



장려상

돼지고기 공급량 예측 모형 연구

양유경 · 부수빈 · 남주현 · 김솔희

돼지고기 공급량 예측 모형 연구

의사결정나무, 인공신경망, 서포트 벡터 회귀 활용을 중심으로

양유경* 부수빈* 남주현** 김솔희*

Key words

돼지고기 공급량(Pork supply), 기계학습(Machine learning), 예측 모형(Prediction model), 의사결정나무(Decision tree), 인공신경망(Artificial neural network), 서포트 벡터 회귀(Support vector regression)

ABSTRACT

The aim of this study is to develop a prediction model on pork supply in Korea using several machine learning methods. Factors affecting the amount of pork supply were classified into two groups: stock and production. Data collection and feature selection were performed based on this classification. Instead of deriving an optimal set of features for each model, stepwise selection was applied to a multiple linear regression model to unify the subset of features that were to be used in training for each model. As a result, 3 out of 8 features that were classified in the stock group, and 5 out of 18 features that were classified in the production group were selected. A total of 8 features were used to estimate the amount of pork supply using decision tree, artificial neural network(ANN), and support vector regression(SVR) methods. The performance of each optimal model was compared, and the results indicate that the feature selection method used in this study was not significantly effective. However, results showed that the decision tree model using the data set after feature selection showed the better performance than when using the data set before. Also, the model with the best performance was found to be the SVR model to which data before feature selection was applied.

차례

- | | |
|----------|----------|
| 1. 서론 | 3. 분석 결과 |
| 2. 분석 모형 | 4. 결론 |

* 건국대학교 동물자원과학과 소속

** 한남대학교 경영학과 소속

1. 서론

2018년 기준 축산물 생산액은 지난 54년간 꾸준히 증가해왔으며 그중 돼지가 높은 비율을 차지하고 있다. 국산 돼지고기 소비량은 2000년 67만 4,017톤에서 2018년 91만 9,931톤으로 연평균 1.7% 증가하였으며, 같은 기간 수입 돼지고기 소비량은 10만 5,891톤에서 41만 4,539톤으로 연평균 7.9% 증가하였다(통계청, 2020).

소비량과 생산액이 꾸준히 증가함에도 불구하고 돼지의 산지가격은 지난 21년간('99~'19년) 연평균 1.3% 하락하였고, 불안정한 비육돈 두당 수익성은 결과적으로 지나친 공급을 발생시켜 지난 39년간(1981~2019년) 비육돈 농가의 축산소득이 연평균 1.8% 감소하는 결과를 가져왔다(통계청, 2020). 실제로 2020년 5월까지만 해도 2019년 하반기 관측 당시보다 10만두 정도 많이 출하됐다. 한돈팜스의 월간 성적 보고서에 따르면, 11개월 전 번식돈 사료를 지표로 하여 추정한 2020년 한해 돼지 출하두수 또한 당초 전망치를 넘어 1천 800만두 이상이 될 수도 있다는 분석 결과를 통해 공급량 조절을 시행하기도 하였다.

최근에는 해외 수급상황과 코로나19 등의 외부적인 요인들이 국내 돼지가격에 큰 영향을 미치고 있는 것으로 나타났다. 코로나 사태에 영향을 받았던 미국의 육류생산 및 공급체계가 예상보다 빠르게 회복세를 보이고 있는 와중에, 중국의 수요가 이를 뒷받침하지 못하면 국제시장의 혼란은 불가피할 것으로 보인다. 결국, 남는 수요가 한국 시장으로 밀려드는 상황이 전개될 수 있다. 뿐만 아니라, 향후 국제시장이 안정되고 수입량이 증가하게 된다면, 국내 양돈산업 시장은 또다시 불안정해질 위험이 있다.

또한, 앞으로 돼지고기 소비는 가격에 따라 민감하게 반응할 것으로 예상된다고 한다. 돼지고기의 소비자 가격이 일정 수준 이상으로 오르게 되면 소비 방향은 닭고기를 비롯한 대체육으로 전환될 가능성이 크다. 실제로 국내산 삼겹살 소비자 가격이 kg당 2만 5천 원을 넘는 즉시 구매저항으로 이어지는 것이 그간의 추세다. 농촌경제연구원에서 주최한 '농업전망 2020'에서 발표한 소비자 대상 조사 내용에 따르면, 국산 돼지고기 소비를 줄일 경우 대체 육류로 많이 꼽은 것이 닭고기(47.2%)였으며, 수입 돼지고기의 대체 육류로 가장 많이 꼽은 것도 닭고기(40.3%)인 것으로 조사되었다.

한편 연도별 가정 내 돼지고기 구매량 추이를 살펴보면, 구매량이 지속적으로 감소하고 있는 것으로 나타나고 있다. 2014년 4주 평균 구매량이 2.28kg인 것에 반해, 2019년은 1.83kg으로 19.7% 감소했다(한국농촌경제연구원, 2020). 축산물가격과 생산성, 수익성 등은 단기적 수급 여건뿐만 아니라 중장기적인 양돈산업 변화 전망에서 중요한 요소이다. 그렇기 때문에 미래의 양돈산업 변화를 예측하기 위해서는 수요, 공급, 가격의 유기적 관계를 종합적으로 고려하여야 한다.

돼지고기 과잉공급과 가격 불안 상황에 직면해 있으면서도 그동안 돼지고기의 수급구조와 관련한 연구는 돼지 단독으로 진행된 경우가 적으며, 유통구조와 소비 동향을 중심으로 분석하는 연구가 대부분이기에 분명한 한계가 존재한다. 이는 합리적인 수급조절과 출하조절을 위하여 공급량에 대한 변수들을 함께 고려한 돼지 수급모형 구축 필요성을 시사한다.

따라서 본 연구는 돼지고기 총공급량의 변화를 초래하는 원인을 생산량과 재고량의 측면으로 구분하여 분석하고 공급되는 돼지고기의 총량을 예측하는 기계학습 기반의 모델을 설계하고자 한다. 또한, 돼지의 생산과정 및 사육 특성을 고려하여 돼지고기 수급에 장·단기적으로 영향을 미치는 변수 중 생산량과 관련된 변수를 새로 발견함으로써 보다 높은 예측력의 모델을 설계하고자 한다. 이러한 분석 결과는 양돈농가뿐만 아니라 유통업자, 소비자에게 직간접적으로 도움이 될 것으로 예상된다.

1.1. 선행연구 검토

전술한 바와 같이, 그동안 돼지고기의 수급구조에 관한 연구는 다양한 방법론을 사용하여 이루어져 왔지만, 주로 쇠고기 외 기타 축산물과 함께 연구가 되었거나 유통구조와 가격을 중심으로 연구가 진행되었다.

이정희(2001)는 우루과이라운드 협상 이후 육류시장의 유통단계별 인과관계를 확실히 규명하고 유통정책의 효율성을 개선하기 위하여 1995년부터 2000년까지 월별자료와 VAR 모형을 이용한 유통단계별 가격의 인과관계를 분석하였다. 그 결과 돼지고기의 산지가격과 도매가격은 연계성이 높지만, 그에 비하여 소매시장은 연계성이 낮은 것으로 분석하였다.

최정섭(2001)은 육류의 유통단계별 가격 연동성을 개선하고 소비자에게 육류 구매 기준가격을 제공하기 위하여 육류 소비자 권장가격 산정모형을 개발하였다. 육류 소비자 권장가격이 실효성과 공정성을 가지기 위해서는 가격 조사방법의 체계화, 지역별 권장가격의 차별화 그리고 자료제공자와 권장가격 공표가 간의 지속적인 조사 협조 등이 필요하다는 것을 제안하였다.

김혜영 등(2003)은 ‘육류 수요체계의 내생성과 구조변화’ 연구에서 소비자 가격과 분기별 소비량을 이용하여 LA/AIDS(준이상수요체계)와 LA/IAIDS(역준이상수요체계)로 한국의 육류 수요체계를 분석하였다.

심춘수, 정경수, 김민정(2006)은 가격과 관련된 데이터가 시계열 데이터임을 고려하여, 시계열의 비정상성을 고려하는 비대칭적 오차정수모형을 이용하여 돼지고기의 경우에는 산지와 도매 사이 단계에서 가격전달의 비대칭성이 관찰된다는 결과를 제시하였다.

김민정 등(2007)은 ‘돼지고기 일반광고에 대한 가격효과 분석’에서 돼지고기 일반광고가 돼지가격 상승으로 이어지기 때문에 양돈농가의 수익을 향상시키는 것에 기여하는 것으로 추정하였다.

윤병삼(2008)은 2000년부터 2006년까지 농협서울공판장의 일일 지육경락가격을 사용하여 쇠고기 및 돼지고기의 가격변동성을 분석하였다. 그 결과 요일효과와 계절효과가 쇠고기와 돼지고기의 가격변동성에 영향을 미치고 큰 변동 뒤에는 한동안 큰 변동이 이어지고 작은 변동 뒤에는 작은 변동이 이어지는 변동성 집적현상이 있음을 밝히고, 가격위험을 해소하기 위한 수단이 필요함을 제시하였다.

정민국과 김현중(2011)은 주로 공급 측면의 요인으로 인해 돼지고기 가격에 계절성이 나타남을 제시하였다. 돼지는 여름철 더위로 인한 스트레스로 수태율이 하락하기 때문에 10개월 후인 5~6월 출하물량을 감소시켜 산지 및 도매가격을 상승시키며, 겨울철 돼지의 수태율 상승은 다음해 10~11월 출하물량을 증가시켜 연중 산지가격의 최저 수준을 형성한다. 그랜저의 동태적 인과성 분석결과, 돼지산지가격은 돼지 도매가격과 소매가격에 많은 영향을 받으며, 돼지 소매가격은 돼지 산지가격과 쇠고기 소매가격에 많은 영향을 받는다. 공적분 관계의 구조변화에 대한 검정결과 산지와 도매 사이 단계에서 유통가격 간 공적분 관계가 구조변화를 겪은 것으로 나타났다.

지인배, 황운재, 이형우, 한봉희(2015)는 돼지의 경우 캠핑문화가 확산되고 일본의 방사능 오염수가 유출된 상황, HPAI 발생 등이 돼지고기 수요 증가에 영향을 주는 것으로 추정하였다. 육류 수요변화 요인인 캠핑, 마트, 정육점, 할인행사, 육가공품 구매 등이 육류 소비형태에 미친 영향을 조사한 결과 마트 육류 할인 행사가 육류소비에 가장 큰 영향을 미쳤으며, 육가공품 구매, 캠핑, 정육점형 식당 순으로 나타났다.

남국현과 최영찬(2016)은 돼지고기의 부위별 수요함수를 추정하고, 남편의 직업이 돼지고기의 부위별 소비량에 미치는 영향을 살펴보았다. 또한, 구입처에 따른 부위별 소비량 차이를 확인하였으며 재래시장이 마트에서 구입한 경우에 비해 소비량이 더 많은 것으로 분석하였다. 또한, 준이상수요체계 모형을 이용한 돼지고기 부위별 가격탄력성과 지출탄력성을 추정하였으며, 그 결과 삼겹살은 갈비를 제외한 모든 부위와 대체관계를 보였다.

2. 분석 모형

2장의 구성은 다음과 같다. 제1절은 본 연구에서 사용하는 3가지 머신러닝 방법의 이론을 간단하게 설명한다. 제2절은 분석에 사용된 자료에 대해 설명하며, 이어서 제4절은 분석 방법을 설명한다. 이후 이어지는 3장과 4장에서 분석 결과를 정리하여 보여주고 의미를 해석 및 요약하여 결론을 내린다.

2.1. 머신러닝 방법론

본 연구는 돼지고기 공급량 예측 모형을 제시하기 위해 의사결정나무(decision tree), 인공신경망(artificial neural network, ANN), 서포트 벡터 회귀(support vector regression, SVR) 알고리즘을 사용하였으며, 각 머신러닝 방법론을 다음과 같이 정리하였다.

2.1.1. 의사결정나무(Decision Tree)

의사결정나무는 의사결정 규칙을 나무 구조로 나타내어 전체 자료를 여러 소집단으로 분류하거나 예측을 수행하는 방법이다(최종후, 1999). 상위 노드에서 하위 노드로 이동하면서 트리구조를 형성하는데, 노드가 분기되는 모든 단계마다 하나의 변수와 값을 분류의 기준으로 선택한다. 분류변수와 기준값은 해당 기준에 의해 분기되는 노드에서 노드 내 동질성과 노드 간의 이질성이 최대화되도록 선택된다. 의사결정나무는 목표변수가 이산형인 경우의 분류나무와 목표변수가 연속형인 경우의 회귀나무로 구분되며, 회귀나무의 경우에는 분류변수와 분류 기준값의 선택 방법으로 F-통계량의 F-값, 분산의 감소량이 사용된다(김규범, 2019).

의사결정나무는 화이트박스 모델의 대표적인 예로, 분석 과정이 직관적이고 이해하기 쉽다는 장점이 있는 반면에 가지가 나뉘질 때 분류변수 한 개만 고려하기 때문에 예측력이 떨어지는 편이다. 또한, 데이터가 조금만 변화해도 나무의 구성이 달라져 모형 불완전성 문제가 제기될 수 있다(장제훈, 2021).

2.1.2. 인공신경망(Artificial Neural Network, ANN)

인공신경망 모형은 특정 자극이 발생하면 신경세포에서 신경 전달물질에 의한 신경세포 간의 자극 전달과정을 거쳐 학습과 의사결정 등이 이루어지는 인간의 신경세포 연결 구조와 유사하게 만들어졌다. 생물학적인 뇌의 감각 입력에 대한 처리와 출력 원리를 모사하여 인공 뉴런과 이러한 뉴런을 연결시켜주는 매개체로 구성된 모형이다. 뇌가 뉴런이라고 하는 상호 연결된 세포 네트워크를 사용하여 정보를 처리하는 것처럼, 인공신경망은 인공 뉴런(노드) 네트워크를 사용하여 데이터를 분석한다(류상모, 2020).

인공신경망의 구조는 입력층, 은닉층, 출력층 3개의 layer로 구성된다. 입력층은 각 입력변수에 대응되는 노드로 구성되며, 입력노드의 수는 입력변수의 수와 같다. 입력노드에 전달된 입력 신호를 출력노드에 전달되는 출력신호로 변환하는 활성화함수를 통해 출력신호의 범위가 결정된다. 출력신호는 입력층으로부터 전달되는 변수값을 작은 범위의 출력값으로 압축하기 때문에 출력신호에 영향을 미치는 입력층으로부터 전달되는 변수값의 범위가 상대적으로 좁

다. 이때 입력값에 대한 왜곡 또는 오류가 발생할 수 있다. 이런 문제를 해결하기 위해 입력값을 '0' 근처의 작은 범위 안으로 들어오도록 변환하는 표준화 혹은 정규화를 동원할 수 있다.

은닉층은 입력층으로부터 전달되는 변수값들의 활성화함수로 처리하고 이를 출력층 또는 다른 은닉층에 전달하는 기능을 수행한다. 이때, 은닉층이 2개 이상인 경우를 다층 퍼셉트론 모형이라고 한다. 출력층은 목표변수에 대응되는 노드로 구성된 층으로 예측 또는 분류된 결과를 나타낸다(이슬기, 2011).

인공신경망 모형은 블랙박스 프로세스의 예로, 입력을 출력으로 변환하는 메커니즘이 불투명하여 왜 그러한 결과가 나오는지에 대한 연관 관계를 설명하기 어렵다. 또한, 뉴런의 수가 많으면 훈련 데이터를 더욱 긴밀하게 반영하는 모델이 생성되지만 과적합화 가능성의 문제가 가장 큰 단점이다(남영우, 2006). 하지만 은닉층의 존재로 자료의 타입에 상관없이 처리 가능하며 구조적인 변화 또는 외부 충격으로 시장이 급변하는 경우에 복잡한 시장 추세를 유사하게 예측할 수 있다는 점에서 분명한 장점이 있다(김태훈, 2004).

2.1.3. 서포트 벡터 회귀(Support vector regression)

서포트 벡터 머신(support vector machine, SVM)은 훈련 데이터를 고차원의 형상 공간으로 대응시킨 후, 해당 공간에서 최적의 의사결정 영역을 찾는 학습기법이다(Hearst et al., 1998). 이때 최적의 의사결정 영역이라 함은, 데이터를 분류하는 경계인 초평면 중에서 두 클래스 간의 거리를 최대화시키는 최대 마진 초평면을 의미한다. 본 개념은 통계학자인 Vapnik에 의해 선형 분류기로 처음 등장하였다(Vapnik, 1992). 이후 변수 간의 관계가 비선형적인 경우에 원래 데이터를 고차원 공간으로 사상시킴으로써 변수 간의 관계를 선형적으로 나타내는 커널 함수가 소개되면서 다양한 경계면을 만들 수 있게 된다. 일반적인 커널의 예로는 다항식 커널, 가우시안 방사 기저 함수 커널, 라플라스 방사 기저 함수 커널 등이 있다(정성훈, 2020).

SVM은 구조적 위험에 대하여 최소화 원칙을 구현함으로써 과적합 문제에서 자유롭다는 점에서 장점을 갖는다. 또한, SVM은 블록함수를 최소화하는 학습을 진행하여 전역 최적해를 구할 수 있는데, 이런 점으로 인해 인공신경망보다 우수한 성능을 보이는 모형으로 주목받고

있다(Son et al., 2004; Ahn and Lee, 2009). 이에 따라, 최근에는 SVM을 활용한 다양한 연구가 진행되었으며 문서분류, 문자인식, 주가예측 등에서 뛰어난 일반화 성능을 보인 것으로 나타났다.

위와 같이 SVM은 분류문제에 주로 활용되지만, 임의의 실수를 예측할 수 있도록 SVM의 회귀 모형에 ϵ -무감도 손실함수(ϵ -insensitive loss function)를 도입한 서포트 벡터 회귀(support vector regression, SVR)가 등장하였다(Vapnik, 1998). SVR에서는 비선형 회귀 문제를 해결하기 위해 데이터를 고차원의 형상 공간으로 대응되고, 결과값과 연관된 선형 함수를 찾는 학습기법이다(Lu et al., 2009). 즉, SVR도 SVM과 마찬가지로 커널함수를 사용하여 훈련 데이터를 특징공간의 점으로 대응시키고 학습을 수행하지만, SVR은 클래스가 아닌 임의의 실수를 예측한다는 점에서 차별화된다.

2.2. 분석 자료

본 연구에서는 머신러닝 방법을 사용하여 돼지고기의 공급량을 예측하기 위해 통계청, 농협축산물정보센터, 축산물품질평가원, 축산과학원, 농림축산검역본부, 한국육류유통수출협회 등에서 보유하고 있는 돼지고기 생산 및 소비 관련 데이터를 확보하였다. 먼저 종속변수인 돼지고기의 공급량은 재고량, 생산량, 수입량의 합이다. 재고량은 국내산 재고량과 수입산 재고량의 합으로, 한국육류유통수출협회의 월간 육류유통실태조사와 농림축산검역본부를 통해 구득하였다. 생산량은 축산물품질평가원에서 제공하는 도축두수 데이터와 농촌진흥청 축산과학원에서 2004년에 발표한 돼지 도체 수율조사의 부위별 생산율을 통해 추정하였다. 수입량은 농림축산검역본부에서 제공하는 돼지고기 수입 데이터를 사용하였다. 본 연구는 돼지고기 공급량의 변동 요인을 재고량과 생산량의 측면으로 구분하여 변수를 선정하기 위해 관련 변수들의 데이터를 두 개의 그룹으로 나누어 구득하였다.

재고량 관련 속성은 총 8개이며, 변수 구성 및 정의는 <표 1>과 같다. 이용된 데이터는 2006년 5월부터 2014년 3월, 2015년 3월, 2015년 6월, 2015년 9월, 2015년 12월부터 2019년 9월까지의 자료이다. 국내 재고량의 경우 한국육류유통수출협회에서 회원사를 대상

으로 조사한 재고수치를 근거로 전체 국내산 재고량을 추정한 값이며, 수입 재고량의 경우 농림축산검역본부에서 제공한 전국 축산물검역시장 창고에 보관 중인 재고물량 조사 데이터를 사용하였다. 산지가격은 돼지 산지가격 조사 중단으로 2009년 3월부터는 전국도매시장 지육 평균경락가격을 기초로 환산한 수치이다. 도매가격은 탕박가격을 기준으로 한 값이며, 소비자 가격은 삼겹살 가격을 기준으로 산정한 값이다. 돼지고기 대체재는 닭고기로 선정하였다. 이는 농촌경제연구원이 주최한 ‘농업전망 2020’에서 발표된 조사결과를 참고하였다. 소비자 물가지수(costumer price index, CPI)는 소비자가 구입하는 상품과 서비스의 가격변동을 측정하기 위한 자료로서, 소비지출목적에 따라 분류한 지수 중 식료품 물가지수(food price index, FPI) 데이터를 추가하여 정확성을 높이고자 하였다. 변수 GNI(gross national income)는 1인당 국민총소득 데이터로서 명목 GNI를 인구수로 나누어 구한 값을 나타낸다. 재고량 관련 변수 중 가격과 관련된 변수는 수요에 영향을 미치는 변수임을 고려하여 데이터 기간을 조정하였다. 예를 들어 2021년 5월 재고량은 2021년 4월의 공급량에서 4월의 수요량(소비량)을 제외한 값이기 때문에 5월의 재고량을 구할 때에는 4월의 수요에 영향을 미친 변수와 대응시켜야 한다.

표 1. 재고량 관련 변수 구성 및 정의

변수명	단위	변수 설명	자료 출처
공급량	톤	돼지고기 재고량, 생산량, 수입량의 합	-
재고량	톤	국내산 재고량과 수입산 재고량의 합	한국육류유통수출협회 농림축산검역본부
월	월	월	-
산지가격	1,000원/110kg	1개월 전 농가가 받은 수취가격	농협축산물정보센터
도매가격	원/kg, 지육	1개월 전 전국 돼지고기 도매가격	
소비자가격	원/100g, 삼겹살	1개월 전 돼지고기 소비자가격	
대체재가격	원/kg, 지육	1개월 전 닭고기 소비자가격	
CPI	2015=100.0	소비자 물가지수	통계청
FPI	2015=100.0	소비자 식료품 물가지수	
GNI	명목, 만 원	1인당 국민 총소득	

생산량 관련 속성은 총 18개이며, 변수 구성 및 정의는 <표 2>와 같다. 이용된 데이터는 2003년 7월부터 2020년 12월까지의 자료이다. 돼지고기 생산량은 전술한 바와 같이, 축산

물품질평가원에서 제공하는 등급판정 두수 데이터를 성별과 도체중 구간별로 나누고, 각 구간의 평균 도체중에 농촌진흥청 축산과학원에서 2004년에 발표한 돼지 도체수율조사의 부위별 생산율을 곱하여 추정하였다. 사육두수 관련 변수 데이터는 2003년 7월부터 2016년 12월까지 농림어업총조사를 기반으로 한 통계 결과이며, 2017년 1월 이후의 자료는 돼지이력제를 기반으로 한 통계 결과이다. 총사육두수와 월령별 사육두수 변수들은 양돈산업에서 돼지의 사육기간이 일반적으로 180일인 것을 고려하여 데이터의 기간을 조정하였다. 예를 들어, 2021년 5월에 국내에서 사육한 2~4개월령 돼지는 당월에 도축되지 않고 일반적으로 2~4개월 이후에 도축되기 때문에 2021년 5월의 2개월령 돼지 사육두수는 2021년 9월의 공급량과 생산량과 대응시켜야 한다.

표 2. 생산량 관련 변수 구성 및 정의

변수명	단위	변수 설명	자료 출처
공급량	톤	돼지고기 생산량, 재고량, 수입량의 합	-
생산량	톤	(도체중 × (도체중 대비 부위별 생산율/100))의 합	축산물품질평가원, 축산과학원
월	월	월	-
총사육두수	마리	국내에서 사육하는 모든 돼지의 수	통계청
총사육두수.3		3개월 전 국내에서 사육한 모든 돼지의 수	
총사육두수.6		6개월 전 국내에서 사육한 모든 돼지의 수	
두수A.4		4개월 전 국내에서 사육한 2개월 미만 월령 돼지의 수	
두수A.6		6개월 전 국내에서 사육한 2개월 미만 월령 돼지의 수	
두수B.2		2개월 전 국내에서 사육한 2~4개월령 돼지의 수	
두수B.4		4개월 전 국내에서 사육한 2~4개월령 돼지의 수	
두수C		국내에서 사육한 4~6개월령 돼지의 수	
두수C.2		2개월 전 국내에서 사육한 4~6개월령 돼지의 수	
두수D.전체		국내에서 사육하는 6~8개월령 돼지의 수	
두수D.암컷		국내에서 사육하는 6~8개월령 돼지 중 암돼지의 수	
두수D.수컷		국내에서 사육하는 6~8개월령 돼지 중 수돼지의 수	
두수E.전체		국내에서 사육하는 8개월 이상 월령 돼지의 수	
두수E.암컷		국내에서 사육하는 8개월 이상 월령 돼지 중 암돼지의 수	
두수E.수컷		국내에서 사육하는 8개월 이상 월령 돼지 중 수돼지의 수	
방역치료비	원	비육돈 두당 사육비 중 평균 방역치료비	통계청
질병발생	0 또는 1	돼지열병, 아프리카돼지열병, 구제역 중 발생 보고 있음 = 1, 발생 보고 없음 = 0	농림축산검역본부

표 3. 전체 데이터 기초통계량

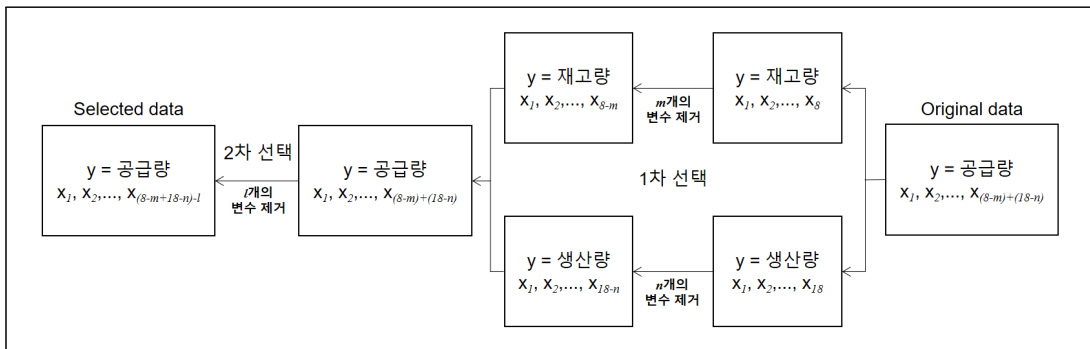
그룹	변수명	관측수	평균값	표준편차	최솟값	최댓값
-	공급량	144	189788.93	62850.76	99803.17	379520.33
-	재고량	144	64336.41	36631.68	18473.9	170730
재고량	월	144	6.45	3.43	1	12
	산지가격	144	341.75	77.71	185	581
	도매가격	144	4163.19	892.18	2538	7165
	소비자가격	144	1782.55	248.54	1292	2460
	대체재가격	144	5155.57	823.89	3302	6983
	CPI	144	94.50	7.85	80.013	105.65
	FPI	144	92	13.01	68.772	115.2
	GNI	144	2927.76	534.37	2069.984	3743.5
-	생산량	210	68784.33	13592.45	40684.28	100450.28
생산량	월	210	6.59	3.45	1	12
	총사육두수	210	9843409.97	1012073.7	7036116	11712607
	총사육두수.3	210	9814453.16	1005309.92	7036116	11712607
	총사육두수.6	210	9781053.31	992005.06	7036116	11712607
	두수A.4	210	3028770.32	317092.08	2149439	3660284
	두수A.6	210	3021473.67	3022738.45	2149439	3660284
	두수B.2	210	3045382.74	285300.37	2079354	3598490
	두수B.4	210	3040849.96	283332.43	2079354	3598490
	두수C	210	2740878.50	378759.76	2003077	3374192
	두수C.2	210	2731527.80	378014.97	2003077	3374192
	두수D.전체	210	103650.17	19096.62	70769	144370
	두수D.암컷	210	93766.53	16501.98	65885	127731
	두수D.수컷	210	9883.63	5287.35	2890	24950
	두수E.전체	210	908447.64	56037.64	688023	980577
	두수E.암컷	210	887688.74	54678.73	672289	961839
	두수E.수컷	210	20758.90	4732.90	15698	34355
	방역치료비	210	8148.77	2018.27	4217	10984
	질병발생	210	0.24	0.43	0	1

2.3. 분석 방법

본 연구에서는 의사결정나무, 인공신경망, 서포트 벡터 회귀 방법 각각에 대한 최적의 돼지고기의 공급량 예측모형을 구축하고, 각 최적 모형들의 성능을 비교한다.

본 연구에서는 모형별로 최적의 속성 집합을 도출하는 대신에, 다중 선행 회귀모델에 단계별 선택(stepwise selection)을 적용하여 각 모형에 학습시킬 변수들의 조합을 하나로 통일한다. 이는 변수들의 조합을 다르게 하면서 각 조합을 학습시켜야 하는 래퍼 기법(wrapper method)의 특성상 시간이 오래 걸리고 과적합(overfitting)의 위험성이 증가한다는 단점을 보완하기 위함이다. 예측하고자 하는 돼지고기의 공급량이 생산량, 재고량, 수입량의 합인 것을 고려하여, 생산량과 재고량을 종속변수로 하여 각각의 속성 집합을 1차적으로 구한다. 이후 공급량을 종속변수로 하고 두 속성 집합에 속한 속성들과의 상관관계를 2차적으로 분석하여 의사결정나무, 인공신경망, 서포트 벡터 회귀 모형에 적용할 변수 조합을 최종적으로 선택한다. 변수 조합을 선택하는 기준은 AIC(Akaike Information Criteria) 값으로 한다.

그림 1. 변수 선택 과정



머신러닝에 사용되는 변수들의 단위가 상이할 경우 학습되는 모형의 성능이 저하될 수 있으므로, 수치형 변수를 0과 1 사이의 값으로 정규화(normalization)하고 범주형 변수와 병합한다. 또한, 과적합 현상을 방지하기 위해 데이터는 7:3의 비율로 훈련 데이터(train data)와 검증 데이터(validation data)로 나눈다.

이어, 방법별 초모수를 다양하게 변화시켜가며 훈련 데이터를 학습에 사용하고, 검증 데이터에 대한 평균제곱근오차(root mean absolute error, RMSE)와 평균절댓값오차(mean absolute error, MAE)가 가장 작은 최적의 모형을 구축한다. 의사결정나무 모형은 교차 검증 수(xval)와 복잡성 매개지수(cp)를, 서포트 벡터 회귀 모형은 C(cost)와 입실론(epsilon)을 조정한다. 인공신경망의 경우는 은닉층의 개수를 3개로 고정하고 각 층의 노드 개수를 차례대로 조정한다.

3. 분석 결과

3장의 구성은 다음과 같다. 제1절은 속성 선택 결과를 정리하여 보여준다. 제2절은 알고리즘별 분석 결과를 정리한 것으로, 의사결정나무, 인공신경망, 서포트 벡터 회귀 순으로 설명한다. 이후 이어지는 4장에서 본 장에 정리된 내용에 대한 해석을 하며, 이를 요약하여 결론을 내린다.

3.1. 속성 선택

〈표 4〉와 〈표 5〉는 재고량과 생산량을 종속변수로 하여 각각의 속성 집합을 1차적으로 구한 결과이다. 재고량 그룹에 속한 변수들 중 선택된 변수들은 월, 산지가격, 도매가격, GNI이며 해당 조합의 ACI는 -663.52이다. 생산량 그룹에 속한 변수들 중 선택된 변수들은 월, 총사육두수, 총사육두수.6, 두수A.6, 두수B.4, 두수C, 두수D.전체, 두수E.전체, 두수E.암컷이며 해당 조합의 ACI는 -993.66이다.

〈표 6〉은 1차적으로 선정된 2개의 조합을 합하고 공급량을 종속변수로 하여 상관관계를 2차적으로 분석한 결과이다. 이때, 재고량 그룹의 1차 속성 집합 중 산지가격과 도매가격은 1개월 전의 가격으로 기간을 조정한다. 최종적으로 선정된 변수 조합의 ACI는 -761.81로, 산

지가격, 도매가격, GNI, 총사육두수, 두수A.6, 두수C, 두수E.전체, 두수E.암컷이 포함되었다. 이렇게 최종적으로 선택된 변수 조합 데이터의 구성과 기초통계량은 <표 7>과 <표 8>에 정리되어 있다.

표 4. 재고량 그룹의 1차 속성 선택 결과

변수 조합	ACI
산지가격, 도매가격, 소비자가격, 대체재가격, CPI, FPI, GNI, 월	-657.84
산지가격, 도매가격, 대체재가격, CPI, FPI, GNI, 월	-659.6
산지가격, 도매가격, 대체재가격, CPI, GNI, 월	-661.33
산지가격, 도매가격, 대체재가격, GNI, 월	-663.16
산지가격, 도매가격, GNI, 월	-663.52

표 5. 생산량 그룹의 1차 속성 선택 결과

변수 조합	ACI
총사육두수, 두수C, 두수D.전체, 두수D.암컷, 두수D.수컷, 두수E.전체, 두수E.암컷, 두수E.수컷, 총사육두수.3, 총사육두수.6, 두수A.4, 두수A.6, 두수B.2, 두수B.4, 두수C.2, 방역치료비, 월, 질병발생	-980.07
총사육두수, 두수C, 두수D.전체, 두수D.암컷, 두수D.수컷, 두수E.전체, 두수E.암컷, 두수E.수컷, 총사육두수.3, 총사육두수.6, 두수A.4, 두수A.6, 두수B.4, 두수C.2, 방역치료비, 월, 질병발생	-982.06
총사육두수, 두수C, 두수D.전체, 두수D.암컷, 두수D.수컷, 두수E.전체, 두수E.암컷, 두수E.수컷, 총사육두수.3, 총사육두수.6, 두수A.4, 두수A.6, 두수B.4, 두수C.2, 방역치료비, 월	-984.05
총사육두수, 두수C, 두수D.전체, 두수D.암컷, 두수D.수컷, 두수E.전체, 두수E.암컷, 두수E.수컷, 총사육두수.3, 총사육두수.6, 두수A.6, 두수B.4, 두수C.2, 방역치료비, 월	-985.77
총사육두수, 두수C, 두수D.전체, 두수D.암컷, 두수D.수컷, 두수E.전체, 두수E.암컷, 두수E.수컷, 총사육두수.3, 총사육두수.6, 두수A.6, 두수B.4, 방역치료비, 월	-987.49
총사육두수, 두수C, 두수D.전체, 두수D.암컷, 두수D.수컷, 두수E.전체, 두수E.암컷, 두수E.수컷, 총사육두수.3, 총사육두수.6, 두수A.6, 두수B.4, 월	-989.05
총사육두수, 두수C, 두수D.전체, 두수D.암컷, 두수D.수컷, 두수E.전체, 두수E.암컷, 두수E.수컷, 총사육두수.6, 두수A.6, 두수B.4, 월	-990.53
총사육두수, 두수C, 두수D.전체, 두수D.암컷, 두수D.수컷, 두수E.전체, 두수E.암컷, 총사육두수.6, 두수A.6, 두수B.4, 월	-991.51
총사육두수, 두수C, 두수D.전체, 두수D.수컷, 두수E.전체, 두수E.암컷, 총사육두수.6, 두수A.6, 두수B.4, 월	-992.17
총사육두수, 두수C, 두수D.전체, 두수E.전체, 두수E.암컷, 총사육두수.6, 두수A.6, 두수B.4, 월	-993.66

표 6. 2차 속성 선택 결과

변수 조합	ACI
산지가격.2, 도매가격.2, GNI, 총사육두수, 두수C개월, 두수D.전체, 두수E.전체, 두수E.암컷, 총사육두수.6, 두수A.6, 두수B.4, 월	-752.37
산지가격.2, 도매가격.2, GNI, 총사육두수, 두수C개월, 두수D.전체, 두수E.전체, 두수E.암컷, 총사육두수.6, 두수A.6, 두수B.4	-756.26
산지가격.2, 도매가격.2, GNI, 총사육두수, 두수C개월, 두수D.전체, 두수E.전체, 두수E.암컷, 두수A.6, 두수B.4	-758.25
산지가격.2, 도매가격.2, GNI, 총사육두수, 두수C개월, 두수D.전체, 두수E.전체, 두수E.암컷, 두수A.6	-760.5
산지가격.2, 도매가격.2, GNI, 총사육두수, 두수C개월, 두수E.전체, 두수E.암컷, 두수A.6	-761.81

표 7. 속성 선택 후 전체 데이터 구성

변수명	단위	변수 설명
공급량	톤	돼지고기 생산량, 재고량, 수입량의 합
산지가격.2	1,000원/110kg	2개월 전 농가가 받은 평균 수취가격
도매가격.2	원/kg, 지육	2개월 전 전국 돼지고기 평균 도매가격
GNI	명목, 만 원	1인당 국민 총소득
총사육두수	마리	국내에서 사육하는 모든 돼지의 수
두수A.6		6개월 전 국내에서 사육한 2개월 미만 월령 돼지의 수
두수C		국내에서 사육한 4~6개월령 돼지의 수
두수E.전체		국내에서 사육하는 8개월 이상 월령 돼지의 수
두수E.암컷		국내에서 사육하는 8개월 이상 월령 돼지 중 암돼지의 수

표 8. 속성 선택 후 전체 데이터 기초통계량

변수명	관측수	평균값	표준편차	최솟값	최댓값
공급량	144	189022.49	62395.93	99803.17	379520.33
산지가격.2	144	341.67	77.98	185	581
도매가격.2	144	4163.08	895.30	2538	7165
GNI	144	2932.61	530.85	2069.98	3743.50
총사육두수	144	9892431.41	1063066.39	7036116	11712607
두수A.6	144	3029770.29	319883.32	2149439	3618510
두수C	144	2771251.04	371810.39	2003077	3374192
두수E.전체	144	906049.98	64566.98	688023	980577
두수E.암컷	144	886539.80	63471.40	672289	961839

3.2. 모형별 학습 및 예측

3.2.1. 의사결정나무(Decision Tree)

의사결정나무 방법을 이용하여 최적의 모형을 도출하기 위해 xval(교차 검증 수)를 1부터 10까지, cp(복잡성 매개지수)를 0.001, 0.01, 0.1, 1로 조절하였고 minsplit(노드에 존재해야 하는 최소 관측치)은 (행의 길이)*0.05로 고정하였다. 그 결과로 xval이 10, cp가 0.001인 경우에 가장 오차가 낮은 것으로 나타났다. 속성 선택 전 데이터와 속성 선택 후 데이터를 각각 모형에 적용한 결과, RMSE와 MAE 모두 속성 선택 후 데이터를 적용한 모형이 더 낮은 것으로 나타났다.

표 8. 의사결정나무 모형의 초모수 설정별 성능 비교(속성 선택 후 데이터 적용)

초모수		RMSE	MAE
xval	cp		
1	0.001	0.06570	0.06570
⋮	⋮	⋮	⋮
10	0.001	0.04965	0.03920
10	0.01	0.05418	0.04211
10	0.1	0.06042	0.04464
10	1	0.08333	0.06981

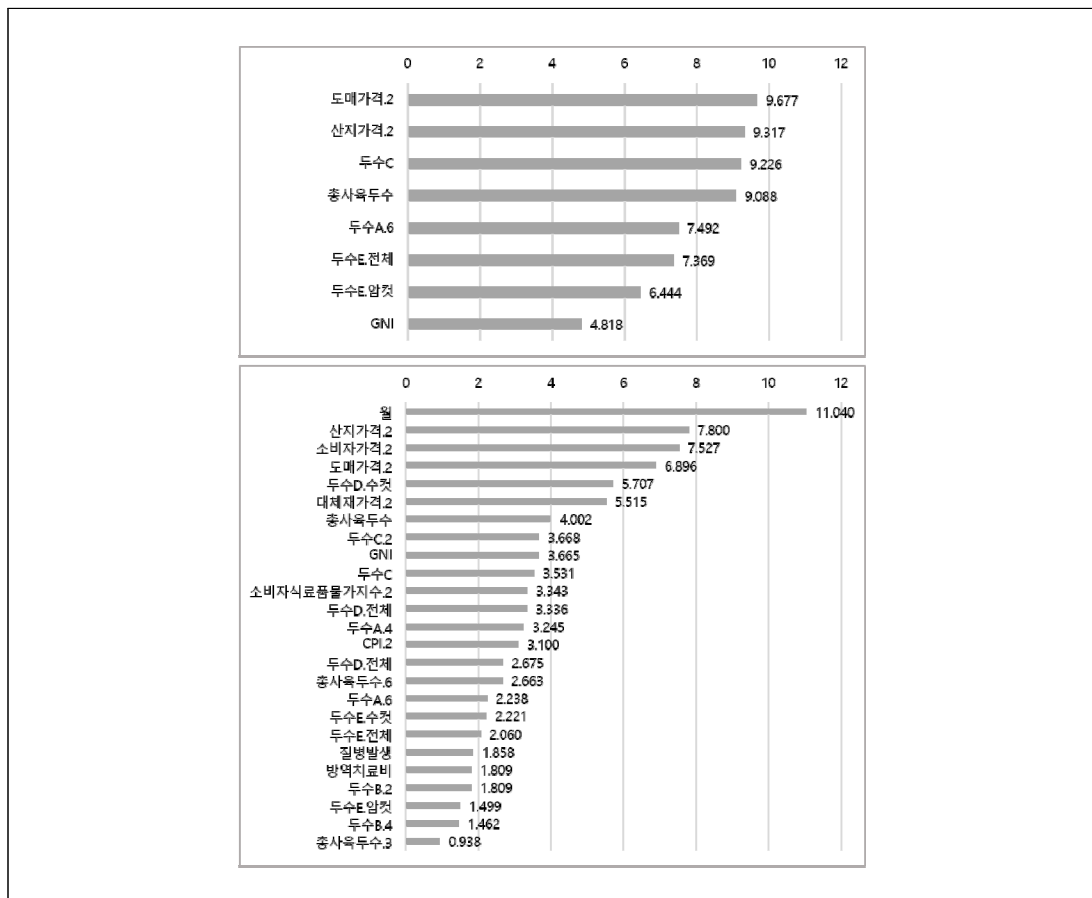
표 9. 의사결정나무 모형의 초모수 설정별 성능 비교(속성 선택 전 데이터 적용)

초모수		RMSE	MAE
xval	cp		
1	0.001	0.05158	0.04080
⋮	⋮	⋮	⋮
10	0.001	0.05093	0.04053
10	0.01	0.06042	0.04464
10	0.1	0.08333	0.06981
10	1	0.18930	0.16258

속성 선택 전 경우의 변수중요도는 월이 가장 큰 중요도를 나타내며 산지가격, 소비자가격, 도매가격, 두수D.수컷, 총사육두수, 두수C.2, GNI, 두수C, FPI, 두수D.전체, 두수A.4, CPI, 두수D.전체, 총사육두수.6, 두수A.6, 두수E.수컷, 두수E.전체, 질병발생, 방역치료비, 두수B.2, 두수E.암컷, 두수B.4, 총사육두수.3 순으로 중요성을 갖는다. 속성 선택 후 경우의 변수 중요도는 도매가격이 가장 큰 중요도를 나타내며 산지가격, 두수C, 총사육두수, 두수A.6, 두수E.전체, 두수E.암컷, GNI 순으로 중요도를 보인다.

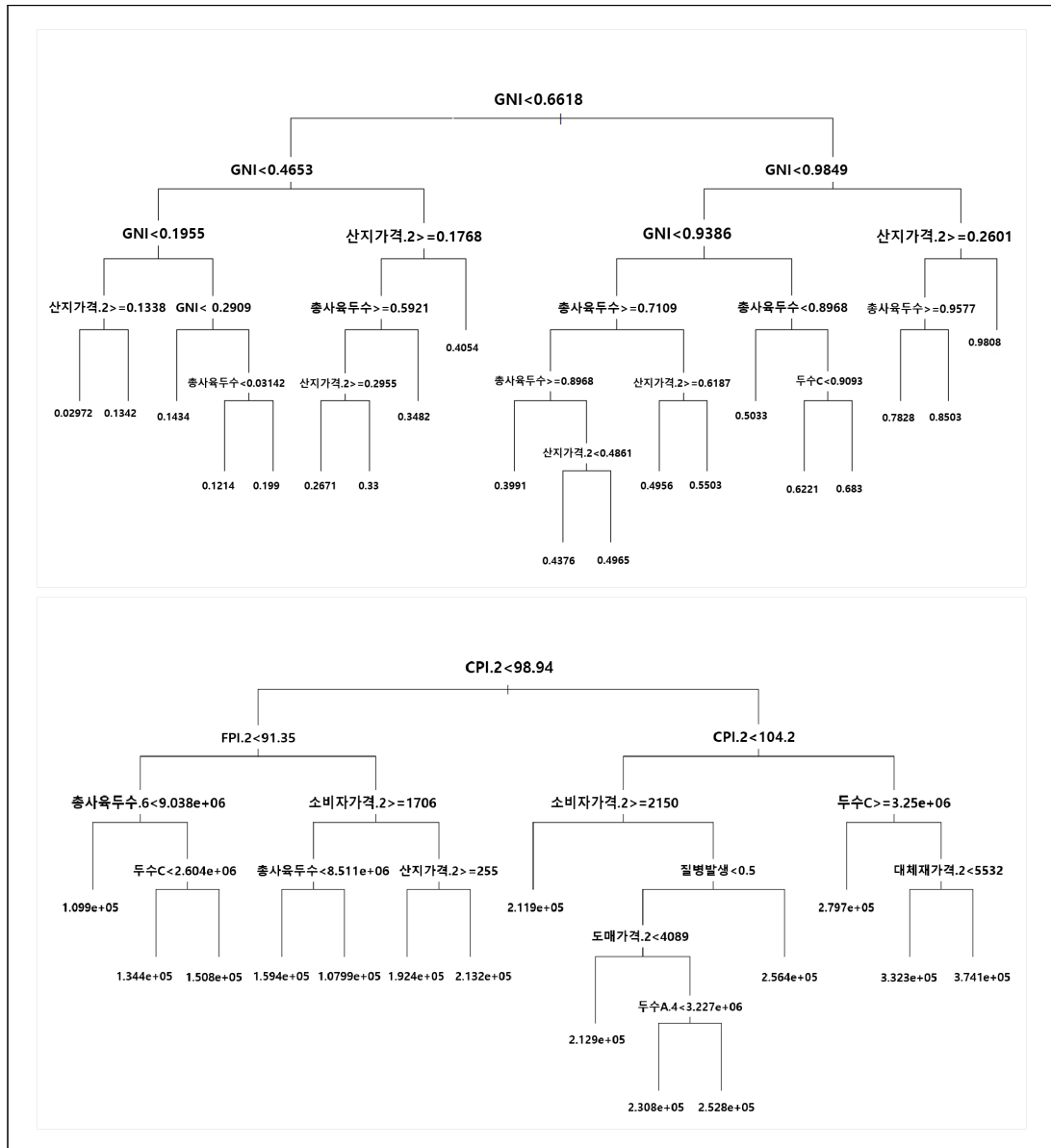
이때 속성 선택 후 변수의 중요도 순위가 달라진 것을 확인할 수 있으며, 이는 의사결정나무 모형이 데이터가 조금만 달라져도 나무의 전체적인 구성이 달라지기 때문이다.

그림 2. 의사결정나무 모형의 변수 중요도



주: 속성 선택 후 데이터 적용(위), 속성 선택 전 데이터 적용(아래)

그림 3. 의사결정나무 모형 시각화



주: 속성 선택 후 데이터 적용(위), 속성 선택 전 데이터 적용(아래)

3.2.2. 인공신경망(Artificial Neural Net, ANN)

인공신경망 방법을 이용하여 최적의 모형을 도출하기 위해 은닉층의 개수를 3개로 고정하고 각 층의 노드 개수를 1개부터 10개까지 차례대로 조절하였다. 또한, stepmax(최대 훈련 단계 수)를 1×10^6 , threshold(다음 뉴런으로 전파되는 역치)를 0.1로 고정하였다. 그 결과로 각 은닉층의 노드 개수가 5, 3, 1개인 경우에 오차가 가장 낮은 것으로 나타났다. 속성 선택 전 데이터와 속성 선택 전 데이터를 각각 모형에 적용한 결과, RMSE와 MAE 모두 속성 선택 전 데이터를 적용한 모형이 더 낮은 것으로 나타났다.

표 10. 인공신경망 모형의 은닉층 노드 개수별 성능 비교(속성 선택 후 데이터 적용)

노드 개수			RMSE	MAE
제 1 은닉층	제 2 은닉층	제 3 은닉층		
1	3	1	0.09084	0.06904
⋮	3	1	⋮	⋮
5	3	1	0.05252	0.04036
⋮	3	1	⋮	⋮
10	3	1	0.08909	0.06552
5	1	1	0.06879	0.05047
5	2	1	0.05789	0.04601
5	3	1	0.05252	0.04036
5	⋮	1	⋮	⋮
5	10	1	0.08240	0.05661
5	3	1	0.05252	0.04036
5	3	⋮	⋮	⋮
5	3	10	0.09275	0.07459

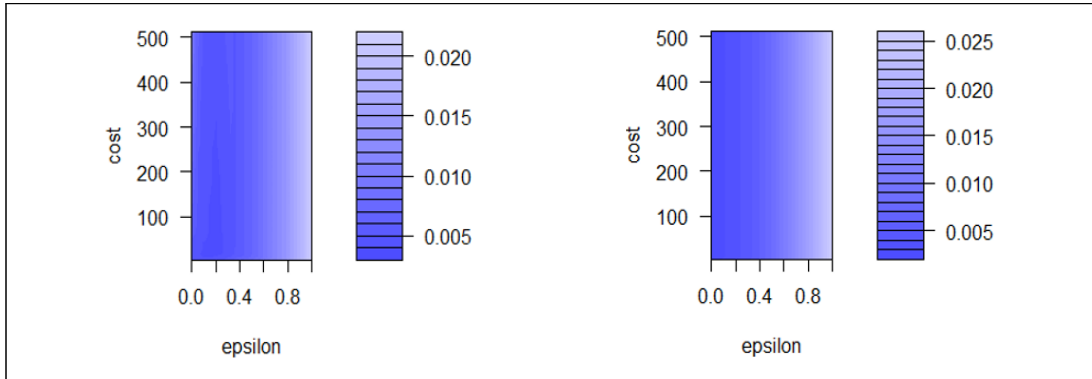
표 10. 인공신경망 모형의 은닉층 노드 개수별 성능 비교(속성 선택 전 데이터 적용)

노드 개수			RMSE	MAE
제 1 은닉층	제 2 은닉층	제 3 은닉층		
1	3	1	0.06185	0.04864
⋮	3	1	⋮	⋮
5	3	1	0.04988	0.04050
⋮	3	1	⋮	⋮
10	3	1	0.07811	0.05502
5	1	1	0.08373	0.06273
5	2	1	0.06766	0.05070
5	3	1	0.04988	0.04050
5	⋮	1	⋮	⋮
5	10	1	0.07870	0.06501
5	3	1	0.04988	0.04050
5	3	⋮	⋮	⋮
5	3	10	0.07107	0.05619

3.2.2. 서포트 벡터 회귀(Support Vector Regression, SVR)

서포트 벡터 회귀 방법을 이용하여 최적의 모형을 도출하기 위해 cost(과적합을 방지하기 위한 모수)를 0부터 500까지, epsilon(모수)을 0부터 1까지 조정하였으며, gamma(선형을 제외한 모든 커널에 요구되는 모수)는 변수의 개수에 따라 설정되는 기본값을 사용하였다. 그 결과로 속성 선택 전 데이터를 적용한 모형에서는 ‘gamma=0.0278, cost=32, epsilon=0’인 경우에 오차가 가장 낮은 것으로 나타났으며, 속성 선택 후 데이터를 적용한 모형에서는 ‘gamma=0.125, cost=32, epsilon=0.2’인 경우에 오차가 가장 낮은 것으로 나타났다. 또한, RMSE와 MAE 모두 후자의 경우가 더 낮은 것으로 나타났다.

그림 4. 서포트 벡터 회귀 모형의 초모수 설정별 성능

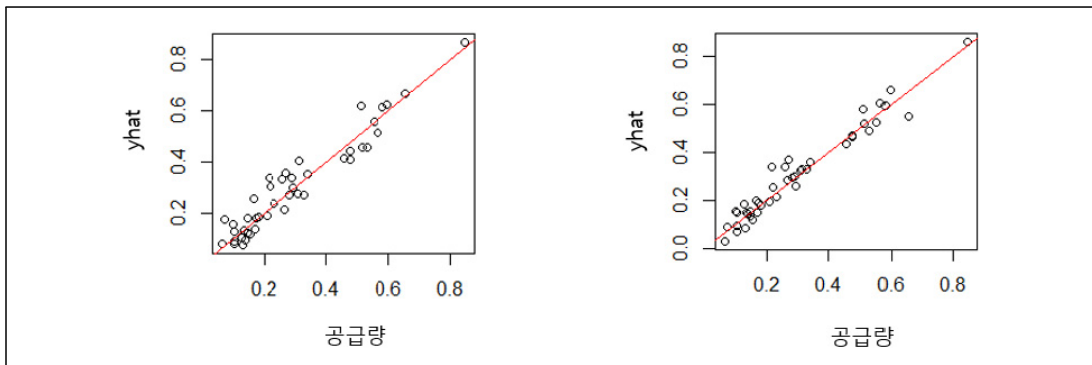


주: 속성 선택 후 데이터 적용(왼), 속성 선택 전 데이터 적용(오)

표 10. 서포트 벡터 회귀 모형의 초모수 설정별 성능 비교

데이터 구분	초모수			RMSE	MAE
	gamma	cost	epsilon		
속성 선택 후 데이터	0.125	32	0.2	0.05305	0.04220
속성 선택 전 데이터	0.0278	32	0	0.04305	0.03221

그림 5. 적용 데이터에 따른 서포트 벡터 회귀 최적 모형의 성능 비교

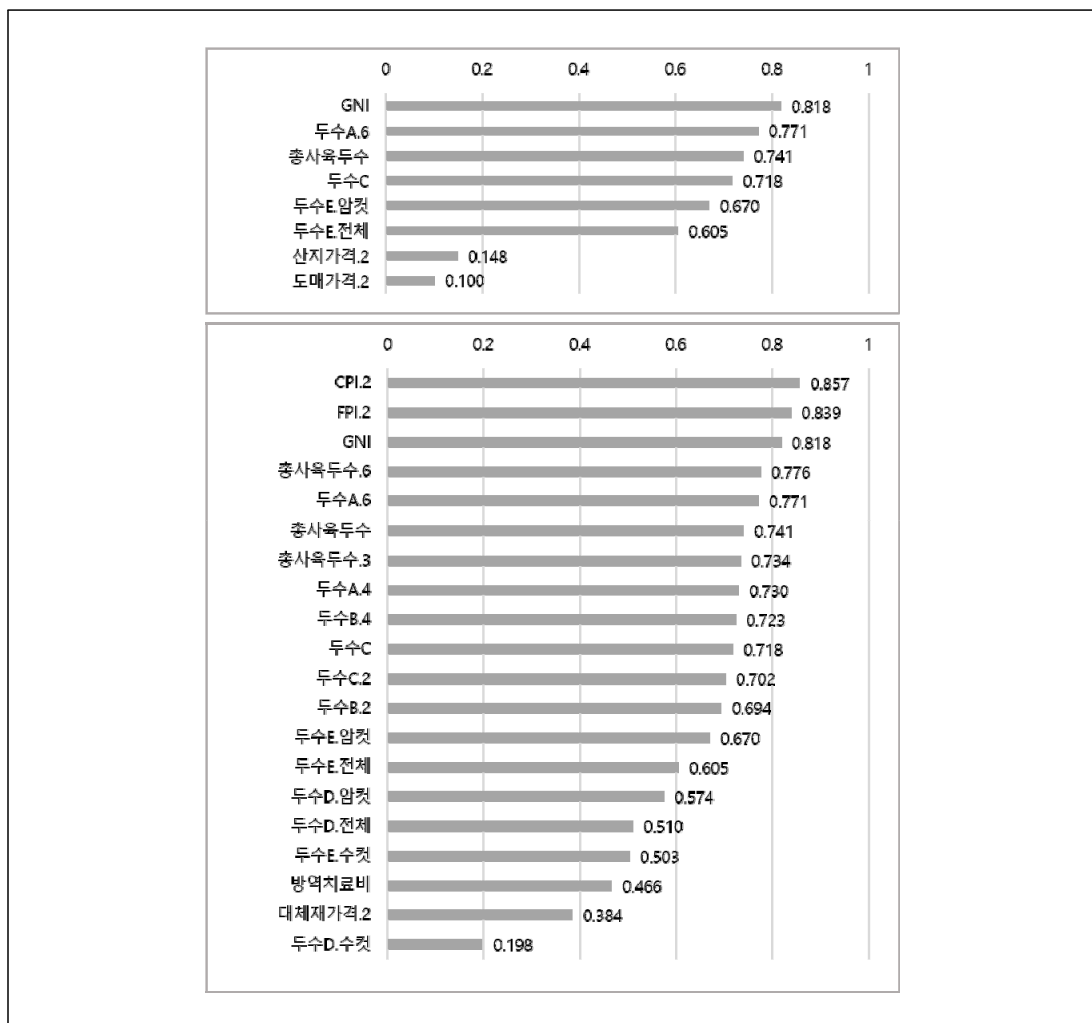


주: 속성 선택 후 데이터 적용, gamma=0.125, cost=32, epsilon=0.2(왼)
속성 선택 전 데이터 적용, gamma=0.0278, cost=32, epsilon=0(오)

속성 선택 전 경우의 변수중요도는 CPI가 가장 큰 중요도를 나타내며 FPI, GNI, 총사육두수.6, 두수A.6, 총사육두수, 총사육두수.3, 두수A.4, 두수B.4, 두수C, 두수C.2, 두수B.2, 두

수E.암컷, 두수E.전체, 두수D.암컷, 두수D.전체, 두수E.수컷, 방역치료비, 대체재가격, 두수D.수컷 순으로 중요성을 갖는다. 속성 선택 후 경우의 변수중요도는 GNI가 가장 큰 중요도를 나타내며 두수A.6, 총사육두수, 두수C, 두수E.암컷, 두수E.전체, 산지가격, 도매가격 순으로 중요도를 보인다. 이때 속성 선택 전과 후 데이터를 각각 적용한 모형에 모두 포함되어 있는 변수들의 순위가 변하지 않았으며, 그 중요도 값 또한 동일한 것을 확인할 수 있다.

그림 6. 서포트 벡터 회귀 모형의 변수 중요도



주: 속성 선택 후 데이터 적용(위), 속성 선택 전 데이터 적용(아래)

4. 결론

본 연구는 다양한 머신러닝 기법을 활용하여 돼지고기 공급량 예측 모형을 구축하고, 이들의 적용 가능성을 검토하였다. 돼지의 생산과정 및 사육 특성을 고려하여 변수를 선정하였으며, 이를 위하여 통계청, 농협축산물정보센터, 축산물품질평가원, 축산과학원, 농림축산검역본부, 한국육류유통수출협회 등에서 보유하고 있는 데이터를 확보하였다.

예측 모형을 구축하기 전에 다중 선형 회귀 모델에 단계별 선택(stepwise selection)을 적용하여 각 모형에 학습시킬 변수들의 조합을 사전에 선정하였다. 예측하고자 하는 돼지고기의 공급량이 생산량, 재고량, 수입량의 합인 것을 고려하여, 생산량과 재고량을 종속변수로 하여 각각의 속성 집합을 1차적으로 구하였다. 그 결과로 재고량 그룹에 속한 변수들 중 ‘월, 산지가격, 도매가격, GNI(국민총소득)’가 선정되었고, 생산량 그룹에 속한 변수들 중 ‘월, 총사육두수, 6개월 전 총사육두수, 6개월 전 2개월 미만 월령 사육두수, 4개월 전 2~4개월령 사육두수, 4~6개월령 사육두수, 6~8개월령 사육두수, 8개월 이상 월령 사육두수, 8개월령 이상 암돼지 사육두수’가 선정되었다. 두 속성 집합에 속한 속성들과 공급량 간의 상관관계를 2차적으로 분석한 결과, ‘산지가격, 도매가격, GNI, 총사육두수, 6개월 전 2개월 미만 월령 사육두수, 4~6개월령 사육두수, 8개월 이상 월령 사육두수, 8개월령 이상 암돼지 사육두수’가 최종적으로 선정되었다. 1차적으로 선택된 재고량과 생산량 속성 집합에 ‘월’ 변수가 모두 포함된 것으로 보아, 재고량은 돼지고기가 소비되는 계절적 패턴이 있고, 생산량은 계절에 따른 질병 발생률, 분만 및 폐사율로부터 기인한 패턴이 있는 것으로 추정된다. 또한, 2차적으로 선택된 속성 집합에 ‘월’ 변수가 제거된 것은 돼지고기 수급 안정화를 위해 수입량을 조절하면서 계절에 따른 공급량의 변동이 둔화된 것으로 판단된다.

속성 선택 후의 데이터를 적용하여 의사결정나무, 인공신경망, 서포트 벡터 회귀 방법 각각의 최적 모형을 구축한 결과, 최적 모형별 성능의 순위는 의사결정나무, 서포트 벡터 회귀, 인공신경망 모형 순으로 나타났다. 의사결정나무 모형의 경우, 속성을 학습 이전에 선택함에 따라 RMSE가 0.05093에서 0.04965로, MAE가 0.04053에서 0.0392로 오차가 낮아져 성능이 개선된 것으로 나타났다. 속성 선택 후 데이터를 적용한 의사결정나무의 RMSE와 MAE가

속성 선택 전 데이터를 적용한 인공신경망 모형보다 낮은 것으로 나타났지만, 서포트 벡터 회귀 모형보다 높은 것으로 나타났다. 따라서 성능이 가장 좋은 모형은 속성 선택 전 데이터를 적용한 서포트 벡터 회귀 모형인 것으로 나타났다. 다만, 속성 선택 이전 데이터를 적용한 경우 세 가지 모형 중에서 성능이 가장 낮았던 의사결정나무 모형이 속성 선택 이후 데이터를 적용한 경우에 성능이 가장 높은 것으로 보아, 속성 선택에 따라 의사결정나무의 성능이 크게 개선되는 것으로 판단할 수 있다.

본 연구의 의의는 다음과 같다. 첫째, 다양한 머신러닝 방법을 사용하여 돼지고기 공급량 예측 모형을 구축하고자 했다는 것이다. 돼지고기 수급구조에 관한 연구는 돼지 단독으로 진행된 경우가 드물고, 돼지고기의 유통구조와 소비 동향을 분석하여 돼지고기의 가격을 예측하는 것을 중심으로 이루어졌다. 반면에, 돼지의 사육 및 도축 과정을 중심으로 분석하여 국내 시장에 공급되는 돼지고기의 양을 산정하는 연구는 찾아보기 힘든 실정이다. 이에 본 연구는 돼지고기 수급 조절에 활용할 수 있는 돼지고기 공급량 예측치를 산정할 방법으로 의사결정나무, 인공신경망, 서포트 벡터 회귀의 적용 가능성을 검토하고자 했다는 점에서 의의를 갖는다. 다양한 머신러닝 방법들을 통해 돼지고기 공급량을 예측하고, 이를 가격 예측에 활용하고 수입량 조절 혹은 입식조절을 하는 등의 다양한 의사결정을 내릴 수 있을 것이다.

둘째, 각 모형별로 최적의 속성 집합을 도출하는 대신에, 다중 선형 회귀 모델에 단계별 선택을 적용하여 각 모형에 학습시킬 변수들의 조합을 학습 이전에 선택했다는 것이다. 속성 선택 전 데이터와 속성 선택 후 데이터를 적용한 모형들의 성능을 비교하여, 직관적으로 이해하기 쉬운 종속변수와 그 외 변수들 간의 선형관계를 기반으로 선정한 변수들을 다른 모형에 적용했을 때의 효과를 검증해볼 수 있다. 이는 변수들의 조합을 다르게 하면서 각 조합을 학습시켜야 하는 래퍼 기법(wrapper method)의 특성상 시간이 오래 걸리고 과적합(overfitting)의 위험성이 증가한다는 단점을 보완하기 위함이다.

본 연구는 의사결정나무, 인공신경망, 서포트 벡터 회귀 알고리즘을 활용하여 기계학습 기반의 수급예측 모형을 구축하고자 하였다. 하지만 최근 MARS, DNN, XGboost 등의 새로운 머신러닝 알고리즘들이 등장하고 있는 것을 감안하면, 이러한 신규 기계학습 방법론들의 적용 가능성을 검토하는 연구가 추가적으로 필요하다. 또한, 본 연구에서 학습 단계 이전에 사용한 속성 선택 방법은 다른 모형의 성능을 개선하는 유의한 효과가 없다. 속성 선택 후 데이

터를 적용한 의사결정나무 모형의 성능이 속성 선택 전 데이터를 적용한 인공신경망 모형보다 높은 것으로 나타났지만, 서포트 벡터 회귀 모형보다 낮은 것으로 나타났다. 따라서 속성 선택에 따른 성능 개선 효과가 모든 모형에게 적용되는 것으로 해석할 수 없다. 또한, 속성 선택 후 데이터를 적용한 모형의 변수 중요도에서 높은 순위를 차지하는 변수들이 속성 선택 과정을 통해 최종 선정된 변수 조합에 포함되지 않은 것으로 보아, 선택된 변수 조합이 각 모형의 성능을 최대화한 조합이 아닐 가능성이 높다. 그러므로 학습 단계 이전에 속성 선택을 위해 사용하는 알고리즘의 종류를 다양하게 검증하고 학습 및 예측에 사용하는 모형에 래퍼 기법을 적용한 변수 선정 과정을 거친 결과와 비교를 하는 후속 연구 또한 필요하다.

참고문헌

- 김규범, 김지옥, 신현정. 2019. “회귀나무(Regression tree)를 이용한 생활용(가정용 및 일반용) 지하수 이용량 추정 방법.” 지질학회지 55(6): 683-691.
- 김태훈, 홍한국. 2004. “회귀 모형과 신경망모형을 이용한 아파트 가격 모형에 관한 연구.” 국토연구: 183-200.
- 김혜영, 김배성. 2003. “육류 수요체계의 내생성과 구조변화.” 『농촌경제』 26(3): 39-53.
- 남국현, 최영찬. 2016. “돼지고기 부위별 수요함수 추정-수도권 소비자 중심으로.” Journal of Agricultural Extension & Community Development Vol.23 No.1. 27-37.
- 남영우, 이정민. 2006. “아파트시장예측을 위한 신경망분석 적응가능성에 대한 연구.” 한국건설관리학회 논문집 7.2: 162-170.
- 대한한돈협회. 2020. 「월간한돈」.
- 류상모. 2020. “인공신경망 채소류 가격·단수 예측 모형 연구.” (사)농식품신유통연구원. 농산물 가격* 단수 예측모형 고도화를 위한 조사용역.
- 심춘수, 정경수, 김민경. 2006. “한국 축산물 가격의 비대칭성.” 『농업경영 정책연구』 33(4): 1113-1126.
- 윤병삼. 2008. “육류시장의 가격변동성, 계절성 및 요일 효과.” 『농업경영 정책연구』 35(1): 21-38.
- 이슬기, 정성관, 이우성, 박경훈. 2011. “인공신경망을 이용한 도시기온 예측모형 구축.” 『국토계획』 46(1): 129-142.
- 이정희. 2001. “육류시장의 유통단계별 연계성 분석.” 『한국유통학회 동계학술대회 발표문』.
- 장재춘. 2021. “머신러닝 기법을 활용한 자영업자 폐업 예측 모형 연구: 서울시 25개 자치구를 중심으로.” (사)아시아문화학술원. 『인문사회21』 12(1).
- 정경수, 박종수, 김민경. 2007. “돼지고기 기초광고에 대한 가격효과 분석.” 『농업경영 정책연구』 34(2): 389-401.
- 정경수, 박종수, 김민경. 2007. “돼지고기 수요에 대한 일반광고 효과.” 『농업경영 정책연구』 34(3): 566-580.
- 정민국, 김현중. 2011. “쇠고기, 돼지고기 수급구조 분석 및 정책 시뮬레이션 모형 개발.” 『한국농촌경제연구원 기본보고서』 1-251.
- 정성훈, 진창하. 2020. “머신 러닝 방법을 이용한 오피스 임대료 산정-랜덤 포레스트, 인공신경망, 서포트 벡터 머신 활용을 중심으로.” 부동산학연구 26.2: 23-53.
- 지인배, 황운재, 이형우, 한봉희. 2015. “한우와 돼지고기 수요변화 요인 분석.” 『한국농촌경제연구원 정책연구보고서』 1-117.
- 최종후, 서두성. 1999. “데이터마이닝 의사결정나무의 응용.” 『통계분석연구』 4(1): 61-83.
- 통계청. 2020. 「통계로 본 축산업 구조 변화」.
- 한국농촌경제연구원. 2020. 「돼지고기 수급 및 유통체계 분석」.

- Ahn, H., and H. Y. Lee. 2009. "A combination model of multiple artificial intelligence techniques based on genetic algorithms for investment decision support aid: An application to KOSPI." *The e-Business Studies* 10.1: 215-236.
- Hearst, Marti A., et al. 1998. "Support vector machines." *IEEE Intelligent Systems and their applications* 13.4: 18-28.
- Lu, Chi-Jie, Tian-Shyug Lee, and Chih-Chou Chiu. 2009. "Financial time series forecasting using independent component analysis and support vector regression." *Decision support systems* 47.2: 115-125.
- Vapnik, Vladimir. 1992. "Principles of risk minimization for learning theory." *Advances in neural information processing systems*.
- Vapnik, V. 1998. "Statistical learning theory new york." NY: Wiley.
- Son, T. S., et al. 2004. "A Study on the Covert Channel Detection in the TCP/IP Header based on the Support Vector Machine." *Journal of The Korea Institute of Information Security and Cryptology* 14.1: 35-45.