Journal of Korea Society of Industrial Information Systems Vol.28 No.5, Oct. 2023 :109-126
http://dx.doi.org/10.9723/jksiis.2023.28.5.109 ISSN:1229-3741

인적요인을 고려한 머신러닝 활용 산림화재 예측+

(Predicting Forest Fires Using Machine Learning Considering Human Factors)

장 진 명[1], 김 주 찬[2], 김 화 중[3]*, 김 광 태[4]
(Jin-Myeong Jang, Joo-Chan Kim, Hwa-Joong Kim, and Kwang-Tae Kim)

요 약 대형 산림화재를 예방하기 위해 산림화재의 조기발견은 매우 중요하다. 조기발견을 위한 하나의 방안으로 산림화재 발생 예측이 고려되고 있으며 다양한 관련 연구가 진행되었다. 그러나 대다수의 선행연구가 산림화재의 주요 발화 원인 중의 하나인 인적요인을 고려하지 않고 기상요인과 지리적 요인만을 주로 다루고 있다. 따라서 본 연구는 기상 및 지리적 요인뿐만 아니라 인적요인을 고려한 산림화재 예측모형을 개발하기 위해 2003년부터 2020년까지의 강원도 산림화재 데이터를 활용하여 로지스틱 회귀모형과 다양한 머신러닝 기법 기반의 예측모형을 개발하고 성능을 비교분석하였다. 성능분석 결과, 머신러닝 기법인 랜덤 포레스트(AUC=0.920)와 XG Boost 모형(AUC=0.925)이 가장 우수한 성능을 나타냈다. 운영시사점을 도출하기 위해 순열특성중요도 분석을 활용하여 요인들의 상대적 중요도를 분석하였으며, 기상요인이 인적요인보다 높은 영향도를 나타냈지만 다양한 인적요인도 유효한 것으로 확인되었다.

핵심주제어: 산림화재 예측, 머신러닝 기법, 로지스틱 회귀모형, 순열특성중요도 분석

Abstract Early detection of forest fires is essential in preventing large-scale forest fires. Predicting forest fires serves as a vital early detection method, leading to various related studies. However, many previous studies focused solely on climate and geographic factors, overlooking human factors, which significantly contribute to forest fires. This study aims to develop forest fire prediction models that take into account human, weather and geographical factors. This study conducted a comparative analysis of four machine learning models alongside the logistic regression model, using forest fire data from Gangwon-do spanning 2003 to 2020. The results indicate that XG Boost models performed the best (AUC=0.925), closely followed by Random Forest (AUC=0.920), both of which are machine learning techniques. Lastly, the study analyzed the relative importance of various factors through permutation feature importance analysis to derive operational insights. While meteorological factors showed a greater impact compared to human factors, various human factors were also found to be significant.

Keywords: Forest fire prediction, Machine learning model, Logistic regression model, Permutation feature importance analysis

* Corresponding Author: hwa-joong.kim@inha.ac.kr
+ 이 논문은 2020년 대한민국 교육부와 한국연구재단의 인문사회분야 중견연구자지원사업의 지원을 받아 수행된 연구임 (NRF-2020S1A5A2A01045577). Q-GIS를 활용하여 지도를 그려준 이성민 학부학생에게 감사를 표함.
Manuscript received September 09, 2023 / accepted October 09, 2023
1) 인하대학교 물류전문대학원, 제1저자
2) CJ대한통운 데이터 솔루션그룹, 제2저자
3) 인하대학교 아태물류학부, 교신저자
4) LG CNS Analytics&Optimization컨설팅팀, 제4저자

1. 서 론

전 세계적인 이상 기후로 인해 대형 산림화재가 빈번히 발생하고 발생빈도는 지속적으로 증가할 것으로 예측된다(Korea Forest Service, 2017). 우리나라는 국토의 약 60%가 산림으로 구성되어 있고, 특히 산림의 수목들이 단순림이고 밀도가 높아 산림화재의 위험성이 높은 상황이다(Kwak et al., 2010). 2011년부터 2018년까지 국내에서 발생한 산림화재는 연평균 474건이 발생하였으며, 산림파괴와 산림환경의 기능 손실, 산업 교란을 일으켰다(Korea Forest Service, 2020). 특히, 대형 산림화재는 진압이 어렵고 인접한 주거지에 큰 피해를 줄 수 있어 큰 주의가 필요하다(Chae et al., 2019).

대형 산림화재는 화재의 불똥이 비산하여 조기발견을 통해 신속하고 적절한 대응을 통해 충분히 예방할 수 있다. 하지만 대부분의 산림화재는 도심과는 멀리 떨어진 장소에서 발생하여 조기발견이 어렵고, 감시 범위를 넓히면 인력투입의 어려움이 존재한다(Lee and Sang, 2012). 따라서 산림화재의 대형화를 예방하기 위해서는 정교한 예측기법을 활용하여 산림화재의 가능성을 예측하고 해당 지역을 중심으로 한 감시활동이 필요하다.

대형 산림화재가 주로 발생하는 미국과 캐나다에서는 산림화재의 예측을 위해 기상 및 지리적 요인을 고려하고 있으며, 우리나라도 지역별 지형조건과 기상요인을 중심으로 산림화재 예측 시스템을 운영하고 있다(Won et al., 2018). 그러나 기상 및 지리적 요인은 일반적으로 산림화재의 발생요인이 아닌 확산요인으로 구분된다(Won et al., 2006). 두 요인을 고려한 예측모형은 산림화재 확산을 예방하는데 효과적이지만 직접적인 발생요인을 고려하지 않았다는 측면에서 한계점이 있다.

실제로 우리나라를 비롯한 전 세계 산림화재 대부분의 원인이 자연적 요인이 아닌 인적요인에 의해 발생하므로(Kwak et al., 2010; Won et al., 2010; Piao et al., 2022; Tariq et al., 2022), 발생요인으로 인적요인을 고려하는 것은 매우 중요하다고 할 수 있다. 인적요인은 기상 및 지리적 요인과 달리 지역의 특수성을 고려해야 하는 특징이 있다. 하지만, 인적요인은 명확한 고려요인과 선정 기준이 없어 상대적으로 많은 연구가 이루어지지 않았고, 기상요인과 지리적 요인을 활용한 연구가 주로 진행되었다(Romero-Calcerrada et al., 2008; Arndt et al., 2013; Pham et al., 2020).

따라서 본 연구는 인적요인을 추가로 고려한 산림화재 예측모형을 개발하고 다양한 분석을 통해 시사점을 도출하고자 한다. 이를 위해 지난 강원도 산림화재 원인을 분석하여 요인을 선정하고 다양한 산림화재 예측모형을 제시하였다. 산림화재 발생 여부에 관한 데이터 불균형 문제를 해소하기 위해 데이터 비율을 조정하고 다양한 샘플링 기법들을 적용하여 모형을 정교화 하였다. 또한, 제안한 여러 예측모델의 성능을 분석하기 위해 강원도 지역의 산림화재 발생자료, 기상자료와 인적자료를 활용하여 성능평가를 실시하였다.

본 논문의 구성은 다음과 같다. 2장에서는 산림화재 예측 및 감시 관련 선행연구를 살펴보고 본 연구의 차별성에 대해 기술한다. 3장은 예측에 활용한 자료와 인적요인에 대해 다루고 4장에서는 본 연구에서 활용된 방법론에 대해 상세히 기술한다. 5장은 수집한 자료에 관해 설명하고, 제안한 모델들의 성능분석 결과를 기술한다. 마지막으로 6장은 결론 및 시사점을 기술한다.

2. 문헌연구

산림화재 예측과 감시 관련 선행연구를 중심으로 문헌조사를 실시하였다. 기존 연구에서 기상요인과 인적요인을 함께 고려한 연구는 극히 제한적임을 확인할 수 있었다. 본 연구에서는 연구자 판단으로 선행연구 분석을 기상 관측 요인 및 인적요인을 고려한 연구로 구분하여 분석하고 본 연구의 의의와 선행 연구와의 차별성을 기술하였다.

2.1 기상 관측 요인에 관한 연구

과거 산림화재 발생과 기상관측자료를 활용한 논문을 살펴보면 Lee et al.(2004)은 산림화재 발생확률 모형의 개발을 위해 산림화재가 주로 발생하는 봄 시기를 대상으로 기온, 풍속, 상대습도, 운량 등 기상데이터를 활용하여 로지스틱 회귀모형을 개발하였다. 산림화재는 계절에 관계없이 발생하여 특정 계절에 국한되지 않는 예측이 필요하다. 하지만, 해당 연구는 특정 계절을 대상으로 하였고, 산림화재의 직접적인 발생요인인 인적요인을 고려하지 않았다.

Won et al.(2006)은 국내 산림화재의 직접적인 원인은 주로 인적요인에 의해 발생되지만, 기상요인이 연소 환경을 구성하는데 강한 상관관계가 있으므로 기상요인을 고려해야 한다고 주장하였다. 실제로 기상요인과 산림화재 발생에 따른 연관성 분석을 위해 봄 시기에 발생한 산림화재 데이터를 활용하여 회귀분석을 실시하였다. 해당 연구도 예측을 특정 시기(봄)로 한정하였고, 인적요인을 고려하지 않았다.

Preisler et al.(2007)은 미국 서부지역의 대형 산림화재 발생을 예측하기 위해 기상데이터를 활용한 로지스틱 회귀모형을 개발하였다. 월별 기상데이터와 월 Parmer 가뭄지수를 활용하였고, 산림화재 발생과 확산 가능성으로 구분하여 예측을 실시하였다. 이러한 연구는 대형 산림화재를 보다 정교하게 예측할 수 있다는 특징이 존재한다. 하지만 해당 연구는 산림화재 발생을 예측하는 연구와 다르게 대형 산림화재가 아닌 산림화재를 예측하는데 있어 정교함이 떨어질 수 있을 것으로 사료된다.

Gudmundsson et al.(2014)은 가뭄지수와 산림화재 발생의 관련성에 주목하였다. 남유럽지역의 산림화재 발생자료와 가뭄지수를 활용하여, 발생확률을 예측하기 위한 로지스틱 회귀모형을 개발하였다. 모델을 활용한 분석을 통해 각 지역에 적합한 요인을 고려할 필요성을 확인하였다. 해당 논문은 남유럽의 지역적 특성을 반영하고 월별 가뭄지수만을 고려하여 일별 예측, 감시 운영 편성 등의 현실 적용 관점에서는 적용의 한계점이 있다.

Won et al.(2012)은 산림화재 발생 예측을 위해 디지털 예보자료를 활용하여 기상관측자료를 수집하였다. 온도, 습도, 풍속을 활용하였으며 상대습도를 통해 실효습도를 산출하였다. 또한 Lee et al.(2004)이 개발한 로지스틱 회귀모형과 국립과학산림과학원에서 개발한 산식을 활용하였고, 산림화재 예측에서 디지털 예보자료의 활용성에 대한 검증을 실시하였다. 본 연구는 해당 연구에서 실시한 요인 검증뿐만 아니라 효과적인 예측모형의 개발이 목적이다. 이를 위해 로지스틱 회귀모형을 활용한 검증과 다양한 머신러닝 기법을 활용하여 예측모형을 개발하고 검증하였다는 측면에서 해당 연구와의 차별성이 있다.

Won et al.(2016)은 국내 산림화재의 발생확률 변화에 대한 주요 원인으로 기상변화를 주목하였다. 산림화재와의 연관성을 밝히기 위해 2000년대 봄 시기의 기상자료 중에서 유의성이 높은 최고기온, 상대습도, 실효습도, 평균 풍속을 선별하여 로지스틱 회귀모형을 활용하여 예측모형을 개발하였다. Won et al.(2016)은 Lee et al.(2004), Won et al.(2006)과 동일하게 봄 시기의 데이터를 활용하였지만 본 연구는 특정 계절을 대상으로 하지 않는다는 점에서 차별성이 있다.

Chae et al.(2018)은 산림화재에 취약한 강원도를 구역 9개로 구분하고 각 구역별로 산림화재 예측모형을 개발하였다. 캐나다산불기상지수 중에서 일부 지수를 파생변수로 활용하여 로지스틱 회귀모형과 머신러닝 기법인 RF(Random forest) 모형, XG Boost(eXtreme gradient boosting) 모형을 활용하였다. 또한 산림화재 발생 여부에 따른 데이터의 불균형을 해소하기 위해 샘플링 기법들을 적용하였다. 해당 연구는 확산요인인 기상 관측 자료만을 활용하였으므로, 발생요인인 인적요인과 관련한 운영시사점을 제시하기 어렵다는 한계점이 존재한다.

Kim et al.(2019)은 기상예보모델과 위성데이터를 활용하여 지리적 요인인 식생 건조지수를 추가로 고려하여 산불위험지수 모형을 개발하였고, 2019년 발생한 고성군 화재에 대한 실제 데이터를 활용하여 실증분석을 실시하였다. 해당 논문은 산림화재의 위험성을 다룬 논문으로 위험도의 높고 낮음을 판단하는 기준에 관해 세밀

한 조정이 필요하다는 한계점이 존재한다. 분석 결과에서 산불위험지수가 높아도 실제 발화의 원인이 없을 경우 산림화재가 발생하지 않은 것으로 나타났다. 이런 한계점은 실제 발생요인인 인적요인을 고려할 필요가 있다는 것을 뒷받침해 준다.

2.2 인적요인에 관한 연구

산림화재 발생자료와 인적요인을 활용한 선행연구로 An et al.(2004)은 국내의 산림화재발생의 주요 원인인 인적요인을 활용하여 산림화재 발생확률 모형을 개발하였다. 농경지, 산림지와 같은 인적요인을 고려한 지형적 특성을 반영하기 위해 GIS를 활용하여 공간분석을 실시하였으며, 로지스틱 회귀모형을 활용하여 예측모형을 개발하였다. 의성군 지역을 대상으로 산림화재와 관련성이 있을 것으로 판단되는 요인들을 고려하였다. 하지만 해당 논문은 산림화재의 직접적 요인인 관광객과 거주 인구가 아닌 농경지 또는 산림지와 같은 간접적 요인들로 구성하였다는 점에서 한계점이 존재한다.

Calef et al.(2008)은 미국 알래스카 지역의 산림화재와 인적요인의 관계를 분석하고 예측모델을 개발하였다. 거주지, 고속도로의 거리와 같은 인적요인에 대한 공간분석을 실시하였으며 로지스틱 회귀모형을 활용한 예측모델을 개발하였다. 분석을 통해 거주지에서 멀어질수록 산림화재의 발생확률이 증가하지만 도로에서 멀어질수록 발생확률이 감소하는 특성을 확인하였다. 이러한 연구 결과는 관광객과 같은 유동인구의 실화로 인해 산림화재의 대부분이 발생하기 때문으로 판단된다. 따라서 본 연구에서는 유동인구를 인적요인으로 고려하여 이러한 특성을 반영하고자 하였다.

Romero-Calcerrada et al.(2008)은 스페인 마드리드 지역의 산림화재 발생을 예측하기 위해 인구, 가축 밀도, 지역 간 거리 등의 데이터를 수집하고, 베이지안(Bayesian) 통계 기법인 증거 가중치 모델(Weights of evidence model)을 활용하였다. Ye et al.(2017)도 산림화재 예측을 위해 증거 가중치 모델을 활용하였으며, 해당 연구는 과거 윈난성의 산림화재에 대해 거주 위치, 도로, 농경지, 강의 위치 등 다양한 인적요인을 활용하여 공간분석을 실시하였다. 이러한 통계기법을 활용한 예측은 명시적인 모델을 제시한다는 점에서 장점이 있지만 광범위한 데이터를 다루는데 한계점이 존재한다(Ii, 2018). 따라서 본 연구에서는 광범위한 데이터를 다룰 수 있고 예측 분야에서 활용되는 머신러닝 기법을 활용하여 연구를 실시하였다.

Sadasivuni et al.(2013)은 미국 동남부 미시시피 지역의 산림화재 위험도를 예측하기 위해 인적요인과 산림자료를 함께 고려하였다. 인적요인으로 인구 수, 산림자료로는 산림밀집 및 생식연도를 고려하여 중력모델을 통해 위험도를 예측하였다. 일반적으로 중력모델은 요인별 상관관계가 미래에도 동일하다고 가정하는 한계점이 존재한다(Choi and Rho, 2015; Dicky, 2018). 이런 한계점은 산림화재 발생의 원인이 다양한 요인들이 복합적으로 작용한다는 점에서 모형의 활용도가 낮을 수 있다. 따라서 본 연구는 여러 요인을 복합적으로 고려할 수 있는 머신러닝 기법을 활용하였다.

Rodrigues and de la Riva(2014)은 스페인 지역의 산림화재 발생빈도를 예측하기 위해 산림화재 발생 데이터, 기계 밀집도, 농지, 전압선, 철도와 같은 다양한 인적요인을 활용하여 로지스틱 회귀모형과 다양한 머신러닝 기법을 적용하였다. 로지스틱 회귀분석과 비교하여 랜덤 포레스트 모형, BRT(Boosted regression tree) 모형, SVM(Support vector machines) 등 머신러닝 기법이 정확도 측면에서 상대적으로 좋은 성능을 도출하는 것을 확인하였다. 하지만 해당 연구에서 제안한 머신러닝 기법의 예측 정확도가 높지 않으므로 이를 개선해야 할 필요가 있음을 확인하였다. 이를 위해서 추가적인 요인 선정을 비롯한 다양한 접근방법이 필요하다. 본 연구에서는 머신러닝 기법의 예측 정확도 향상을 위해 기상요인뿐만 아니라 지역 특수성을 고려한 인적요인을 선별하였다.

Martín et al.(2019)은 스페인 동북지역을 대상으로 산림화재 예측모형을 개발하였다. 산림화재 발생자료와 도심지 거리, 농경지 거리, 전

선, 및 도로 거리 등 다양한 인적요인과 지표면 온도를 고려하였고, 해당 데이터를 계절과 휴무일로 분류하였다. 최대 엔트로피 알고리즘(Maximum entropy algorithm)을 활용하여 산림화재 발생을 예측하였으며, 선정된 요인들은 공간특성자료로 시간의 흐름에 따른 변화가 크지 않은 요인들로 구성하였다. 이는 산림화재 발생확률이 각 요인별 측정값의 변화를 활용하여 계산되므로, 시간의 변화에 따른 산림화재 발생확률의 변화가 크지 않다는 것을 의미한다. 본 연구에서는 시간에 따라 측정값이 변화되고, 산림화재의 직접적인 원인인 사람의 소실을 반영하기 위해 관광객과 같은 유동인구를 추가적으로 고려하였다. 특히 이러한 부분은 본 연구에서 분석한 인적요인을 고려한 기존 연구들의 공통적인 한계점이라 할 수 있는데, 산림화재의 위험도가 높은 지역들을 선별할 수 있지만 요인별 측정값의 변화가 크지 않아 실시간으로 산림화재를 예측하기 어렵다.

2.3 연구의의

선행연구 조사를 토대로 본 연구의 의의는 다음과 같다. 첫째, 산림화재 예측에 주로 활용되는 기상요인과 화재의 주요 원인인 인적요인을 함께 고려한다는 점이다. 본 연구의 저자들의 판단으로는 기상 및 인적요인을 함께 고려한 논문은 극히 일부에 불과하였으며, 인적요인과 기상요인을 함께 다루어도 기상요인 중 지표면의 온도를 고려한 연구이다(Martín et al., 2019). 그러나 산림화재의 예측 정확도를 높이기 위해 두 요인을 함께 고려해야 한다. 그러므로 본 연구에서는 일반적으로 통용되는 기상요인들과 지역 특성을 반영한 인적요인을 활용하였다.

둘째, 산림화재 예측에서 고려해야 하는 인적요인을 상세히 다루었다는 점이다. 산림화재는 주로 인적요인으로 인해 발생하지만 고려해야 하는 요인을 설정하는데 있어 명확한 기준이 존재하지 않아, 연구의 어려움이 존재한다(Romero-Calcerrada et al., 2008). 특히 지역의 특성을 고려한 인적요인을 선정할 필요가 있어 연구의 어려움이 더욱 크다고 할 수 있다. 그러나 선행연구 중에서 강원도 지역의 산림화재 예측을 다룬 연구에서 대부분 기상요인을 고려하였다. 또한 인적요인 중에서 공간적 특성만을 고려하였다. 본 연구는 강원도의 특성을 반영한 인적요인을 선별하고 고려했다는 측면에서 기존 연구와 차별성이 있다.

셋째, 예측모형에서 일반적으로 활용되는 로지스틱 회귀분석뿐만 아니라 머신러닝 기반의 예측모델을 개발했다는 점에서 의의가 있다. 산림화재 예측모형을 개발한 연구 중에서 Rodrigues and de la Riva(2014)와 Chae et al.(2018)를 제외한 대부분의 선행연구는 예측을 위해 주로 로지스틱 회귀분석과 같은 통계적 기법을 활용하였고, 통계적 기법은 예측보다는 추론에 초점을 둔다는 특징이 있다. 하지만 머신러닝 기법은 훈련 데이터의 양이 방대할수록 통계적 기법 대비 우수한 성능을 보이는 특징이 있다(Ii 2018; Levy and O'Malley, 2020; Yoo, 2021). 그러므로 상대적으로 산림화재 발생빈도가 높은 강원도 지역은 머신러닝 기법을 고려할 필요가 있다. 머신러닝 기법을 활용한 선행연구 중에서 Rodrigues and de la Riva(2014)과 Chae et al.(2018)은 인적요인과 기상요인을 함께 고려하지 않았다는 것이 한계점이다. 이런 부분은 일별 혹은 시간 단위로 변화하므로 산림화재 예측을 실시간으로 가능한 산림화재의 확산요인과 직접적인 발화 원인이지만 일별 혹은 시간 단위로 변화가 크지 않은 인적요인을 종합적으로 고려한 운영시사점 도출에 한계가 있다고 할 수 있다. 본 연구에서는 두 요인을 함께 고려하여 운영시사점을 제시하고자 하였다.

마지막으로, 산림화재가 상대적으로 적게 발생하는 국내의 특성을 보완하기 위해 실험 데이터의 비율을 조정하였으며 추가로 불균형 발생을 해결할 수 있는 오버샘플링 및 복합샘플링 기법을 적용했다는 측면에서 의의가 있다.

3. 자료수집 방법 및 요인 선정

3.1 자료(데이터)의 범위/종류/특성

강원도는 대한민국의 북동부에 위치하며, 약 156만 명이 거주하는 16,875km^{2}의 면적을 가진 지역이다. 또한 강원도는 전체 면적의 약 81%가 산림으로 구성이 되어 있으며 전국 산림 면적의 약 21%를 차지하고 있다(Chae et al., 2011). 2011-2020년을 기준으로 강원도의 연평균 온도는 11.42도이며, 최고기온은 29.4도, 최저온도는 -8.3도로 계절에 따른 온도 차이가 뚜렷하다. 연평균 강수량은 1216.8mm이지만, 여름의 평균 강수량은 680.9mm으로 여름을 제외한 다른 계절은 건조하다는 특징이 있다. 이런 특징은 강원도의 산림화재가 대형으로 쉽게 번질 수 있는 환경을 가지고 있다는 것이며 실제로 잦은 산림화재와 대형 화재가 발생하였다. 산림청에 따르면 2012년부터 2021년까지 강원도에서 산림화재는 연평균 72건이 발생하였고, 피해면적은 551.60ha였다(Forest Fire Service, 2022). 전국 기준으로 산림화재의 연평균 발생이 28건, 피해면적이 63.94ha인 것을 고려해보면 강원도 지역의 산림화재 발생건수와 피해면적이 상당히 높은 비중을 차지한다. 이는 강원도가 산림화재에 취약한 지역임인 것을 통계 자료가 뒷받침해 준다.

따라서 본 연구에서는 산림화재 예측을 위해 강원도를 분석 대상 지역으로 선정하였으며, 산림화재 발생을 예측하기 위해 산림빅데이터 거래소에서 제공하는 강원도 17개 행정구역(시군구)에서 발생한 산림화재 발생자료 중에서 화재 발생 위치의 좌표 정보가 있는 11년치(2005-2013, 2020-2021.06) 자료인 743건을 활용하였다.

3.2 설명요인

산림화재의 설명요인을 확산요인과 발화요인으로 구분하였으며, 발화요인의 선정을 위해 산림청과 산림빅데이터 거래소에서 제공하는 자료를 바탕으로 2003년도부터 2020년도까지 강원도 지역에서 발생한 산림화재 941건의 원인을 조사하였다. <Table 1>은 조사 결과를 정리한 표이

Table 1 Causes of Forest Fires in Gangwon-do (2003-2022)

Causes	No. occurrences	Percentage (%)	Accumulated percentage (%)
Population flow (tourist)	450	47.82	47.82
Residents (incineration of garbage)	97	10.31	58.13
Residents (incineration of crops)	95	10.10	68.23
Cigarette fiasco	52	5.53	73.75
Industrial fire	44	4.68	78.43
Residents (housing)	42	4.46	82.89
Military unit	33	3.51	86.40
Spontaneous ignition	26	2.76	89.16
Residents (boiler)	24	2.55	91.71
Residents (negligence)	24	2.55	94.26
Population flow (grave guest)	16	1.70	95.96
Residents (electric short circuit)	13	1.38	97.34
Arson (intentional)	11	1.17	98.51
Vehicle fire	6	0.64	99.15
Arson (playing with fire)	5	0.53	99.68
Etc.	3	0.32	100.00
Total	941	100%	100%

다. 강원도 지역 산림화재의 발생 원인 중에서 약 47.82%가 입산자 실화 등 관광객에 의한 화재였으며, 거주민들의 쓰레기 소각과 농작을 위한 소각이 각각 10.31%와 10.10%를 차지하였다. 이는 강원도 지역의 대부분이 산림으로 구성되어 있어 관광객의 방문이 많고 거주민들의 농업 종사 비율이 높기 때문으로 판단된다. 본 연구에서는 이런 특징을 반영하기 위해 관광 및 농업 관련된 인적요인들을 발화요인으로 선별하였으며, 확산요인은 불씨를 화재로 확산시킬 수 있는 요인으로 여러 연구에서 통용되고 있는 기상요인들을 중심으로 선정하였다.

본 연구에서는 강원도의 특성을 고려하여 관광객, 농업인구와 관련된 인적요인으로 농업인구 수, 관광객 수, 관광지와의 평균 인접거리를 설명변수로 선정하였고, 확산요인은 기온, 습도, 풍속, 강수량을 고려하였다. 산림화재 발생 원인이 관광객 및 거주민에 의한 산림화재가 주요 원인이므로 관광객 수와 농업인 수를 고려하였다. 또한 관광객들의 방문 목적지인 관광지와의 평균 인접거리를 추가로 고려하여 활동반경을 반영하였다.

4. 방법론

4.1 로지스틱 회귀모형

회귀분석은 설명변수와 피설명변수의 관계, 오차항에 대한 정보 등을 제공하며, 인과관계 분석과 예측에 많이 사용되는 방법론이다(Chatterjee and Hadi, 2015). 설명변수와 피설명변수의 관계를 선형으로 가정하며, 식 (1)과 같이 표현할 수 있다.

$$E(yx) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx \quad (1)$$

산림화재 예측모형은 피설명변수가 화재 발생 여부를 의미하는 이항 변수이므로 관계를 선형으로 가정할 수 없다. 그러므로 변수 간의 관계를 비선형으로 가정하고 회귀계수를 추정하는 로지스틱 회귀분석이 적합하다(Menard, 2002). 로지스틱 회귀분석에서 설명변수 x의 범위가 음의 무한대에서 양의 무한대를 가진다고 가정하면 확률은 0에서 1의 값을 갖게 되어 로지스틱 회귀분석의 추정은 사건의 발생확률을 추정하게 된다(Atkinson and Massari, 1998). 로지스틱 함수를 표현하면 식 (2)와 같다.

$$E(yx) = \frac{\exp(b_0 + b_1x)}{1 + \exp(b_0 + b_1x)} \quad (2)$$

식(2)에서 b_0와 b_1x은 비선형이지만, $E(yx)=P$로 가정하면 함수를 선형으로 변환이 가능하다.

$$P' = \ln\left(\frac{P}{1-P}\right) - \ln\left(\frac{E(yx)}{1-E(yx)}\right) \quad (3)$$

해당 식(3)을 식(2)에 대입하면 아래와 같이 선형화가 된다.

$$P' = b_0 + b_1x_1 \quad (4)$$

이러한 변환을 로지스틱 변환이라 하며, P'은 로짓(Logit)이라고 한다. 최소자승법을 활용하여 회귀계수를 추정하는 선형회귀분석과 달리 로지스틱 회귀분석에서는 최대 우도 추정법을 사용하여 계수를 추정한다.

4.2 머신러닝 기법

머신러닝 기법은 비선형 추정기법으로 수집된 데이터를 중심으로 복잡한 관계를 학습하고 분석하는 기법이다(Bishop, 2006). 머신러닝 기법은 분류 및 예측 분야에서 주로 활용되고 있으며, 금융, 보건, 강수량, 미세먼지, 생산성 등 다양한 분야에서 적용되고 있다(Lee et al., 2020; Kang et al., 2022; Lee et al., 2023). 앞에서 언급하였듯이, 산림화재 예측모형은 변수들을 선형으로 가정한다는 측면에서 한계가 있어 비선형 추정기법인 머신러닝 기법을 적용하는 것이 적절할 것으로 판단된다. 본 연구는 머신러닝 기법 중에서 신경망 모델인 다층 퍼셉트론 모형(Multi-layer perception, MLP), 의사결정나무

기법인 RF(Random forest) 모형, 앙상블 모형인 XG Boost 모형과 SVM 모형을 활용하였다.

4.2.1 다층 퍼셉트론 모형

다층 퍼셉트론 모형은 인공신경망의 한 종류로 입력층(Input layer)과 은닉층(Hidden layer), 출력층(Output layer)으로 구성된 신경망 모형이며 비선형 관계를 분석하는 모형이다(Noriega, 2005). 각 층들은 여러 노드로 구성되고, 노드 간은 서로 상호적으로 연결되어 입력값과 출력값을 계산하고 전달하게 된다. 각 층들을 살펴보면 먼저 입력층에는 학습을 위한 데이터가 배치되고 입력값(x_i)을 은닉층으로 전달한다. 은닉층은 가중치(w_{ij})와 바이어스($bias_i$), 활성함수(σ)를 활용 및 조정하여 식(5)에 의해 새로운 값으로 변환하게 된다.

$$\sum(PVH)=\sum_{i=1}^{n}\sigma\left(\sum_{j=1}^{m}x_i w_{ij}+bias_i\right) \quad (5)$$

여기서, PVH은 은닉층의 퍼셉트론 값, n은 은닉층의 노드 수, m은 출력층의 노드 수이다. 출력층에서는 실제 값과 비교하기 위해 식(6)을 활용하여 출력층 값을 계산하며, 결과값이 가장 적합한 값이 출력되도록 가중치를 변화시키며 반복적으로 학습을 수행한다.

$$y=\sigma\left(\sum(PVH)+bias\right) \quad (6)$$

4.2.2 RF 모형

RF(Random forest) 모형은 여러 모형의 예측치를 종합하여 하나의 예측치를 구하는 앙상블 모형이다. 부트스트랩(Bootstrap) 방식을 통해 다수의 표본을 생성하고 의사결정나무(Decision tree)모형으로 결과를 종합하는 방식이다(Biau and Scornet, 2016). 설명변수들을 무작위로 추출하고 결합하여 표본선택에 대한 무작위성을 최대한으로 부여한다는 특징이 있으며, 상관관계가 감소하게 되어 분류 예측의 안정성을 확보하게 된다. 또한, 확률추정이 아닌 데이터의 분류 결과를 바탕으로 예측을 실시하여 예측력이 높으며, 데이터 잡음과 과적합(Overfitting) 문제로부터 상대적으로 자유롭다는 특징이 있다(Siroky, 2009). 그러나 RF 모형은 수학적 관점에서 명확하게 밝혀지지 않은 단점이 존재하는 기법이다(Breiman, 2002; Biau et al., 2008). 이러한 단점에도 RF 모형은 많은 설명변수를 사용해도 높은 예측력을 가질 수 있다는 장점이 있어 빅데이터 분석과 예측에 매우 효과적인 기법이다.

4.2.3 XG Boost 모형

XG Boost(eXtreme gradient boosting) 모형은 의사결정나무의 부스팅 성능을 개선하기 위해 그래디언트 부스팅(Gradient boosting)을 적용한 앙상블 모형이다(Chen and Guestrin, 2016). 부스팅 기법은 앙상블 기법 중 하나로 트리 분류기를 순차적으로 실행하면서 이전 분류기의 오차에 가중치를 반영하여 성능을 높이는 과정으로 이루어진다. 이 과정에서 기존의 부스팅 방식은 분류기의 순차적 학습으로 속도가 느려지는 단점이 있다. 하지만 XG Boost는 부스팅 기법을 병렬처리하여 학습속도를 개선하고, 동시에 과적합 문제도 해결하기 때문에 일반적으로 예측 또는 분류에 활발히 사용되고 있다. 또한, 대용량 데이터에서도 안정적으로 작동하여 의료, 경제, 마케팅 등 다양한 분야에서 예측모형으로 활용되고 있다(Hah et al., 2019; Liang et al., 2019; Du et al., 2020).

4.2.4 SVM 모형

Cortes and Vapnik(1995)이 제시한 SVM(Support vector machine)은 분류와 회귀 문제 해결에 활용되는 통계적 학습 이론의 머신러닝 기법이다. 일반적으로 SVM은 범주를 분류하는 최적의 초평면(Hyperplane)을 찾기 위해 서포트 벡터(Support vector)와 초평면 간의 거리(Margin)를 최대화하는 과정으로 이뤄진다(Cortes and Vapnik, 1995). 여기서 최적의 초평면과 가장 가까운 훈련 데이터를 서포트 벡터라고 부르며 SVM 모형은 서포트 벡터만 계산과정에 포함하므로 학습속도가 짧고, 커널함수(Kernel function)를 활용하여 선형 분리가 불가능한 데이터는 고차원으로 변형하여 비선형 데이터도 분류가 가능한 장점이 있다.

4.3 샘플링 기법

샘플링 기법은 학습데이터의 불균형으로 발생하는 데이터 과적합 문제를 해결하기 위한 기법으로 언더샘플링과 오버샘플링으로 분류된다(Kim et al., 2014). 오버샘플링 기법은 기존 데이터의 특성을 활용하여 새로운 데이터를 생성하며 기존의 모든 데이터를 활용하므로 학습 시간이 증가한다. 반면 언더샘플링은 기존의 데이터 수를 감소시켜 학습 시간을 감소시킬 수 있지만 데이터 손실이 발생한다. 본 연구에서는 두 기법 중에서 데이터 손실이 발생하지 않고 상대적으로 성능이 좋은 예측모형을 만들 수 있는 오버샘플링 기법인 SMOTE(Synthetic minority oversampling technique), ADASYN(Adaptive synthetic sampling) 기법과 두 샘플링 기법의 단점을 상호보완 하는 복합샘플링 기법인 SMOTE-ENN, SMOTE-Tomek기법을 활용하였다(Mohammed et al., 2020).

4.3.1 오버샘플링 기법

오버샘플링은 학습에 필요한 데이터를 충분히 확보하기 위해 적은 레이블을 가진 데이터 집합 S_{minor}에서 무작위로 데이터를 추출하고 기존 데이터 집합 S에 추가하는 기법이다(He and Garcia, 2009). 오버샘플링은 학습 시간의 증가뿐만 아니라 데이터를 단순 복제하는 과정에서 과적합 문제를 야기할 수 있어 이를 보완하기 위한 다양한 기법들이 개발되었다.

SMOTE는 과적합 문제를 완화시키는 대표적인 오버샘플링 기법으로 데이터를 단순 복제하는 것이 아닌 기존 데이터를 활용하여 새로운 인공 데이터를 생성한다(Chawla et al., 2002). SMOTE는 S_{minor}내 개별 데이터들의 k-최근접 이웃(k-nearest neighbor)을 탐색하고 개별 데이터와 k개 이웃들을 보간하여 새로운 데이터를 합성하는 과정으로 이루어진다. 또 다른 대표적인 기법인 ADASYN은 S_{minor}내 개별 데이터들을 생성하는 데이터 개수가 동일한 SMOTE와 달리 각 데이터들마다 다른 양의 샘플 데이터를 생성하는 기법이다(He et al., 2008). 이 과정에서 데이터들의 밀도분포를 이용하여 인공 데이터를 합성함으로써 왜곡된 샘플 데이터의 분포를 방지하고 더 현실성 있는 데이터를 생성한다는 특징이 있다.

4.3.2 복합샘플링 기법

복합샘플링은 먼저 데이터에 오버샘플링을 수행하고 이후에 언더샘플링 기법을 수행하는 방식이다. 두 기법의 단점인 계산시간 증가와 데이터 손실 문제를 최소화할 수 있는 기법이다. 특히 SMOTE 기법을 토대로 한 SMOTE-Tomek과 SMOTE-ENN 기법은 소수의 양성 레이블만 존재하는 불균형 데이터에서 좋은 예측 정확도가 나올 수 있게 머신러닝을 학습시키는 샘플링 기법으로 알려져 있다(Batista et al., 2004). 먼저 SMOTE-Tomek은 SMOTE와 Tomek Links의 합성어로 SMOTE로 오버샘플링을, Tomek Links로 언더샘플링을 수행한다. 여기서 Tomek Links는 소수 레이블을 가진 S_{minor}의 샘플 데이터와 다수 레이블을 S_{major}의 샘플 데이터 사이의 데이터 중 S_{major}에 속하는 데이터만 제거하는 방법이다(Ivan, 1976). SMOTE-ENN은 SMOTE와 ENN(Edited nearest neighbors)를 결합한 복합샘플링 기법이다. ENN은 S_{minor}의 k-최근접 이웃을 찾아 이웃한 데이터 중에서 S_{major}의 데이터 수가 많을 경우 해당 데이터들을 제거해주는 과정으로 이루어진다(Wilson, 1972). 상기한 방식들은 가용 데이터의 수를 최대한 활용함과 동시에 S_{minor}와 가까운 S_{major}를 집중적으로 제거하여 데이터 분류 성능을 높인다는 특징이 있다.

5. 실험 자료 및 결과

5.1 실험 자료

기상요인 데이터는 기상청의 기상자료개방포털에서 제공하는 동네예보 초단기실황 자료를 활용하였고, 산불발생일시, 발생지역을 기준으로 시간 단위 데이터를 수집하였다. 인적요인인 농업인 수, 관광객 수, 주요 관광지와의 평균 인접거리를 고려하기 위해 농업인구와 관광객 수는

Table 2 Summary of AUC Results by Model and Sampling Technique

Sampling Technique	LR	MLP	RF	XG Boost	SVM
Without Sampling Technique	0.67	0.77	0.92	0.93	0.78
SMOTE	0.64	0.79	0.87	0.88	0.77
ADASYN	0.63	0.78	0.87	0.88	0.78
SMOTEEN	0.62	0.81	0.87	0.86	0.78
SMOTETomek	0.65	0.80	0.87	0.87	0.78
Mean	0.64	0.79	0.88	0.88	0.78

* LR: Logistic Regression
* MLP: Multi-Layer Perceptron
* RF: Random Forest
* SVM: Support Vector Machine

농림축산식품부에서 제공하는 농업 경영체 현황 자료와 관광지식정보시스템에서 제공하는 주요 관광지 입장객 통계를 활용하였으며 농업인 수는 연별, 관광객 수는 월별로 각각 산정하였다. 산림화재 발생자료와 기상자료는 시간 단위 데이터이므로 단위를 맞추기 위해, 월별 관광객 수를 시간 단위에 대한 평균값으로 산출하였다. 또한 농업인구는 시간별 변동성이 작은 것으로 판단하여 연도별 인구수와 시간별 인구수가 동일하다고 가정하였다. 다음으로 관광지와의 평균 인접거리는 관광지 좌표와 산불발생위치 좌표 간의 유클리디언 거리를 상용 GIS 패키지인 Q-GIS를 활용하여 도출하였다. 관광지와의 평균 인접거리도 시간대별로 인접거리가 동일하다고 가정하였고, 앞에서 제시한 설명요인 7개를 활용하여 로지스틱 회귀분석과 다층 퍼셉트론 모형을 포함한 총 5개의 모형에 대한 실험을 진행하였다. 이때 각 데이터들은 수치형 데이터로 각 분포 및 단위의 차이가 존재하여 계수 추정이 어렵고 오차가 크게 나타난다는 문제점이 존재하여 데이터 정규화 과정을 통해 문제를 해결하였다. 또한 전체 데이터에서 산림화재가 발생한 데이터 비율을 약 5%로 조정하였으며 데이터의 불균형 문제가 발생할 수 있어 오버샘플링과 복합샘플링 기법을 적용하였다.

5.2 예측모형의 성능평가

본 연구에서는 산림화재의 확산요인인 기상요인과 발화요인인 인적요인을 고려한 산림화재 예측모형을 개발하고 성능평가를 수행하였다. 예측모형의 성능평가를 단순히 정확도를 기준으로 평가하는 것은 한계가 있다. 이를 보완하기 위해서 이진 분류의 예측성능 측정에서 중요한 평가지표로 활용되고 정확도(Accuracy)의 단점을 보완하기 위해 활용되는 ROC(Receiver operation characteristic) 곡선과 AUC(Area under curve) 값을 활용하였다(Kim et al., 2019). ROC 곡선은 위양성율(False positive rate, FPR)이 변동될 때 재현율(True positive rate, TPR)이 어떻게 변동되는지를 나타내는 곡선이다. 위양성율은 실제로 음성인 값을 양성으로 틀리게 예측한 비율이고, 재현율은 실제 양성인 값을 양성으로 올바르게 예측한 비율이다(Cook, 2008). 본 연구에서 양성 값은 산림화재 발생을 의미하며 음성 값은 산림화재가 발생하지 않음을 의미한다. ROC 곡선은 명확한 수치가 존재하지 않아 성능을 평가하기 위한 지표로 곡선 아래의 면적 값인 AUC 값을 활용하는데, AUC 값의 범위는 0.5부터 1까지 값이다. 0.5는 완전히 무작위한 예측을 의미하며 1은 완벽한 예측을 의미한다. 0.5에서 0.7 사이는 낮은 성능, 0.7에서 0.9 사이의 값은 중간 정도의 성능을 의미하고, 0.9보다 큰 값은 우수한 모델 성능을 나타낸다(McCune and Grace, 2002).

<Table 2>는 개발한 여러 예측모형의 성능 비교를 분석하기 위한 AUC 값으로 각 샘플링 기법을 적용여부를 구분하여 표기하였다. 먼저, AUC 값을 살펴보면 XG Boost와 RF 모형이 샘플링기법 적용에 관계없이 다른 모형들에 비

해 높은 예측 성능을 보이는 것으로 나타났다. 특히 샘플링기법을 적용하지 않은 XG Boost와 RF 모형의 AUC 값은 0.9 이상으로 성능이 우수한 것으로 분석되었다. 또한 MLP 모형을 제외한 다른 모형들은 샘플링기법을 적용하지 않은 것이 적용한 것보다 AUC 값이 높은 것을 확인할 수 있었다. 이는 산림화재 발생 비율이 약 5%가 적절했던 것으로 판단된다. 또한, 예측모형으로 가장 많이 활용되는 로지스틱 회귀분석에서 가장 높은 AUC가 0.672로 다른 모형들에 비해 낮은 성능을 나타냈는데, 이는 산림화재 예측은 로지스틱 회귀분석 보다 머신러닝 기법들을 활용하는 것이 더욱 효과적일 것으로 판단된다.

XG Boost와 RF 같은 머신러닝 모형은 예측 성능이 우수하지만 예측에 대한 설명이 어렵다는 단점이 존재한다. 따라서 인적요인 선별과 관련된 운영시사점을 도출하기 위해서 산림화재 예측에 어떤 변수가 주요한 영향을 보이는지 확인하였다. 이를 위해 모형 성능평가에서 가장 우수한 성능을 보인 XG Boost에 대해서 순열특성중요도 분석(Permutation feature importance analysis)을 시행하여 요인들의 상대적 중요도를 측정하였다. 여기서, 순열특성중요도 분석은 개별 특성(요인) 중에서 측정을 위한 특성에 무작위로 노이즈를 발생시켜 해당 특성(요인)이 기능하지 못하게 만든다. 그 이후에 예측 오차의 변화를 산정하여 요인의 상대적 중요도를 평가하는 분석기법이다(Breiman, 2001; Fisher et al., 2019). 이때 특성(요인)의 중요도(Weight) 값이 양수(+)이면 모형의 예측 성능에 영향을 주는 것을 의미하며, 영(0) 또는 음수(-)이면 예측에 중요도가 낮은 요인으로 판단할 수 있다.

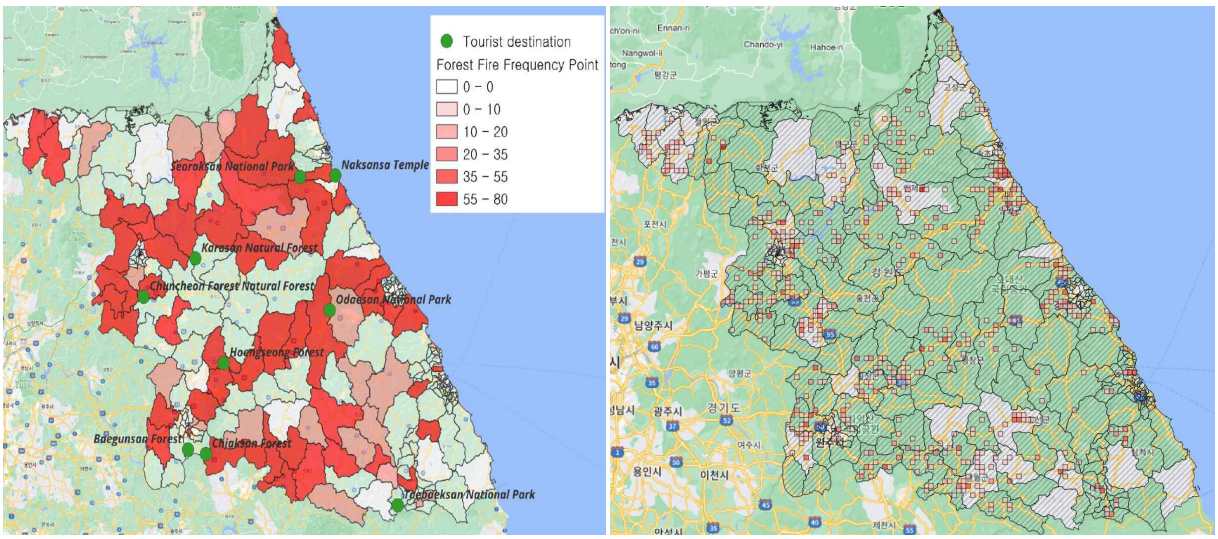
<Table 3>은 본 연구에서 제안한 요인들에 대해 중요도를 크기순으로 정렬한 것이다. 요인별 중요도 결과값을 살펴보면 강수, 풍속, 습도, 기온, 농업인, 관광객, 관광지 인접거리 순으로 기상요인이 인적요인 대비 상대적으로 중요한 요인으로 나타났다. 이러한 결과는 산림화재 및 기상요인 데이터의 경우 시간 단위로 데이터를 책정한 반면, 인적요인 데이터는 월 단위 또는 연 단위로 책정한 결과값이 영향을 미친 것으로

Table 3 Results of Permutation Feature Importance Analysis in XG Boost

Feature	Weight
Precipitation	0.1157 ± 0.0388
Wind velocity	0.0699 ± 0.0220
Humidity	0.0694 ± 0.0264
Temperature	0.0375 ± 0.0093
Agricultural population	0.0332 ± 0.0133
Tourist	0.0113 ± 0.0089
Distance from tourist attraction	0.0098 ± 0.0157

보인다. 순열특성중요도는 개별 요인에 노이즈를 주어 측정한 예측 오차를 활용하는 방식이므로, 시간별로 변화하는 기상요인 대비 변동성이 작은 인적요인들은 설명력의 한계가 있을 것으로 판단된다. 특히 요인 중요도의 전반적인 크기를 볼 때 관광지 인접거리는 다른 요인들 대비 중요도 값이 작은 것으로 나타났는데, 이는 관광지 인접거리는 다른 요인 대비 시간의 흐름에 영향을 받지 않아 데이터의 변동성이 작기 때문으로 파악된다. 하지만, 본 연구에서 제안한 모든 요인들의 중요도 값이 양수임을 확인할 수 있었고, 이는 제안한 인적요인들이 산림화재 예측에 주요한 요인임을 의미한다. 따라서 강원도 지역의 산림화재 예측을 위해서는 거주 중인 농업인 수, 관광인 수, 관광지 거리 등을 고려해야 할 것으로 판단된다. Fig. 1은 강원도 지역에서 발생한 산림화재에 대한 빈도를 중심으로 색으로 표현하였으며, 주요 도로는 노란색 선, 주요 거주지는 회색으로 표현되었다. Fig. 1.(a)에서는 산림화재 빈도가 상대적으로 높은 지역을 구분하였으며(진한 적색일수록 빈도가 높음), 관광지는 녹색 원으로 표기하였다. Fig. 1.(b)는 산림화재가 발생한 좌표(네모 점)를 표기하였다. Fig. 1에서 볼 수 있듯이, 산림화재 빈도가 높은 지역은 대부분 거주 밀집도가 높은 지역 주변이거나 도로 혹은 관광지 인접한 지역이었다.

이러한 결과는 산림화재의 예방 업무를 수행하는 담당자가 인적요인을 통해 감시 지역 선정에 활용할 수 있다는 것을 의미한다. 담당자는



(a) Forest Fire Frequency by Area (b) Forest Fire Frequency by Point

Fig. 1 Forest Fire Occurrence in Gangwon-do

도심지, 관광지 인접지역과 주요 도로를 중점으로 감시와 조기진압을 위한 효율적 인력 배치에 의사결정을 지원할 수 있을 것으로 판단된다. 또한 본 연구에서는 이러한 산림화재가 주로 상기한 지역을 중심으로 발생하였으나, 해당 지역 외에도 산림화재가 발생함을 확인할 수 있었다. 이러한 부분들은 도심지, 관광지 인접지역, 주요 도로와 같이 변동성이 작은 요소 외에도 기상요인과 관광객 등 변동성이 큰 요소들이 많은 영향을 미친 것으로 파악된다. 따라서 담당자는 본 연구에서 제안한 예측모형과 앞에서 제시한 요인 외에도 일별 혹은 시간대별 관광객 및 기상요인 등을 종합적으로 고려해야 한다. 즉, 직접적인 발화원인인 관광객이 많이 방문하는 관광지를 대상으로 기상요인에 인해 확산 가능성이 높은 지역들을 선제적으로 선별하고 해당 지역에 인력을 배치가 필요할 것으로 판단된다. 따라서 화재예방 업무 담당자는 기상요인 외에도 관광지별 예상되는 일별 및 시간대별 관광객 수를 파악해야하고, 농업인의 소각 작업 시간 등을 효율적으로 확인하는 방안을 고안할 필요가 있을 것으로 사료된다.

6. 결 론

본 연구는 강원도 지역의 산림화재 데이터를 활용하여 산림화재에 영향을 미치는 인적요인을 제안하고 머신러닝 기법을 활용하여 산림화재 예측모형을 개발하였다. 모형 개발을 위해 2003년-2022년 강원도 산림화재 발생 원인을 분석하였고 인적요인으로 농업인, 관광객, 관광지 인접거리를 선정하였다. 또한, 산림화재 예측과 관련한 연구에서 활용하는 강수, 풍속, 습도, 기온을 활용하여 로지스틱 회귀분석과 4개의 머신러닝 모형을 활용하여 산림화재를 예측하고 AUC 값을 통해 성능평가를 실시하였다. 이때 산림화재 발생 여부의 불균형 문제를 해소하기 위해 산림화재의 발생 비율을 약 5%로 조정하고 데이터 샘플링 기법들을 적용하였다. 성능평가 결과에서 샘플링 기법과 관계없이 XG Boost와 RF가 우수한 성능을 보였고, MLP 모형은 다른 모형과 달리 샘플링 기법을 적용하면 성능이 개선되었다.

산림화재에 영향을 미치는 주요 요인들을 파악하기 위해 순열특성중요도 분석을 시행하여 요인별로 영향 값을 도출하였다. 영향 값을 살펴보면 기상요인이 인적요인보다 높은 영향 값을 가지지만, 이는 데이터별 책정 기준이 산림

화재 및 기상요인은 시간별, 인적요인은 월별 혹은 연별로 인해 데이터의 변동성의 차이가 영향을 주었기 때문으로 판단된다. 또한, 제안한 모든 요인들이 모두 유의미한 것을 확인할 수 있었다. 이를 통해, 강원도 지역의 산림화재 예측 과정에서 제안한 인적요인들이 통상적으로 활용되는 기상요인과 함께 주요한 요인으로 고려해야 될 것으로 사료된다. 또한 산림화재 예측모형 개발을 위해 머신러닝 기법을 활용할 경우 화재 발생 여부의 비율이 데이터 학습에 있어 주요한 기준이 될 수 있을 것으로 판단된다.

본 연구의 의의는 다음과 같다. 첫째, 강원도 산림화재 예측에 있어 일반적으로 활용되는 기상요인과 더불어 직접적인 발화 원인인 인적요인을 고려하였다는 측면에서 학문적 의의가 있다. 특히 인적요인은 산림화재 예측에 있어 주요한 요인이지만 기준이 명확하지 않고 지역의 특성을 고려해야 하는 어려움이 있었다. 본 연구에서 제안한 인적요인들은 강원도 및 유사 지역의 산림화재 예측 연구에 주요한 기준이 될 것으로 사료된다. 또한 강원도 지역의 산림화재 예방 담당자는 제시한 인적요인들을 활용하여 정밀한 산림화재 예측을 하고 선행적으로 주요 감시 지역을 결정할 수 있다는 점에서 실무적인 측면에서 의의가 있다.

둘째, 산림화재 발생 여부에 따른 데이터 불균형 문제를 해결하기 위해 산림화재 발생과 미발생의 비율 조정을 실시하였으며 데이터 샘플링 기법을 활용했다는 점에서 학술적 의의가 있다. 더불어, MLP 모형을 제외한 모든 모형에서 비율을 조정한 후 AUC 값이 샘플링 기법을 활용한 AUC 값보다 우수하므로 앞에서 제시한 비율이 산림화재를 예측하는데 기준점이 된다는 측면에서 실무적으로 의의가 있다고 할 수 있다.

마지막으로 이진 분류와 예측에서 주로 활용하는 로지스틱 회귀분석 이외에도 최근 주목 받고 있는 다양한 머신러닝 기법들을 적용하였다는 점에서 학술적 의의가 있다. 또한 개발한 모형이 높은 예측 성능을 보여 담당자가 해당 모형을 활용 가능하다는 점에서 실무적인 측면에서 의의가 있다.

향후 연구에서는 정밀한 예측과 중요도 평가를 위해 일별 혹은 시간별 단위로 책정할 수 있는 요인들을 선별할 필요가 있다. 본 연구에서 제시한 인적요인들은 월 혹은 연단위로 측정되어 시간대별로 발생하는 산림화재 발생 데이터를 측정하는데 한계점이 있었다. 실제 산림화재에 있어 기존에 제시한 인적요인 이외에도 관광객과 유동인구를 추정할 수 있는 공휴일 여부, 시간대별 교통량 등을 추가적인 인적요인으로 선별하여 활용하면 더 개선된 예측모형을 개발할 수 있을 것으로 판단된다.

References

An, S. H., Lee, S. Y., Won, M. S., Lee, M. B. and Shin, Y. C. (2004). Developing the Forest Fire Occurrence Probability Model Using GIS and Mapping Forest Fire Risks, *Journal of the Korean Association of Geographic Information Studies,* 7(4), 57-64.

Arndt, N., Vacik, H., Koch, V., Arpaci, A. and Gossow, H. (2013). Modeling Human-caused Forest Fire Ignition for Assessing Forest Fire Danger in Austria, *iForest-Biogeosciences and Forestry*, 6(6), 315. https://doi.org/10.3832/ifor0936-006.

Atkinson, P. M. and Massari, R. (1998). Generalized Linear Modeling of Susceptibility to Landsliding in the Central Apennines, Italy, *Computer & Geosciences,* 24(4), 373-385.

Batista, G. E., Prati, R. C. and Monard, M. C. (2004). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data, *ACM SIGKDD Explorations Newsletter*, 6(1), 20-29. https://doi.org/10.1145/1007730.1007735.

Biau, G. and Scornet, E. (2016). A Random Forest Guided Tour, *Test*, 25(2), 197-227. https://doi.org/10.1007/s11749-016-0481-7.

Biau, G., Devroye, L. and Lugosi, G. (2008).

Consistency of Random Forest and Other Averaging Classifiers, *Journal of Machine Learning Research*, 9, 2015-2033.

Bishop, C. M. (2006). *Pattern Recognition and Machine learning*, Berlin, Springer.

Breiman, L. (2001). Random Forests, *Machine learning*, 45, 5-32.

Breiman, L. (2002). *Manual on Setting up, Using, and Understanding Random Forests,* California, Berkeley: Statistics Department University of California Berkeley.

Calef, M. P., McGuire, A. D. and Chapin III, F. S. (2008). Human Influences on Wildfire in Alaska from 1988 through 2005: An Analysis of the Spatial Patterns of Human Impacts, *Earth Interactions*, 12(1), 1-17. https://doi.org/10.1175/2007EI220.1.

Chae, H. M., Um, G. J. and Lee, S. Y. (2011). The Vulnerability Assessment of Forest Fire in Gangwon Province Using CCGIS, *Korean Society of Hazard Mitigation*, 11(4), 123-30.

Chae, J. S., Kim B. K., Lee, J. H. and Lee, S. Y. (2019). A Study on Mitigation of Facilities Damage Caused by Forest Fire, *Journal of Wellness*, 14, 39-51. http://dx.doi.org/10.21097/ksw.2019.08.14.3.39.

Chae, K. J., Lee, Y. L., Cho, Y. J. and Park, J. H. (2018). Development of a Gangwon Province Forest Fire Prediction Model Using Machine Learning and Sampling, *The Korea Journal of BigData*, 3(2), 71-78. https://doi.org/10.36498/kbigdt.2018.3.2.71.

Chatterjee, S. and Hadi, A. S. (2015). *Regression Analysis by Example*, New Jersey, John Wiley & Sons.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique, *Journal of artificial intelligence research*, 16, 321-357. https://doi.org/10.1 613/jair.953.

Chen, T. and Guestrin, C. (2016). Xgboost: A Scalable Tree Boosting System, *Proceeding of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, Aug. 13-17, San Francisco California, USA.

Choi, S. and Rho, J. H. (2015). Development and Implementation of a 2-Phase Calibration Method for Gravity Model Considering Accessibility, *Journal of Korean Society of Transportation*, 33(4), 393-404. https://doi.org/10.7470/jkst.2015.33.4.393.

Cook, N. R. (2008). Statistical Evaluation of Prognostic versus Diagnostic Models: Beyond the ROC Curve, *Clinical chemistry*, 54(1), 17-23. https://doi.org/10.1373/clinchem.2007.096529.

Cortes, C. and Vapnik, V. (1995). Support-vector Networks, *Machine learning*, 20(3), 273-297.

Dicky, J. W. (2018). *Metropolitan Transportation Planning*, London, Routledge.

Du, M., Yu, Z., Wang, T., Wang, X. and Jiang, X. (2020). XGBoost Based Strategic Consumers Classification Model on E-commerce Platform, *Proceeding of the 6th International Conference on E-Business and Applications*, Feb. 25-27, Kuala Lumpur, Malaysia.

Fisher, A., Rudin, C. and Dominici, F. (2019). All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously, *Journal of Machine Learning Research*, 20(177), 1-81.

Gudmundsson, L., Rego, F. C., Rocha, M. and Seneviratne, S. I. (2014). Predicting above Normal Wildfire Activity in Southern Europe as a Function of Meteorological Drought, *Environmental Research Letters*, 9(8), 084008. https://doi.org/10.1088/1748-9326/9/8/084008.

Hah, D. W., Kim, Y. M. and Ahn, J. J. (2019). A Study on KOSPI 200 Direction Forecasting Using XGBoost Model, *The*

Korean Data & Information Science Society, 30(3), 655-669. https://doi.org/10.7465/jkdi.2019.30.3.655.

He, H., Bai, Y., Garcia, E. A. and Li, S. (2008). ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning, *In 2008 IEEE International Joint Conference on Neural Networks,* Jun. 1-8, Hong Kong, China, pp. 1322-1328.

Ij, H. (2018). Statistics versus Machine Learning, *Nature Methods*, 15(4), 233. https://doi.org/10.1038/nmeth.4642.

Ivan, T. (1976). Two Modifications of CNN, *IEEE transactions on Systems, Man and Communications, SMC*, 6, 769-772. https://doi.org/10.1109/TSMC.1976.4309452.

Kang, S. A., Kim, S. H. and Ryu, M. H. (2022), Analysis of Hypertension Risk Factors by Life Cycle Based on Machine Learning. *Journal of the Korea Industrial Information Systems Research*, 27(5), 73-82. http://dx.doi.org/10.9723/jksiis.202 2.27.5.073.

Kim, Y. H., Kong, I. H., Chung, C. Y., Shin, I. Cheong S., Jung. W. C., Mo, H. S., Kim, S. I. and Lee, Y. W. (2019). Wildfire Risk Index Using NWP and Satellite Data: Its Development and Application to 2019 Kangwon Wildfires, *Korean Journal of Remote Sensing*, 35(2), 337-342. https://doi..org/10.7780/kjrs.2019.35.2.12.

Kim, K. M., Jang, H. Y. and Zhang, B. T. (2014). Oversampling-based Ensemble Learning Methods for Imbalanced Data, *KIISE Transactions on Computiong Practices*, 20(1), 549-554.

Korea Forest Service. (2017). *The 6th Basic Forest Policies 2018-2037*, Daejeon, Korea Forest Service.

Korea Forest Service. (2020). *Comprehensive Measurements to Prevent Forest Fires Nationwide in 2020*, Daejeon, Korea Forest Service.

Korea Forest Service. (2022). *Forest fire occurrence status*, Daejeon, Korea Forest Service.

Kwak, H. B., Lee, W. K., Lee, S. Y., Won, M. S., Koo, K. S., Lee, B. D. and Lee, M. B. (2010). Cause-specific Spatial Point Pattern Analysis of Forest Fire in Korea, *Journal of Korean Forest Society*, 99, 259-266.

Lee, D., Byun, K., Lee, H. and Shin, S. (2023). The Prediction of Survival of Breast Cancer Patients Based onMachine Learning Using Health Insurance Claim Data, *Journal of the Korea Industrial Information Systems Research*, 28(2), 1-9. http://dx.doi.org/10.9723/jksiis.2023.28.2.001.

Lee, S. Y., Han, S. Y., Won, M. S., An, S. H. and Lee, M. B. (2004). Developing of Forest Fire Occurrence Probability Model by Using the Meteorological Characteristics in Korea, *Korean Journal of Agricultural and Forest Meteorology*, 6(4), 242-249.

Lee, W. C., Kim, Y. S., Kim, J. M. and Lee, C. K. (2020). Forecasting of Iron Ore Prices Using Machine Learning, *Journal of the Korea Industrial Information Systems Research*, 25(2), 57-72. http://dx.doi.org./10.9723/jksiis.2020.25.2.057.

Lee, Y. S. and Sang, H. S. (2012). Problems and Improvement Measures for Extinguishing Wildfires, *Journal of International Studies*, 18, 97-132.

Levy, J. J. and O'Malley, A. J. (2020). Don't Dismiss Logistic Regression: the Case for Sensible Extraction of Interactions in the Era of Machine Learning, *BMC Medical Research Methodology*, 20(1), 1-15. https://doi.org/10.1186/s12874-020-01046-3.

Liang, Y., Wu, J., Wang, W., Cao, Y., Zhong, B., Chen, Z. and Li, Z. (2019). Product Marketing Prediction Based on XGboost and LightGBM Algorithm, *Proceeding of the 2nd International Conference on Artificial*

Intelligence and Pattern Recognition, Aug. 16-18, Beijing, China.

Martín, Y., Zúñiga-Antón, M. and Rodrigues Mimbrero, M. (2019). Modelling Temporal Variation of Fire-occurrence towards the Dynamic Prediction of Human Wildfire Ignition Danger in Northeast Spain, *Geomatics, Natural Hazards and Risk,* 10(1), 385-411. https://doi.org/10.1080/19475705.2018.1526219.

McCune, B. and Grace, J. (2002). A*nalysis of Ecological Communities*, Ohio, MJM Software Design.

Menard, S. (2002). *Applied Logistic Regression Analysis*, California, Sage.

Mohammed, R., Rawashdeh, J. and Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results, *Proceeding of the 11th International Conference on Information and Communication Systems*, Apr. 7-9, Irbid, Jordan.

Noriega, L. (2005). *Multilayer Perceptron Tutorial. School of Computing*, Stoke-on-Trent, Staffordshire University.

Pham, B. T., Jaafari, A., Avand, M., Al-Ansari, N., Dinh Du, T., Yen, H. P. H., Phong, T. V., Nguyen, D. H., Le, H. V., Mafi-Gholami, D., Prakash, I., Thuy, H. T. and Tuyen, T. T. (2020). Performance Evaluation of Machine Learning Methods for Forest Fire Modeling and Prediction, *Symmetry*, 12(6), 1022. https://doi.org/10.3390/sym12061022.

Piao, Y., Lee, D., Park, S., Kim, H. G. and Jin, Y. (2022). Forest Fire Susceptibility Assessment Using Google Earth Engine in Gangwon-do, Republic of Korea, *Geomatics, Natural Hazards and Risk*, 13(1), 432-450. https://doi.org/10.1080/19475705.2022.2030808.

Preisler, H. K. and Westerling, A. L. (2007). Statistical Model for Forecasting Monthly Large Wildfire Events in Western United States, *Journal of Applied Meteorology and Climatology*, 46(7), 1020-1030. https://doi.org/10.1175/JAM2513.1.

Rodrigues, M. and de la Riva, J. (2014). An Insight into Machine-learning Algorithms to Model Human-caused Wildfire Occurrence, *Environmental Modelling & Software*, 57, 192-201. https://doi.org/10.1016/j.envsoft.2014.03.003.

Romero-Calcerrada, R., Novillo, C. J., Millington, J. D. and Gomez-Jimenez, I. (2008). GIS Analysis of Spatial Patterns of Human-caused Wildfire Ignition Risk in the SW of Madrid (Central Spain), *Landscape ecology*, 23(3), 341-354. https://doi.org/10.1007/s10980-008-9190-2.

Sadasivuni, R., Cooke, W. H. and Bhushan, S. (2013). Wildfire Risk Prediction in Southeastern Mississippi Using Population Interaction, *Ecological Modelling*, 251, 297-306. https://doi.org/10.1016/j.ecolmodel.2012.12.024.

Siroky, D. S. (2009). Navigating and Random Forest and Related Advances in Algorithmic Modeling, *Statistics Survey*, 3, 147-163. https://doi.org/10.1214/07-SS033.

Tariq, A., Shu, H., Siddiqui, S., Munir, I., Sharifi, A., Li, Q. and Lu, L. (2022). Spatio-temporal Analysis of Forest Fire Events in the Margalla Hills, Islamabad, Pakistan Using Socio-economic and Environmental Variable Data with Machine Learning Methods, *Journal of Forestry Research*, 33(1), 183-194. https://doi.org/10.1007/s11676-021-01354-4.

Vilar, L., Woolford, D. G., Martell, D. L. and Martín, M. P. (2010). A Model for Predicting Human-caused Wildfire Occurrence in the Region of Madrid, Spain, *International Journal of Wildland Fire*, 19(3), 325-337. https://doi.org/10.1071/WF09030.

Wilson, D. L. (1972). Asymptotic Properties of Nearest Neighbor Rules Using Edited Data,

IEEE Transactions on Systems, Man, and Cybernetics, 3, 408-421. https://doi.org/10.1109/TSMC.1972.4309137.

Won, M. S., Jang, K. C. and Yoon, S. H. (2018). Development of Fire Weather Index Model in Inaccessible Areas Using MOD14 Fire Product and 5㎞-resolution Meteorological Data, *Journal of the Korean Association of Geographic Information Studies*, 21(3), 189-204. https://doi.org/10.11108/kagis.2018.21.3.189.

Won, M. S., Koo, K. S. and Lee, M. B. (2006). An Analysis of Forest Fire Occurrence Hazards by Changing Temperature and Humidity of Ten-day Intervals for 30 Years in Spring, *Korean Journal of Agricultural and Forest Meteorology*, 8(4), 250-259.

Won, M. S., Lee, M. B., Lee, W. K. and Yoon, S. H. (2012). Prediction of Forest Fire Danger Rating over the Korean Peninsula with the Digital Forecast Data and Daily Weather Index (DWI) Model, *Korean Journal of Agricultural and Forest Meteorology*, 14(1), 1-10. https://doi.org/10.5532/KJAFM.2012.14.1.001.

Won, M. S., Miah, D., Koo, K. S., Lee, M. B. and Shin, M. Y. (2010). Meteorological Determinants of Forest Fire Occurrence in the Fall, South Korea, *Journal of Korean Society of Forest Science*, 99(2), 163-171.

Won, M. S., Yoon, S. H. and Jang, K. C. (2016). Developing Korean Forest Fire Occurrence Probability Model Reflecting Climate Change in the Spring of 2000s, *Korean Journal of Agricultural and Forest Meteorology*, 18(4), 199-207. https://doi.org/10.5532/KJAFM.2016.18.4.199.

Ye, J., Wu, M., Deng, Z., Xu, S., Zhou, R. and Clarke, K. C. (2017). Modeling the Spatial Patterns of Human Wildfire Ignition in Yunnan Province, China, *Applied Geography*, 89, 150-162. https://doi.org/10.1016/j.apgeog.2017.09.012

Yoo, B. J. (2021). A Study on the Performance Comparison and Approach Strategy by Classification Methods of Imbalanced Data, *Journal of The Korean Data Analysis Society*, 23(1), 195-207. https://doi.org/10.37727/jkdas.2021.23.1.195.

장 진 명 (Jin-Myeong Jang)

- 인하대학교 아태물류학부 물류학사
- 인하대학교 물류전문대학원 물류학 학술석사
- (현재) 인하대학교 물류전문대학원 물류학박사과정 재학

• 관심분야: 물류 최적화, 머신러닝, 재고관리, 운송 네트워크

김 주 찬 (Joo-Chan Kim)

- 가천대학교 글로벌경제학과 경제학사
- 인하대학교 물류전문대학원 물류학 학술석사
- (현재) CJ대한통운 데이터솔루션그룹 선임연구원

• 관심분야: 네트워크 최적화, 배차 최적화, 머신러닝, 선형계획법

김 화 중 (Hwa-Joong Kim)

- 한국과학기술원 산업공학 석사
- 로잔연방공과대학 산업공학 박사
- (현재) 인하대학교 아태물류학부 교수

• 관심분야: 기업물류, 해상운송, 공급사슬계획

김 광 태 (Kwang-Tae Kim)

- 인하대학교 물류전문대학원 물류학 학술석사
- 고려대학교 산업경영공학부 박사수료
- (현재) LG CNS D&A사업부 Analytics&Optimization 컨설팅팀 총괄컨설턴트

• 관심분야: SCM 및 물류 시스템, 최적화, 머신러닝 등