수록하고자 하였다. 이를 위해 부적합 사례에 대한 분석과 실제 데이터셋을 활용한 전처리 과정을 작성하여 데이터 기반 의사결정에서 가장 기본적으로 수행되어야 할 데이터 전처리 기법 등을 작성하였다. 데이터 분석을 통한 의사결정은 기존에 데이터 분석이 활용되지 않았던 분야에서도 점차 다양하게 활용되고 있는 추세이므로, 이 연구를 통해 다양한 학술 분야에서 데이터 전처리 과정을 참고할 수 있는 도구로써 활용될 수 있기를 기대한다.

Ⅱ. 문헌 연구

2.1 데이터 전처리 및 전처리 목적

데이터 전처리는 데이터 분석 도구, 즉 모델들을 활용하기 위해 적절한 방법으로 데이터를 가공하는 것을 의미한다[9]. 따라서 데이터 전처리는 특정한 공식에 의해 수행되는 것이 아니라 데이터 분석이 필요한 과제와 상황에 따라서 데이터 분석을 하는 주체의 주관적인 해석을 기반으로 수행된다. 따라서 데이터 분석 파이프라인에 대한 전반적인 이해는 데이터 관련 과업을 수행하는 데 있어서 중요하다. 데이터 전처리 기법으로는 데이터 정제, 데이터 통합, 데이터 변형, 데이터 축소 및 변환 등이 있다[10]. 아래 표는 각 기법들에 대한 구체적인 방법론과 내용을 다루는 주요한 연구들을 수록하였다<표 1>. 데이터 전처리 기법에 관한 연구 가운데는 효율적인 방법들을 찾는 연구들이 지속적으로 진행되고 있어, 목적에 맞는 상황에서 효과적인 방법론을 찾아 활용하는 것도 효과적인 데이터 분석을 진행하기 위해 필요한 역량이다.

〈표 1〉 데이터 전처리 기법 분류와 관련 방법론

분류	기법	관련 method	연구
데이터 정제	결측치 처리	Missing value imputation	[11]
	이상치 처리	IQR(Interquartile) method	[12]
데이터 변형	데이터 표준화	Scaling	[13]
	데이터 평활	Smoothing	[14]
데이터 축소	변수 선택	Filtering	[15], [16]
	차원 축소	PCA, t-SNE	[17]
데이터 변환	데이터 이산화	Binning	[18]
	데이터 범주화	One-hot encoding	[19]

2.2 국내외 수입식품 처리 및 수입 절차 관련 연구

데이터 기반의 의사결정의 중요성은 수입식품 분류 프로세스에서도 예외가 아니며, 실제 식품 안전성 평가와 같은 주제로 다양한 연구들이 진행되고 있다. 국내외의 사례를 보면, 먼저 국내에서는 모든 수입신고 식품에 대해 서류검사를 실시하고 추가적인 검사가 필요한 경우 현장검사, 정밀검사 등을 실시한다. 또한, 예외 사항에 대응하기 위한 장치로 무작위 표본검사를 일정 비율로 시행하고 있다[20]. 해외의 수입식품 처리 및 수입 절차 관련 연구 사례로는 미국, 일본, EU, 대만의 사례를 찾을 수 있다[21]. 대만의 경우 식품의 분류에 따라 동식물위생검사검역국(BAPHIQ)과 대만 식품의약품안전처(TFDA)가 규정한 방식에 의거한 서류 제출 및 검사를 실시하여야 한다. 일본은 후생노동성(MHLW)에서 식품의 수입관리를 전담하며, 특별히 농산/수산/축산물과 관련한 품목은 농림수산성(MAFF)이 관리한다. 일본 또한 모든 수입신고 건에 대하여 서류검사를 실시하며, 수입업체의 자가검사 등이 요구된다. 미국의 경우 위해도 유의 항목에 대한 risk matrix를 구성하여 feature mapping 아이디어를 도입하였다. 이는 위해도 점수로 분류되는 위해 수입식품 여부 판단에 활용된다. 점수를 산출하는 기준은 Predictive Risk-based Evaluation for Dynamic Import Compliance Targeting (PREDICT) 모델을 활용한다. 이는 수출국과 사업자, 보관 및 배송 상태 등 종합적인 변수를 근거로 산출하는 모델로, 실제 PREDICT 모델로 위해도 점수를 구했을 때 위해도 점수가 높게 나온 구간일수록 수입 부적합 식품이 많이 검출되었다. EU의 경우 유럽연합 보건 및 식품 안전 총국(Directorate-General for Health and Food Safety: DG SANTE)에서 수입식품의 안전성을 검사하며, 동물성 수입 식품과 비동물성 수입 식품 및 사료에 대한 검사를 각각 관련 기관에서 검사하여 통관 여부를 결정한다.

이들 국가의 사례에서 찾아볼 수 있는 공통적인 절차는, 서류검사를 통한 1차 스크린을 거치고 2차적으로 세부검사를 실시한다는 점, 그리고 통관 이후에도 지속적인 감시 체계가 존재한다는 점이다. 세부검사 항목을 분류하는 데 있어서 데이터 기반의 의사결정이 포함되고, 통관 이후의 모니터링은 이러한 데이터 기반 의사결정의 근거로써 다시 활용된다. 이때 주요하게 작용하는 특성은 해당 식품 카테고리의 위반 사례, 수입 국가, 수입 과정 등이 있으며, 서류에 포함된 특성들을 종합적으로 판단하여 정밀검사 대상을 지정하고 있다[21,22]. 특히 미국의 경우 세부 검사의 기준으로 자체 모델의 결과값에 의해 선정한다는 점이 국내의 사례와 유사한데, 국내에서는 위해도 점수를 산출하는 것이 아니라 해당 수입식품에 대한 검사필요 여부만 산출한다는 점이 다르게 나타났다.

Ⅲ. 데이터 처리 파이프라인

본 연구는 식품 수입 과정에서 발생하는 1) 데이터 및 데이터 전처리 방법의 소개, 2) 탐색적 데이터 분석(Exploratory Data Analysis: EDA)을 통한 데이터 특성 발굴 및 시각화, 3) 데이터 정제 및 분석에 필요한 형태로의 데이터 변환 과정, 4) 파생변수의 생성 및 모듈화 과정 등을 담고 있다. 분석의 목적은, 수입식품 통계적 관점으로는 실제 참(결과: 부적합 사례) 사건을 거짓(결과: 국내 유입)으로 판정하는 형태의 오류, 즉 제2종 오류를 범하지 말아야 한다는 것을 의미한다<그림 4>.

따라서 수입식품 검사체계 내에서 데이터 기반 의사결정의 목적은, 부적합 식품 유형의 특성을 적절하게 파악 및 분류함으로써 예측 모델이 적용되는 대상 데이터셋에 실제 부적합 식품을 최대한 많이 포함시킬 수 있는 모델을 만드는 것

수입식품 위해도 모델을 통한 사전 분류 체계		예측 (Predicted by model)	
		적합 (Negative)	부적합 (Positive)
실제 (Actual)	적합 (Negative)	TN	제 1종 오류
	부적합 (Positive)	제 2종 오류 (위험)	TP

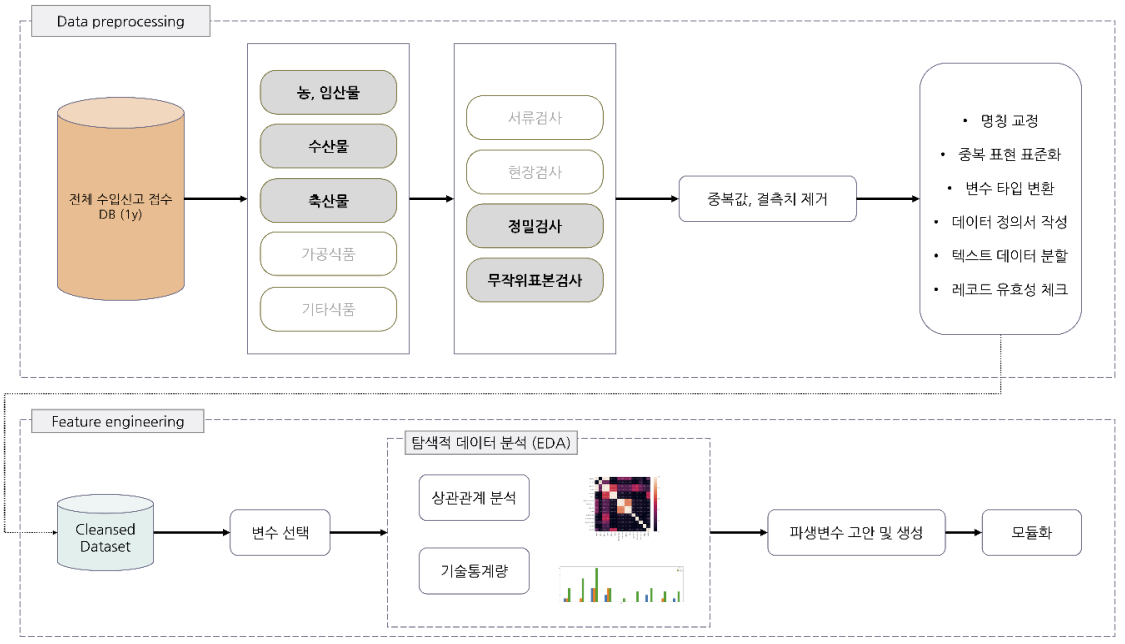
〈그림 4〉 Confusion matrix

이다. 다시 말해, 제2종 오류를 최소화하면서 정확도(accuracy)를 높여야 한다. 이 연구의 후속 연구로 다양한 머신러닝 기반의 분석 기법들을 활용하여 제2종 오류를 최소화하는 모델을 탐색할 계획이므로, 이에 초점을 맞추어 변수 테이블을 작성하고자 하였다.

3.1 연구 방법론의 요약

<그림 5>는 연구의 전체 과정을 나타낸다. 데이터 전처리 과정에서는 데이터 정제, 데이터 통합, 결측값 제거 등을 포함하여 다양한 전처리 기법들을 활용하였다. 데이터 처리가 완료된 통합 데이터에서 변수를 선택하고, 상관관계 분석 등을 활용하여 분석에 필요한 추가 변수들을 고안하고 생성하였다. 마지막으로 변수 생성 과정을 모듈화하여, 변수 체계를 정리하고 변수 제작 로직의 관리 및 활용에 용이한 형태로 구성하였다.

연간 전체 수입신고 건수는 2021년을 기준으로 약 80만 건으로 산출되었다. 이 연구에서의 변수는 2017년부터 2021년 하반기까지 총 4개년 6개월의 데이터가 활용되었다. 품목의 구분은 수산물, 식품첨가물, 축산물, 농・임산물, 기구 또는 용기 및 포장, 가공식품, 건강기능식품의 총 7가지 분류 가운데 농산, 축산, 수산물에 해당하는 연간 기준 약 25만 건의 데이터를 다루고 있다. 부적합 사례가 비교적 많이 포함되어 있으며 실제 테스트 데이터셋으로 활용될 수 있는 정밀검사와 무작위 표본검사의 수입신고 데이터로 범위를 축소하였고, 중복값 및 결측치 제거를 포함한 데이터 정제 과정을 거쳐서 EDA를 실시하였



〈그림 5〉 전체 연구과정 요약도

다. 요약하면 <그림 5>에서 보는 것과 같이, 전체 수입신고 데이터셋에서 '농・임산물', '수산물', '축산물'의 식품 카테고리에서 '정밀검사', '무작위 표본검사'에 해당하는 데이터셋을 추출하였고, 추출된 데이터셋을 바탕으로 전처리 과정 및 EDA를 거쳐 파생변수를 생성한다. 전체 절차에서 수입신고 데이터를 활용하여 제작된 파생변수와 기존의 식품안전정보원 연구 및 모델에서 발견한 변수들의 성능을 확인하여 최종적인 데이터셋을 구축하였다.

3.2 데이터 전처리

3.2.1 데이터셋 소개

개체-관계 다이어그램(Entity-Relationship Diagram: ERD)의 구성을 보면 중심적으로 활용되는 개체인 수입식품 신고정보 DB의 '수입식품 수입신고 접수' 테이블을 바탕으로 데이터 처리를 진행하였다. 파생변수의 제작 과정에서는 전체 ERD에서 파생되는 다른 참조 테이블을 활용하였으나, <표 2>에서는 일부 생략하였다.

〈표 2〉 전처리 대상 데이터 테이블 목록

구분	내 용	
수입식품 신고정보	수입식품수입신고 접수	수입신고서 정보 전반
	수입식품수입신고 정보	처리일자, 부적합 내역 등 참조
	수입식품수입신고 선별	검사 종류 코드 등 참조
	수입식품수입신고 접수 업체	수입 관련 업체 정보

데이터셋의 구성은, 2021년을 기준으로 표의 가장 상단의 수입식품 수입신고 접수 데이터가 (825975, 64)의 차원을 갖는 테이블이며 기타 참조 테이블들을 활용하여 데이터 전처리 및 EDA, 파생변수 생성 작업을 진행하였다.

3.2.2 데이터 축소

2021년의 전체 수입신고 데이터셋의 식품 분류별 분포는 <표 3>과 같다. 전체 데이터셋 가운데 타깃 분류항목인 '농. 임산물', '수산물', '축산물' 분류항목의 데이터를 추출하고, 테이블의 구분자 컬럼인 접수번호(ID)를 기준으로 중복값들을 제거하였다. 또한, 검사 종류를 정밀검사와 무작위 표본으로 축소하고, 유효한 컬럼 내의 값이 없는 경우 등을 제외하여 최종적으로 26,114건의 데이터를 추출하였다.

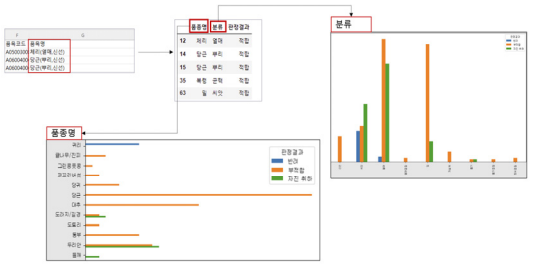
〈표 3〉 수입식품 수입신고 접수(2021) 테이블 분류별 레코드 분포

분류	처리 전(건 수)	처리 후(건 수)
가공식품	271,593	-
기구 또는 용기 및 포장	202,501	-
축산물	107,038	8,321
수산물	88,639	11,704
농・임산물	73,656	6,089
건강기능식품	46,732	-
식품첨가물	35,816	-
계	825,975	26,114

3.2.3 데이터셋 정제 과정

주어진 데이터셋은 64개의 컬럼으로 구성된 테이블이면서 레코드 수가 많으므로 데이터를 적절하게 정제하여 분석과정에 유효한 값들만 남기는 과정이 필요하다. 그 과정들은 데이터 타입의 변환과 입력값의 표준화, 결측값 처리, 텍스트 데이터 처리 등과 같다. <그림 6>은 텍스트 데이터를 처리하고 파생변수를 제작하는 과정의 일부를 보여주고 있다. 테이블에서 '품목명'이라는 변수의 처리 과정에서 하나의 컬럼에 속한 내용들을 분할하여 세부 변수로 만들었을 때 얻어지는 효과를 나타낸다. '품목명'으로 정의되어 있을 때는 품목명의 세부 내용들이 하나의 컬럼 값으로 포함되어 있었으나, 품목에 포함된 텍스

트 내용을 세부 항목으로 구분할 수 있었는데, 예컨대 당근(뿌리, 신선)으로 저장된 품목명 컬럼의 값을 당근, 뿌리로 각각 나누어 '품종명', '분류'로 새로 저장할 경우 각각의 분류가 더 부적합 품목을 설명하는 데 도움을 줄 수 있다. <그림 6>에서 나타나듯이, 텍스트 데이터를 분할한 이후 시각화 분포를 확인하면 특정 품종명, 특정 분류에서 더 눈에 띄는 부적합 분포를 나타내는 것을 확인할 수 있다.



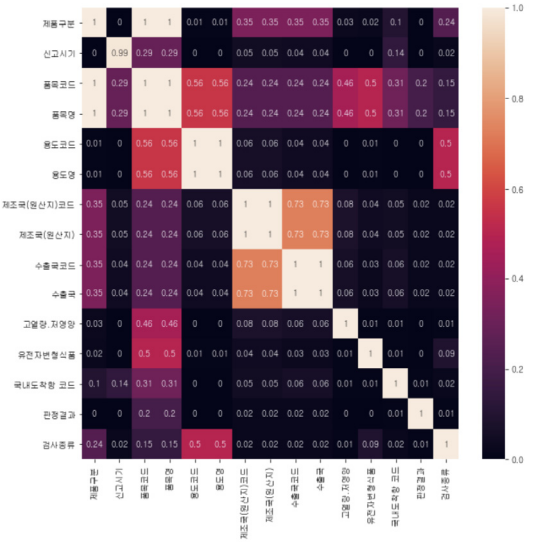
〈그림 6〉 텍스트 데이터 스플릿 및 시각화

이 밖에도 수입신고서에서 중요하게 평가되는 선적일, 입항일, 등록일과 같은 시간 데이터의 구간을 확인하여 '유통 시간' 등의 변수를 제작할 수 있었으며, 부적합 사유서 내의 비고 및 행정조치 사유 등의 비정형 데이터를 활용하여 새로운 변수를 제작하거나 해외 제조업소의 기후 특성이나 기존 위반 사례 등을 활용하여 변수를 제작할 수 있다.

3.3 탐색적 데이터 분석(EDA) 및 파생변수 생성

3.3.1 상관관계 분석

상관관계 분석은 범주형 변수를 대상으로 실시하였다. 데이터의 특성상 정량적으로 표현되지 않는 텍스트, 서술형 데이터가 많았고, 이를 카테고리화하여 판정 결과에 미치는 상관관계 그래프를 작성하였고 이를 변수 선택 과정에서 활용하였다. 본 연구는 범주형 변수에 대해 Cramer's V 상관계수를 활용하였다. Cramer's V 상관계수는 범주형 변수 개수가 2개 이상일 경우에서의 변수간의 관계를 파악하기 위한 상관계수로, 0과 1 사이의 값을 갖으며 값이 0.15 이상일 경우 유의미한 상관관계를 갖고 있다고 해석할 수 있다[23]. 판정결과에 대한 상관관계를 도출한 결과<그림 7>, 품목코드와 품목명이 각각 0.2로 유의미한 상관계수가 도출되었고 기타 변수들은 상관계수가 0.1 이하로 나타났다. 이는 품목에 따라 부적합 데이터가 어느 정도 구분이 되고 있음을 시사하고, 품목을 세분화하고 품목과 관련된 파생변수를 제작하여 활용할 경우 설명력을 갖는 모델을 설계할 수 있음을 시사하고 있다.



〈그림 7〉 범주형 변수 상관계수 heatmap

3.3.2 Feature engineering

식약처의 기존 분류 모델에서 활용하던 변수 세트의 일부를 일반화 선형 모델(Generalized Linear Model: GLM)을 통해 각 변수의 가중치를 확인하였다. 이 과정에서 변수 활용도가 높은 변수들을 추출하였고, 추가적으로 기존 연구에서 활용된 변수와 앞서 추출한 파생변수들을 조합하여 최종 변수 테이블을 구성하는 데 활용하였

다. 기본적으로는 기준 테이블 내의 여러 가지 변수들의 조합으로 구성되었으나, 유의 공정 또는 성분에 대한 정보를 활용한 변수는 추가적인 참조 테이블을 요청하여 변수를 제작하였다. 연속형 변수의 카테고리화 작업은 파이썬의 데이터 이산화(data discretization) 관련 패키지를 활용하여 구성하였다.

3.4 최종 파생변수 테이블 및 모듈화

최종적인 테이블을 구성하여, 변수를 생성하는 로직과 생성 과정을 효율화하기 위해 변수 생성 로직을 함수화하고, 함수들을 패키징하여 모듈화하였다<그림 8>.

〈그림 8〉 변수 생성 과정의 모듈화

<그림 8>은 변수 생성 과정에서 발생하는 함수 코드들을 묶어 클래스 변수로 만들고, 그 클래스들을 패키지로 활용하는 과정을 나타낸다. 농산물, 수산물, 축산물변수를 생성하기 위한 함수 단위에서 패키지 단위까지 모듈화시켜, 변수 생성 로직과 추후 연구 협업 과정에서도 효율적으로 활용할 수 있도록 제작하였다.

Ⅳ. 결론

본 연구는 데이터 분석 과정에서 실제 데이터셋을 활용하여 전처리 기법들을 사용하는 과정을 자세하게 수록하였다. 기존의 연구들은 데이터를 활용하여 연구 결과물을 도출하기까지의 과정이 다소 부족하게 서술되어 있는 경우가 많았으나, 이 연구에서는 그 과정을 상세히 기술하여 데이터 기반 기존 연구의 부족한 부분을 채우고자 하였다. 이 연구에서는 수입식품 안전관리 시스템 ERD에서 기준이 되는 수입식품 수입신고 접수 테이블의 전처리 방법과 테이블 내의 특성들을 분석한 결과를 수록하였다. 분석 내용을 활용하여 변수를 제작하였고, 기타 수입식품 분류체계에 활용되는 변수들을 수집하여 최종 변수 테이블을 구성하였다. 이 연구를 통해 기대할 수 있는 부분들은 다음과 같다. 첫 번째로, 연구에 기술된 데이터 전처리와 변수 생성을 포함한 데이터 처리 파이프라인은 다른 분야의 데이터 관련 연구 설계 단계에서의 가이드라인으로 활용될 수 있을 것으로 기대한다. 두 번째로는 수입식품 검사 과정에서 극복해야 하는 문제를 소개 및 주요 과업으로 설정하고, 이러한 특정 상황에서 데이터 전처리 과정에서 어떻게 접근해야 하는지에 대한 고찰을 작성하였다. 데이터를 활용한 문제 해결방법에 대해서 실질적인 접근을 연구로써 풀어낸 사례는 많지 않았기 때문에, 이 또한 학술적인 기여로 인정받을 수 있기를 기대한다. 세 번째로는 데이터 처리와 탐색적 데이터 분석을 활용하여 새로이 발굴된 변수들은 추후 식품의약품안전처의 현행 수입식품 신고서 기반의 분류 모델의 결과물 개선에 기여할 수 있다는 점에서 실용적인 측면의 기여 또한 있을 것으로 생각된다.

추가적으로 국내 수입식품 검사 과정에서 발생하는 실질적인 문제들을 공유함으로써 보다 고도화되고 안전한 수입식품 접수 과정에 대한 관심과 논의가 확장되기를 기대한다. 식품 수입 과정에서 발생하는 데이터는 특성과 레코드 개수가 많지만, 부적합 데이터 개수는 적은 불균형 분포를 띄는 데이터셋이기 때문에, 부적합 데이

터를 완벽히 설명할 수 있는 모델을 만들기에는 현실적인 어려움이 존재하고 국민의 안전과 직결된 문제이기 때문에 예외 사항을 최소화하는 방안으로 분류 모델을 설계 및 개발해야 한다. 그 과정에서, 수입 식품들 가운데 부적합 데이터에 대한 설명력이 높은 변수들을 앞으로 더 발굴하여 효과적으로 부적합 식품들을 걸러낼 수 있는 절차가 확립될 수 있기를 기대한다.

감사의 글

본 연구는 2022년도 식품의약품안전처의 연구개발비(21163MFDS516-4)로 수행되었으며 이에 감사드립니다.

참 고 문 헌

[1] 이경수, 박예린, 신윤종, 손권상, 권오병, “효율적 수입식품 검사를 위한 머신러닝 기반 부적합 건강기능식품 탐지 방법”, 지능정보연구, 제28권, 제2호, pp. 139-159, 2022.

[2] 조상구, 조승용, “기계학습을 이용한 식품위생 점검 체계의 효율성 개선 연구”, 한국빅데이터학회지, 제5권, 제2호, pp. 53-67, 2020.

[3] Nganje, W.E., *Quality Assurance for Imports and Trade: Risk-Based Surveillance, in US Programs Affecting Food and Agricultural Marketing*, Available at Springer, 2012.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique”, J. of Artificial Intelligence Research, 16: pp. 321-357, 2002.

[5] Han, H., W.-Y. Wang, and B.-H. Mao., “Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning”, Advances in Intelligent Computing: International Conference on Intelligent Computing, ICIC 2005, Proceedings, Part I 1. 2005.

[6] He, H., E.A. Garcia, “Learning from imbalanced data”, IEEE Transactions on Knowledge and Data Engineering, 21(9): pp. 1263-1284, 2009.

[7] Zhang, T., Chen, J., Li, F., Zhang, K., Lv, H., He, S., E, Xu., “Intelligent fault diagnosis of machines with small & imbalanced data: A state-of-the-art review and possible extensions”, ISA Transactions, 119: pp. 152-171, 2022.

[8] Li, C., “Preprocessing methods and pipelines of data mining: An overview”, arXiv preprint arXiv: 1906.08510, 2019.

[9] Garcia, S., Ramirez-Gallego, S., Luengo, J., Benitez, J., Herrera, F., “Big data preprocessing: methods and prospects”, Big Data Analysis, Vol. 1, No. 9, 2016.

[10] Lin, W.-C. and C.-F. Tsai, Missing value imputation: a review and analysis of the literature (2006-2017). Artificial Intelligence Review, 53: pp. 1487-1509, 2020.

[11] Rousseeuw, P.J., M. Hubert, “Robust statistics for outlier detection. Wiley interdisciplinary reviews: Data mining and knowledge discovery”, 1(1): pp. 73-79, 2011.

[12] Patro, S., K.K. Sahu, “Normalization: A preprocessing stage”, arXiv preprint arXiv: 1503. 06462, 2015.

[13] Xiao, Z., Gang, W., Yuan, J., Chen, Z., Li, J., Wang, X., Feng, X., “Impacts of data preprocessing and selection on energy consumption prediction model of HVAC systems based on deep learning”, Energy and Buildings, 258: p. 111832, 2022.

[14] Talavera, L., “Feature selection as a preprocessing

step for hierarchical clustering", in International Conference on Machine Learning (ICML), 1999.

[15] Li, J., Cheng, K., Wang, S., Morstatter, F., P. Trevino, R., Tang, J., Liu, H., "Feature selection: A data perspective", ACM Computing Surveys (CSUR), 50(6): pp. 1-45, 2017.

[16] Van der Maaten, L., G. Hinton, "Visualizing data using t-SNE", J. of Machine Learning Research, 9(11), 2008.

[17] Tsai, C.-F., Y.-C. Chen, "The optimal combination of feature selection and data discretization: An empirical study", Information Sciences, 505: pp. 282-293, 2019.

[18] Li, C., "Preprocessing methods and pipelines of data mining: An overview", arXiv preprint arXiv: 1906.08510, 2019.

[19] 박혜진, 조상구, 수입식품 현지실사 업체 선정을 위한 예측모형 개발. 학술대회 및 심포지엄, 2022.

[20] Ndraha, N., H.-I. Hsiao, and W.C.C. Wang, "Comparative study of imported food control systems of Taiwan, Japan, the United States, and the European Union", Food Control, 78: pp. 331-341, 2017.

[21] Kwak, N.-S., "Comparative analysis of the imported food control systems of the Republic of Korea, Japan, the United States, and the European Union", Food Reviews International, 30(3): pp. 225-243, 2014.

[22] Akoglu, H., "User's guide to correlation coefficients", Turkish J. of Emergency Medicine, 18(3): pp. 91-93, 2018.

저 자 소 개

박 재 형(Jae-Hyeong Park)

・2022년 2월: 아주대학교 e-business학과, 아주대학교 ICT융합전공 (공학사)
・2022년 3월～현재: 아주대학교 비즈니스애널리틱스학과 석사과정

<관심분야> : Data-driven solution, Data Analytics, Business Analytics

송 용 욱(Yong-Uk Song)

・1988년 2월: 서울대학교 국제경제학과 (경제학사)
・1990년 2월: 한국과학기술원 경영과학과 (공학석사)
・1995년 8월: 한국과학기술원 산업경영학과 (박사)
・2000년～현재: 연세대학교 미래캠퍼스 경영학부 교수

<관심분야> : e-비즈니스, 인공지능 응용, 전자 결제 및 보안

강 주 영(Ju-Young Kang)

・1995년 2월: 포항공과대학교 컴퓨터공학과 (공학사)
・1997년 2월: 서울대학교 컴퓨터공학과 (공학석사)
・2005년 2월: 한국과학기술원 경영공학 (박사)
・2005년～현재: 아주대학교 경영대학 e-business학과 교수

<관심분야> : 텍스트 마이닝, 빅데이터 분석, 지능형 전자상거래, 위성활용 비즈니스