The Journal of the Convergence on Culture Technology (JCCT)
Vol. 8, No. 4, pp.355-360, July 31, 2022. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2022.8.4.355
JCCT 2022-7-44

머신러닝 기반 고춧가루 원산지 판별기법

Detection of Red Pepper Powders Origin based on Machine Learning

유성민*, 박민서**

Sungmin Ryu*, Minseo Park**

요 약 최근 국내산 고추의 생산 비용 상승과 수입산 고추의 도입으로 고춧가루 원산지 허위표기 등의 피해사례가 속출하고 있다. 이에 따라 원산지를 신속하고 정확하게 판별하는 문제가 대두되었다. 기존의 고춧가루 원산지 판별법의 경우 무기 및 유기성분을 실험적으로 대조 및 분석하여 비용과 시간이 많이 든다는 한계가 있다. 이를 보완하기 위해, 본 연구는 머신러닝을 도입하여 국내산, 수입산 고춧가루 분류를 제안한다. 고춧가루에 포함된 53가지 성분에 대하여 머신러닝 모델을 설계하고 검증하였다. 본 연구를 통해 어떠한 성분이 원산지 판별 시 중요하게 활용되는지 파악 할 수 있었다. 추후 고춧가루뿐만 아니라 다양한 식품으로 확장하여 원산지 판별에 드는 비용을 보다 줄일 수 있을 것으로 기대된다.

주요어 : 고춧가루, 원산지, 머신러닝, 인공지능, 의사결정나무

Abstract As the increase cost of domestic red pepper and the increase of imported red pepper, damage cases such as false labeling of the origin of red pepper powder are issued. Accordingly we need to determine quickly and accurately for the origin of red pepper powder. The used method for presently determining the origin has the limitation in that it requires a lot of cost and time by experimentally comparing and analyzing the components of red pepper powder. To resolve the issues, this study proposes machine learning algorithm to classifiy domestic and imported red pepper powder. We have built machine learning model with 53 components contained in red pepper powder and validated. Through the proposed model, it was possible to identify which ingredients are importantly used in determining the origin. In the near future, it is expected that the cost of determining the origin can be further reduced by expanding to various foods as well as red pepper powder.

Key words : Red-Pepper Powder, Geographical Origin, Machine Learning, Artificial Intelligence, Decision Tree

Ⅰ. 서 론

고추는 우리나라에서 가장 많이 소비되는 부재료로 대부분 건조 후 가루의 형태로 보관 및 소비하고 있다. 고춧가루는 식품첨가용 향신료뿐만 아니라 조미료로서 1인당 연평균 소비량이 4kg에 달할 정도로 한국인의 식생활에 중요한 역할을 담당하고 있다. 그러나 최근 농촌 노동력 감소와 인건비 증가에 따른 생산 비용 상승과 수입산 고추의 도입으로 인해 지속적으로 국내 고추 재배 면적이 감소하는 추세이다. 이에 필요한 소비량

*준회원, 서울여자대학교 데이터과학전공 학사과정 (제1저자)
**정회원, 서울여자대학교 데이터사이언스학과 조교수 (교신저자)
접수일: 2022년 5월 26일, 수정완료일: 2022년 6월 23일
게재확정일: 2022년 7월 2일

Received: May 26, 2022 / Revised: June 23, 2022
Accepted: July 2, 2022
**Corresponding Author: mpark@swu.ac.kr
Dept. of Data Science, Seoul Women's Univ, Seoul, Korea

조달을 위해 수입산 고추의 수입량이 2000년 1,032톤에서 2010년 이후 36,000톤으로 급증하였다. 수입산 고추는 국내산 고추 가격 대비 평균 86%로 가격경쟁력이 높다[1].

수입산 고추 및 고춧가루 시장의 규모가 증가함에 따라 혼합 고춧가루 원산지 허위표기 사례, 원산지 허위표기 고춧가루 수출, 보따리상인 불법유통 등 피해사례가 빈발하고 있다. 특히 중국산 고춧가루의 경우 가격이 국내산 고춧가루에 비해 2배 이상 저렴하며, 고추씨를 함께 분쇄하여 단가를 낮추는 사례도 있다[2]. 또한 고춧가루의 높은 관세율로 인해 불법유통으로 통관검사 및 안정성 검사를 거치지 않아 위생상 문제도 발생하고 있다[3]. 저렴한 가격을 위해 고춧가루를 혼합하여 유통하는 경우도 빈번히 일어남에 따라 고춧가루의 원산지를 빠르고 신뢰성 있게 판별하는 것이 중요한 문제로 대두되어 왔다[4].

동일 품종의 고추라도 재배지역과 환경, 가공 조건에 따라 고춧가루의 품질과 맛이 크게 좌우된다[5]. 국립농산물품질관리원 원산지식별정보에 따르면 국내산 고춧가루는 붉은빛이 돌고 냄새가 약하며 부드러운 반면, 수입산 고춧가루는 붉은빛이 돌고 냄새가 강하며 거칠다. 그러나 사람의 감각만으로 고춧가루의 원산지를 판별하는 데에 한계가 있다. 현재 원산지 판별을 위해서는 시료의 성분을 분석하여 통계적 방식으로 판별하는 방식을 사용한다[6][7][8]. 또한 보다 간편한 판별을 위하여 외적 요소인 색도 차이 분석 연구[9] 등이 수행되었다.

구체적으로 살펴보면, 고춧가루에 포함된 유기성분 및 무기성분을 분석하여 원산지 판별에 활용하는데 일반적인 경우 통계기법으로 성분을 대조 및 비교한다. 이 방법은 많은 시간과 노동력이 든다[10]. 특히 시료의 양이나 성분의 종류가 많으면 분석의 정확도가 떨어질 가능성이 있다. 또 다른 방법으로는 고춧가루의 색도를 비교하는 방법이다. 시간이 단축된다는 장점이 있지만 정확도가 떨어진다는 한계가 있다[11]. 따라서 본 연구에서는 고춧가루에 포함된 53가지[12]의 분을 기반으로 원산지를 분류하기 위해 인공지능의 한 분야인 머신러닝 알고리즘[13]을 제안한다. 머신러닝의 대표적인 분류 알고리즘인 Logistic Regression(로지스틱 회귀), Decision Tree(의사결정나무), Random Forest(랜덤 포레스트), Support Vector Machine(서포트 벡터 머신)을 적용하여, 실험 및 검증하였다. 이를 통해, 고춧가루 원산지 분류에 가장 효과적인 모델과 중요 성분을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 원산지 판별 방법을 설명하고, 3장에서는 분류에 효과적인 머신러닝 알고리즘을 설명한다. 4장에서는 본 연구에서 제안하는 국내산 및 수입산 고춧가루 선별법을 기술한다. 5장에서는 제안하는 방법론의 검증 및 결과, 6장에서는 토의 및 결과를 요약한다.

II. 원산지 판별 방법에 관한 선행 연구

기존의 식품 원산지 판별에는 전자코, 근적외선 분광분석법 등이 활용된다. 연구 대상에 해당하는 고춧가루의 경우 선행 연구가 극히 적은 편이므로, 고춧가루 판별 연구를 포함하여 고춧가루에 적용가능할 것으로 판단되는 연구를 검토하였다.

근적외선 분광분석기로 고춧가루의 반사 스펙트럼을 측정한 연구의 경우, 측정 결과를 2차 미분하여 PLS (Partial Least Square analysis)로 통계 처리하여 원산지를 판별하였다. 한국산 시료와 중국산 시료를 95~100% 정확도로 구분하였으나, 데이터를 1차, 2차 미분하는 등 일부 복잡한 전처리 과정이 필요하다는 한계가 있다[10].

미량 원소를 분석할 수 있는 X-선 형광분석기를 이용해 숙지황의 원산지를 판별한 연구는 35종의 원소를 정준판별분석하였다. 92.3%의 정확도를 나타냈으나, 35종의 원소 데이터를 모두 활용해야만 정확도가 높아지는 경향을 보였다[6].

전자코는 시료에서 나오는 휘발성 화합물을 채취해 물질을 측정한다. 송이버섯의 원산지를 감별하는 연구의 경우, 측정된 모든 물질 중 14가지를 선택해 그 분포를 다중 판별 분석하여 원산지를 판별하였다. 이는 95%의 판별 정확도를 보였다. 그러나 전자코로 측정된 모든 물질 중 분석에 유의미한 영향을 끼치는 물질을 실험자가 직접 선택해야 한다는 단점이 있다[7].

동위원소 질량분석기를 이용하여 고춧가루의 안정동위원소비를 분석한 연구의 경우 동위원소의 비율을 다중분석하여 원산지를 감별하였다. 그러나 이 방식은 동위원소비를 각각 측정해 시간이 오래 걸리고, 하나의 동일원소비만으로 구분이 어려우며 여러 동일원소비를 다중분석 및 비교하여야 원산지 감별이 가능하다는 한계가 있다[8].

III. 분류에 효과적인 머신러닝 알고리즘

머신러닝은 크게 Supervised Learning(지도 학습)과 Unsupervised Learning(비지도 학습), Reinforcement learning(강화 학습)의 3가지로 구분할 수 있다[14]. 본 연구는 고춧가루의 원산지 판별이 목적이므로, 분류에 효과적인 지도 학습을 활용한다.

지도 학습은 입력에 대한 라벨값을 포함하고 있는 데이터의 집합을 통해 최적의 예측함수를 찾아내고, 이를 바탕으로 새로운 데이터에 대한 결과를 추정하는 방식으로 분류에 효과적인 알고리즘이다[15]. 본 장에서는 대표적인 분류 알고리즘을 살펴보고자 한다.: Logistic Regression(로지스틱 회귀)[16], Decision Tree(의사결정나무)[17], Random Forest(랜덤 포레스트)[18], Support Vector Machine(서포트 벡터 머신)[19]

1. Logistic Regression(로지스틱 회귀)

Logistic Regression은 사건의 발생 확률을 예측하는 모델로서, 반응변수가 0 또는 1인 이진형 변수에서 쓰이는 분류이다. 이에 따라, 두 개의 그룹 중 특정 그룹에 속할 확률을 추정하여 변수와의 상관관계를 파악하는데 효과적이다[16][20].

2. Decision Tree(의사결정나무)

Decision Tree는 하향식 나무 구조 모델을 만들어 데이터를 분류하는 방법으로, 본 연구와 같이 목표 변수가 범주형 변수인 경우, 불순도를 측정해 불순도가 적은 방향으로 자식 가지를 형성하는데 효과적이다. 다른 알고리즘과 달리 나무 구조 모양으로 시각화되기 때문에 알고리즘을 이해하고 설명이 가능하다는 장점이 있다[17][21].

3. Random Forest(랜덤 포레스트)

Random Forest는 트리 기반의 앙상블 기법으로 전체 데이터를 무작위로 샘플링하여 의사결정나무를 여러 개 만든 뒤, 그 결과를 종합하여 예측 성능을 높이는 앙상블 학습 기법이다. 데이터의 양과 고려해야 하는 요소가 많은 경우 뛰어난 성능을 보인다[18][22].

4. Support Vector Machine(서포트 벡터 머신)

Support Vector Machine은 분석 대상 자료를 성질이 유사한 그룹으로 분류할 때 쓰이는 모델로, 차원의 이동으로 새로운 분류기를 만드는 방법이다. 만들어진 분류 모델은 그룹 간의 경계로 표현되는데, 두 클래스 사이에서 가장 큰 마진을 가진 경계를 찾는 모델이다[19][23].

IV. 고춧가루 원산지 판별을 위한 머신러닝 알고리즘 설계

본 연구에서는 고춧가루의 원산지를 판별하는 데 효과적인 머신러닝 모델을 제안한다. 전체 프로세스는 아래 Figure 1과 같다.

1. Data Collection

본 연구에서는 질량분석법을 통해 수집한 국내산 및 수입산 고춧가루 115개의 53가지 무기성분 데이터를 사용하였다[12]. 데이터셋은 각 성분의 함유량을 담고 있다.

2. Data Preprocessing

1) Data Cleansing

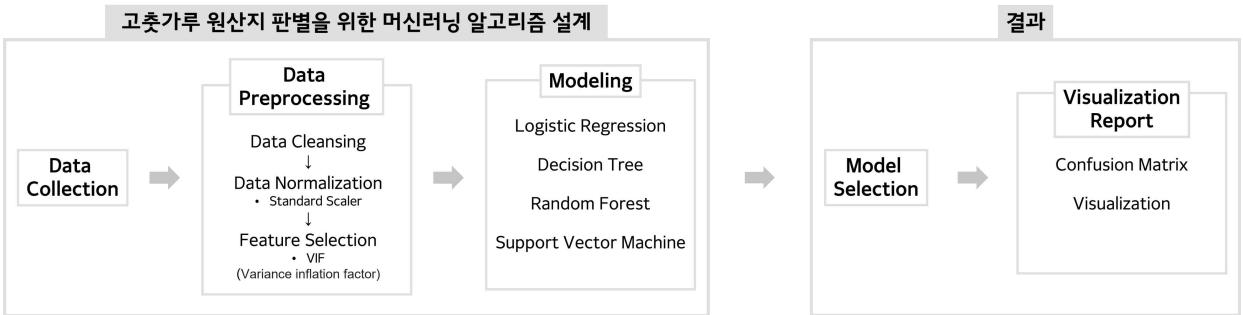

그림 1. 고춧가루 원산지 판별을 위한 머신러닝 알고리즘
Figure 1. ML-based Red-Pepper Powder Origin Detection Process

데이터셋 내에서 성분값이 0인 것이 전체의 50% 이상인 9Be, 101Ru, 125Te를 제거하였다. 변수 값에 0이 다수 존재하는 경우, 모델 학습에 악영향을 줄 수 있어 제외하는 것이 좋다[24].

2) Data Normalization

데이터의 편차가 클 경우, 모델 학습에 방해되기 때문에 이를 줄이기 위해 데이터의 정규화가 필요하다[25]. 데이터의 정규화 방법으로 Standard Scaler를 사용한다. 데이터의 평균을 0, 분산을 1로 변경하여 데이터가 표준정규분포를 따르도록 정규화하였다.

3) Feature Selection

머신러닝 모델은 독립변수와 종속변수 간의 함수를 찾는 과정이다. 모델의 정확도를 높이기 위해서는 변수 간 독립성이 중요하다. 변수 간의 상관 관계는 VIF(Variance Inflation Factor) 지표를 통해 파악할 수 있다. VIF가 10 이상인 경우, 변수 간 독립성이 낮으므로 변수 간에 영향을 미칠 수 있다[26]. 따라서, 본 연구에서는 VIF 10 이상인 변수를 제외하였다. 다음 Table 1은 VIF 값이 10 미만인 변수이며, 해당 변수들만 독립변수로 모델학습의 입력으로 사용되었다.

표 1. VIF 값이 10 미만인 최종 선정된 변수들
Table 1. Features without High Multicollinearity

Features	VIF	Features	VIF
^{182}W	2.386670	^{208}Pb	5.936461
^{95}Mo	2.936833	^{137}Ba	6.457995
^{78}As	3.254948	^{63}Cu	6.820831
^{61}Ni	3.552020	^{185}Re	7.337939
Na	3.560465	^{195}Pt	7.726395
^{181}Ta	3.872838	^{75}As	8.005710
^{52}Cr	4.022767	^{55}Mn	8.894728
^{121}Sb	4.351224	Zn	8.900033
^{107}Ag	4.563482	^{59}Co	9.459108
^{111}Cd	5.407838	P	9.512784
^{205}Tl	5.929047	^{72}Ge	9.845056

3. Modeling

전처리 된 데이터를 7:3의 비율로 Train set와 Test set로 분리하여 모델링을 진행하였다. 분류모델의 대표적인 알고리즘인 Logistic Regression, Decision Tree, Random Forest, Support Vector Machine을 적용하여 모델링을 진행하였다.

Ⅴ. 결 과

학습 결과는 아래 Table 2와 같다. 4가지 머신러닝의 대표적인 분류 모델 중, Decision Tree의 Train set score와 Test set score가 각각 1.00, 0.97로 가장 우수한 성능을 보였다.

표 2. 대표적인 분류 알고리즘 4가지의 성능비교
Table 2. Comparison with 4 ML Algorithms

Model	Train set score	Test set score
Logistic Regression	1.00	0.91
Decision Tree	**1.00**	**0.97**
Random Forest	1.00	0.91
Support Vector Machine	0.98	0.94

따라서, 본 연구에서는 성능이 가장 우수한 Decision Tree를 고춧가루 원산지 판별 모델로 채택하였다. Decision Tree에 대한 성능을 한번 더 검증하기 위해 Accuracy(정확도), Precision(정밀도), Recall(재현율), F1-score(F1 점수)를 사용하여 평가하였다(Table 3).

Accuracy(정확도)는 모델에 입력된 고춧가루 데이터 중 국내산과 수입산을 얼마나 정확하게 예측했는지를 의미한다. Precision(정밀도)는 예측값의 정확성을 나타내는 지표로, 국내산이라고 예측한 고춧가루 중 실제로 국내산 고춧가루의 비율을 나타낸다. Recall(재현율)은 실제 국내산 고춧가루 중 국내산이라고 예측한 비율을 나타낸다. F1-score(F1 점수)는 Precision과 Recall을 모두 고려한 지표이다.

본 연구에서 채택한 Decision Tree의 성능을 평가한 결과 Accuracy가 0.97, 국내산 예측에서 Precision 0.94, Recall 1.00, F1-score 0.97의 매우 우수한 성능을 보였다. 또한 수입산 예측에서도 Precision 1.00, Recall 0.95, F1-score 0.97로 우수한 성능을 나타냈다.

표 3. 의사결정나무의 성능 검증
Table 3. Decision Tree Performance Evaluation: Accuracy, Precision, Recall, F1-score

Accuracy	0.97		
	Precision	**Recall**	**F1-score**
Domestic	0.94	1.00	0.97
Imported	1.00	0.95	0.97

다음 Figure 2는 Decision Tree를 시각화한 그림이다. 53가지 성분 중 ^{63}Cu, Zn, P, Na가 원산지를 감별하는데 효과적임을 알 수 있다.

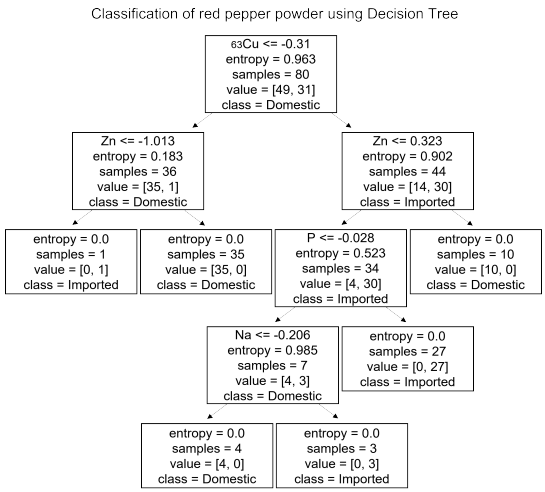

그림 2. 의사결정나무의 시각화
Figure 2. Visualization of Red-Pepper Powder Using Decision Tree

Ⅵ. 결 론

본 연구는 고춧가루의 원산지를 판별하기 위해 분류에 효과적인 머신러닝 알고리즘 중 Decision Tree를 제안하였다.

먼저 고춧가루의 질량분석법을 통해 얻은 53가지의 무기성분을 수집한 후, 변수 간 다중공선성을 제거하기 위하여 VIF 값이 10 미만인 변수만을 선택하였다. 또한 변수의 불균형을 줄이기 위해 데이터를 정규화하였다. 데이터 전처리 후, 최적의 분류 모델을 찾기 위해 머신러닝 기법 중 대표적인 분류 알고리즘인 Logistic Regression, Decision Tree, Random Forest, Support Vector Machine을 모두 활용하여 모델링 및 테스트하였다. 그 결과 Decision Tree가 가장 높은 정확도를 보였다. 국내산 예측의 경우 Precision이 0.94, Recall이 1.00, F1-score가 0.97로 우수한 성능을 보였으며, 수입산 예측의 경우에도 각각 1.00, 0.95, 0.97로 우수한 성능을 보였다. 모델을 설명하기 위해, Decision Tree를 시각화한 결과 ^{63}Cu, Zn, P, Na가 고춧가루 원산지 판별에 주된 영향을 미치는 성분임을 알 수 있었다.

본 연구는 고춧가루 원산지 판별에 미치는 주요 성분을 도출하였다는 데 의의가 있다. 주요 성분만을 집중적으로 검출하여 원산지 판별에 활용하면 저비용으로도 효율적인 판별이 가능할 것으로 판단된다.

해당 연구는 115개의 샘플만을 이용하여 분석을 진행하였기 때문에, 더 많은 시료를 통해 원산지를 보다 신뢰성 있게 판별할 수 있는 지속적인 연구가 필요하다. 또한 본 연구를 기반으로 국내산 및 수입산의 이항분류 뿐만 아니라 국내 및 해외 지역 단위의 세분화된 분류를 수행하는 후속연구가 필요하다. 이를 통해 고춧가루의 빠르고 명확한 원산지 추적이 가능하므로 보다 안전한 유통환경 조성이 가능할 것으로 사료된다.

4차 산업혁명의 발달과 더불어 인공지능 기술이 다양한 분야에서 활용되기 시작하였으며, 식품 분야에서도 활발한 연구가 수행되고 있다[27]. 본 연구를 통해 머신러닝 기법으로 식품의 원산지 판별이 효과적임을 실증적으로 검증하였다. 연구 결과를 기반으로 고춧가루 원산지 판별뿐만 아니라 다양한 식품으로 확대 적용이 가능할 것으로 판단된다.

References

[1] KOREA RURAL ECONOMIC INSTITUTE, “Causes and implications causes of the dried pepper industry in the crisis”, 2017.
[2] S. W. Kim, S. H. Song, S. J. No, J. S. Gang, S. Y. KIM and Y. M. Song, “Research on measures to investigate and improve red pepper powder mixed distribution”, KOREA RURAL ECONOMIC INSTITUTE, 2017.
[3] KOREA COAST GUARD, “Illegal importers caught using Chinese red pepper powder as seasoning”, 2021.
[4] B. S. Park, “Creating the geographical origin of agricultural products : discourses on NIRS analysis in the courts”, Graduate Program of Science Journalism, 2015.
[5] Y. R. Lim et al., “Effects of Drying Methods on Quality of Red Pepper Powder,” Journal of the Korean Society of Food Science and Nutrition, vol. 41, no. 9, pp. 1315 - 1319, Sep 2012. https://doi.org/10.3746/jkfn.2012.41.9.1315
[6] H. R. Bae, S. K. Lee, I. J. Whang, J. M. Kang, J. H. Lee, and J. H. Kim, “Discrimination of Geographic Origin by Trace Elements Contents inRehmannia Radix Preparatusing X-ray Fluorescence Analysis,” Journal of Applied Biological Chemistry, vol. 58, no. 4, pp. 345 - 348, Dec 2015. https://doi.org/10.3839/jabc2015.054
[7] N. Y. Lee, H. R. Bae, C. L. Lim, B. S. Noh, “Discrimination of Geographical Origin of Mushroom (Tricholoma matsutake) using Electronic Nose Based on Mass Spectrometry”. Food

Engineering Progress, vol. 10, no. 4, pp.275-279, 2006.

[8] H. J. Oh, G. A. Ko, M. Y. Yang, Y. J. Kim, “Stable Isotope Ratio Analysis for Origin Authentication of Red Pepper Powders and Soybeans“, Journal of the Korean Society of Food Science and Nutrition, vol. 50, no. 6, pp. 570-576, Jun 2021. https://doi.org/10.3746/jkfn.2021.50.6.570

[9] J. K. Chun, S. K. Park, “Originals:Color Measurement of Red Pepper Powder and its Relationship with the Quality”, Applied Biological Chemistry, vol. 22, no. 1, pp.18-23, 1979.

[10]H. S. Kwon, N. Y. Lee, S. J. Kim, S. S. Chung, J. H. Kim, “RESEARCH PAPER : Discrimination of Geographical Origin and Seed Content in Red Pepper Powder by Near Infrared Reflectance Spectroscopic Analysis”, Journal of the Korean Oil Chemists Society, vol. 16, no. 2, pp.155-315, 1999. https://doi.org/10.12925/jkocs.1 999.16.2.6

[11]S. Kang, K. Lee, J. Lim, C. Mo, J. Park, J. Kim. “Color difference between domestic and imported red pepper powder”. Korea Socierty for Agricultural Machinery, vol. 17, no. 1, pp.405-408, 2012.

[12]J. E Lee et al., “Distinguishing Korean and Chinese red pepper powder using inductively coupled plasma and X-ray fluorescence-based analysis.”, Food Science and Biotechnology, vol. 30, no. 12, pp.1497-1507, Sep 2021. https://doi.org/10.1007/s10068-021-00980-2

[13]S. H. Moon, “Analysis of AI-Applied Industry and Development Direction,” The journal of the convergence on culture technology, vol. 5, no. 1, pp. 77 - 82, Feb 2019. https://doi.org/10.17703/JCCT.2019.5.1.77

[14]S. E. Moon, S. B. Jang, J. H. Lee and J. S. Lee, “Machine Learning and Deep Learning Technology Trends”. Information & communications magazine, vol. 33, no.10, pp. 49.0-56.0, 2016.

[15]Cunningham, Pádraig, Matthieu Cord, and Sarah Jane Delany. “Supervised learning.” Machine learning techniques for multimedia. Springer, Berlin, Heidelberg, pp.21-49. 2008. https://doi.org/10.1007/978-3-540-75171-7_2

[16]WRIGHT, Raymond E. “Logistic regression”. 1995.

[17]Safavian, S. Rasoul, and David Landgrebe. "A survey of decision tree classifier methodology." IEEE transactions on systems, man, and cybernetics, vol. 21, no. 3. pp. 660-674, 1991. DOI: 10.1109/21.97458

[18]Breiman, Leo., “Random forests”, Machine Learning, vol. 45, no. 1, pp. 5-32, 2001, https://doi.org/10.1023/A:1010933404324

[19]Boser, Bernhard E., Isabelle M. Guyon, and Vladimir N. Vapnik. “A training algorithm for optimal margin classifiers.” In Proceedings of the fifth annual workshop on Computational learning theory, pp. 144-152. 1992. https://doi.org/10.1145/130385.130401

[20]J. R. Hosmer et al., Applied logistic regression. John Wiley & Sons, 2013.

[21]P. H. Swain, H. Hauska, “The decision tree classifier: Design and potential”, IEEE Transactions on Geoscience Electronics, vol. 15, no. 3, pp. 142-147, July 1977, DOI: 10.1109/TGE.1977.6498972.

[22]L. Breiman, “Bagging predictors. Machine Learning”. vol. 24, no. 2, pp.123 - 140 1996. https://doi.org/10.1007/BF00058655

[23]W. S. Noble, “What is a support vector machine?.” Nature biotechnology. vol.24, no.12, pp.1565-1567, 2006. https://doi.org/10.1038/nbt1206-1565

[24]S. B. Kotsiantis, D. Kanellopoulos, P. E. Pintelas, "Data preprocessing for supervised leaning." International journal of computer science, vol. 1, no. 2, pp.111-117, 2006.

[25]Raschka, Sebastian, “Python machine learning”, Packt publishing ltd, 2015.

[26]Alin, Aylin. “Multicollinearity.” Wiley Interdisciplinary Reviews: Computational Statistics, vol.2, no.3, pp. 370-374. 2010. DOI: 10.1002/wics.84

[27]S. Lee, M. Han, “Utilization and Analysis of Big-data,” International Journal of Advanced Culture Technology, vol. 7, no. 4, pp. 255 - 259, Dec 2019. DOI: 10.17703/IJACT.2019.7.4.255.