The Korean Journal of Applied Statistics
2022, Vol. 35, No. 4, 543–552

DOI: https://doi.org/10.5351/KJAS.2022.35.4.543
Print ISSN1225-066X / Online ISSN 2383-5818

A study on frost prediction model using machine learning

Hyojeoung Kim[a], Sahm Kim[1,a]

[a]Department of Applied Statistics, University of Chung-Ang

Abstract

When frost occurs, crops are directly damaged. When crops come into contact with low temperatures, tissues freeze, which hardens and destroys the cell membranes or chloroplasts, or dry cells to death. In July 2020, a sudden sub-zero weather and frost hit the Minas Gerais state of Brazil, the world's largest coffee producer, damaging about 30% of local coffee trees. As a result, coffee prices have risen significantly due to the damage, and farmers with severe damage can produce coffee only after three years for crops to recover, which is expected to cause long-term damage.

In this paper, we tried to predict frost using frost generation data and weather observation data provided by the Korea Meteorological Administration to prevent severe frost. A model was constructed by reflecting weather factors such as wind speed, temperature, humidity, precipitation, and cloudiness. Using XGB(eXtreme Gradient Boosting), SVM(Support Vector Machine), Random Forest, and MLP(Multi Layer perceptron) models, various hyper parameters were applied as training data to select the best model for each model. Finally, the results were evaluated as accuracy(acc) and CSI(Critical Success Index) in test data.

XGB was the best model compared to other models with 90.4% ac and 64.4% CSI, followed by SVM with 89.7% ac and 61.2% CSI. Random Forest and MLP showed similar performance with about 89% ac and about 60% CSI.

Keywords: frost prediction, machine learning, XGB, SVM, Random Forest, MLP

1. 서론

최근 지구 평균기온이 상승하면서 작물의 개화 시기가 빨라지는 반면, 기온 변동성이 커지면서 갑작스런 한파가 발생하면서 농작물 피해가 발생하고 있다. 농작물이 저온에 접하면 조직이 동결되는데, 이로 인하여 세포막이나 엽록체의 막이 경화되어 파괴되거나, 세포가 말라 죽는다. 농작물의 직접적인 큰 피해를 야기하는 서리는 습한 공기가 물의 3중점과 공기 이슬점보다 낮은 차가운 표면에 노출될때 발생한다 (Lee 등, 2002). 이러한 서리로 인한 피해는 최근 세계 곳곳에 발생하며 서리 예측에 대한 관심이 증가하고 있다.

2021년 7월 세계 최대 커피 생산국인 브라질의 미나스 제라이스 주 일대에 갑작스러운 영하의 날씨와 서리가 덮쳐 지역 커피나무의 약 30%가 피해를 보았다. 커피 피해로 인한 커피 가격이 6개월만에 파운드당 1.26달러에서 1.94달러로 상승했다. 그뿐 아니라 피해가 심한 농장은 작황이 회복되기까지 3년은 걸려 2024년에나 커피를 생산할 수 있어 장기간 막대한 피해가 예상된다. 또한 프랑스는 21년 4월에 이어 22년 프랑스에서 역대 가장 추운 4월을 기록하며 2년 연속 봄서리로 피해를 겪고있다. 21년 엄청난 서리로 인해 전체 피해금액은 20억 달러(약 2조 2,746억 원)에 달하며, 포도밭은 약 80%로 전년대비 와인 생산량이 27% 감소하였다. 22년은

This research was supported by the Graduate Fellowship Scholarship in 2018.
[1] Corresponding author: Department of Statistics, Chungang University, 84 Heuksukro, Dongjak-Gu, Seoul 06974, Korea. E-mail: sahm@cau.ac.kr

Published 31 August 2022 / journal homepage: http://kjas.or.kr

프랑스 북부 지방에서는 영하 9도까지 내려가며 한겨울 맹추위를 재현하며 포도뿐 아니라 복숭아, 살구 등 다른 과수에도 큰 영향을 미칠것으로 예상된다. 이렇듯 최근 지구온난화로 계절에 맞지 않는 따뜻한 날씨가 이어지며 식물의 성장을 가속했는데, 기온 변동성이 커지면서 갑작스런 한파가 발생하면서 서리 발생 예측이 점점 어려워지고 있다. 이로 인해 시기에 맞지 않게 빨리 자란 식물이 한파에 더욱 취약한 상태가 되어 수십만 헥타르에 달하는 피해가 이어지고 있다. 이에 따라, 서리 발생을 정확히 예측하기 위한 시도들이 계속되고 있다.

서리 발생 예측을 위하여 Sallis 등 (2008)은 SOM(Self-organizing Map)을 사용하여 기온, 습도, 풍속, 강수, 기압 변수들의 데이터 간의 종속성(상관 관계)를 연구하였다. 연구를 통해 서리가 내릴 때의 풍향은 주로 남동쪽, 즉 칠레 오히긴스 지역의 계곡에 있는 로스 안데스산맥의 영향을 받기 때문에 지리 공간적 요인들에 대한 분석이 필요함을 확인하였다. Lee 등 (2016)은 한반도 6개 관측소 데이터로 로직스틱 회귀분석보다 결정트리기법이 POD(Probability of Detection)가 높음에 따라 결정트리 기법을 서리 예측 모델로 선택했다. Casta ??neda-Miranda 등 (2017)은 ARX(Autoregressive models with external input)와 MLP(Multi-Layer Perceptron) 모형으로 외부 공기 온도, 외부 공기 상대 습도, 풍속, 지구 일사량 플럭스, 내부 공기 상대 습도변수를 사용해 온실 내부 온도를 예측했다. MLP모형이 R^2가 여름, 겨울 각각 0.9549, 0.9590으로 더 높은 성능을 보임을 밝혔다. Zendehboudi 등 (2017)은 수평 및 평면에서 서리 두께와 밀도를 예측하였다. Hybrid ANFIS(Adaptive Neuro-Fuzzy Inference System), GA-LSSVM(Least Square Support Vector Machine with Genetic Algorithm), GA-RBF(Radial Basis Function neural network based on genetic algorithm), MLP모델을 비교하였으며, 4가지 경우 모두 MLP모형에서 R^2가 약 0.99를 가지며 가장 우수한 성능을 보임을 보였다. Diedrichs 등 (2018)은 SMOTE(Synthetic Minority Over-sampling Technique)를 적용하면 Random Forest 및 로지스틱 회귀 모델 모두에서 재현율을 높임으로써 향상된 성능을 보여주었다. Ding 등 (2019)은 온도, 습도 및 방사선의 과거값을 취하여 SVM 모형으로 서리 발생 가능성을 예측했다. 기온은 서리 예측 모델에 관여하는 핵심 요소며, 습도는 2시간 또는 3시간 이내와 같이 상대적으로 긴 시간 동안 조기 경보를 발생시키는 데 도움이 되며, 방사선은 단기간에 일부 지역에 대한 변화를 반영하여 민감도를 향상시킴을 보여줬다. 최종적으로 SVM 모형으로 다음 1시간, 2시간 및 3시간 서리발생을 예측할 때, 재현율은 각각 100%, 99.3%, 99.8%로 달성함을 보였다. Tamura 등 (2020)은 온도와 증기압 변수의 과거값으로 단순 이동 평균과 지수 이동 평균을 사용하여 SVM(Support Vector Machine) 결과를 비교했다. 지수 이동 평균을 사용한 모델은 F1 측정 측면에서 단순 이동 평균을 사용한 모델보다 몇 퍼센트 더 성능이 우수함을 밝혔다. Rozante 등 (2020)은 지역 일기 예보 모델에 의해 수치적으로 계산된 변수의 가중치를 보정하여 서리 지수(IG)를 교정했다. 이때 가중치는 온도가 가장 큰 기여도를 가지며 압력과 바람이 그 뒤를 이으며, 다른 변수는 가중치의 합이 1이라는 제한을 따르도록 조정했다. Wassan 등 (2021)은 CNN 모델로 서리 현상 예측했다. 1차원 데이터 분석을 위해 conv1d를 사용했으며, 반복횟수가 30,000회일 때 정확도가 97.6%, 50,000회일 때 정확도가 98.6%임을 보였다.

위에서 언급한 모형들과 같이 최근 연구에서는 서리발생 예측 정확도를 향상시키기 위해 다양한 방법이 적용되고 있다. 본 논문에서는 위의 논문에서 자주 사용되었던 SVM, Random Forest, MLP 모형과 최근 다양한 분야에 자주 사용되는 XGB 모형의 결과를 비교해보려 한다. 사용한 데이터는 2000년 1월 1일부터 2021년 4월 30일까지 대한민국 53개 지점의 서리발생이력데이터와 지상관측자료를 활용하였다. 하이퍼파라미터(Hyper parameter) 선정을 위해 정확도(accuracy)를, ROC(Receiver operating characteristic)를 참조하여 모델을 선정하였고, 최종 모델 성능은 재현율(recall), F-score, 임계성공지수(CSI) 등의 다양한 모델 평가기준으로 종합적으로 모델을 평가하였다.

다음 2장에서는 서리 발생 예측을 위해 사용한 SVM, Random Forest, MLP, XGB 모형에 대하여 소개한다. 3장에서는 본 연구에 활용된 서리발생이력 데이터, 기상 데이터와 전처리 방법에 대하여 설명하고, 4장에서는 위에서 언급한 모형을 적용한 예측 결과를 비교, 분석한다. 5장에서는 결론 및 향후 연구 방향에 대하여 제안할 예정이다.

2. 예측 모형

2.1. Support vector Machine (SVM) 모형

Support vector machine (SVM)은 기계 학습의 분야 중 하나로 패턴 인식, 자료 분석을 위한 지도 학습 모델이며, SVM은 오류 데이터에 대한 영향이 없으며, 과적합되는 경우가 적다 (Halil 등, 2019). SVM은 예측하고자 하는 종속변수의 형태에 따라 2개 그룹을 분류하는 기본모형인 support vector classification (SVC)과 support vector regression (SVR) 두 가지로 나누어진다 (Cao 등, 2009). 본 논문의 목적은 서리 발생 여부를 분류하기 위함이므로 SVC를 이용한다. SVM에서는 최적화된 SVM 함수를 찾기 위해 비용함수와 라그랑주 함수, 제약조건을 통해 마진을 크게 하는 것과 에러에 대한 페널티를 작게 하는 것의 균형을 맞춘다. 이때 제약조건으로 C와 gamma 매개변수를 지정하는데, C는 데이터 샘플들이 다른 클래스에 놓이는 것을 허용하는 정도를 결정하고, gamma는 결정 경계의 곡률을 결정한다. 두 값 모두 커질수록 알고리즘의 복잡도는 증가하고, 작아질수록 복잡도는 낮아진다. 본 논문에서는 C = 1, C = 3 두 가지 값을 사용하였고, gamma는 scale 값으로 지정하였다. 또한 SVM 모형은 결정경계인 커널함수를 지정할 수 있는데, 본 논문에서는 기본 선형 커널과 비선형에서 주로 사용되는 radial basis function (RBF) 커널을 사용한다.

2.2. Random Forest 모형

Random Forest (RF) 모형은 검출, 분류 그리고 회귀분석 등에 활용되는 의사 결정 트리 기반의 알고리즘으로, 앙상블 학습(ensemble learning)방법의 일종이다. 훈련 과정에서 부트스트랩(bootstrap)방식을 이용해 개별 의사결정트리(decision tree)를 구축한다. 다수의 의사 결정 트리로부터 나온 결과를 범주형 데이터인 경우는 다수결(voting)로, 연속형 데이터의 경우는 평균으로 최종 값을 결정한다. Random Forest 모형은 설명변수 간의 상호작용 및 비선형성을 다루기 용이하고 이상치에 강해 회귀 문제에서 광범위하게 사용되고 있다. Random Forest의 주요 파라미터는 트리의 갯수와 트리의 허용 깊이로 연구자가 지정해야 한다. 트리의 갯수를 늘리면 연산량이 증가해 속도가 느려지지만 데이터에 대한 과적합을 피할 수 있다. 반면, 트리의 최대 깊이를 늘리면 과적합이 발생한다. 과적합을 해결하기 위해 적정수준의 파라미터를 결정하게 된다. 본 논문에서는 트리의 갯수(n estimator)를 100과 1000, 트리 허용 깊이(max depth)를 3과 7로 총 4가지 경우로 파라미터를 조정하는 과정을 거쳐 예측력이 우수한 모형을 선정하였다.

2.3. Multi-Layer Perceptron 모형

인공신경망(뉴럴네트워크)은 기계학습과 인지과학에서 생물학의 신경망(perceptron)에서 영감을 얻은 통계학적 학습 알고리즘이다. MLP모형은 인공신경망의 기본형으로 여러 개의 퍼셉트론을 연결해 여러 층을 만들고, 이 층들을 중첩시켜 다층(Multi layer)으로 만든다 (Ghielmi 등, 2006). MLP는 형성한 네트워크에서 인공뉴런(노드)이 학습을 통해 시냅스의 결합 세기를 변화시켜가며 복잡한 계산과정을 통해 모형의 정확도를 높인다. MLP는 입력층(input layer), 은닉층(hidden layer), 출력층(output layer)으로 구성되며, 입력층과 출력층 사이에 하나 이상의 중간층이 존재한다. MLP 하이퍼 파라미터로 은닉층 사이즈(hidden layer sizes), 반복횟수(max iteration)를 조정했으며, 활성화 함수는 ‘relu’, 옵티마이저는 ‘adam’으로 지정하였다.

2.4. XGB 모형

Extreme Gradient Boosting (XGB) 모형은 의사결정나무 기반의 알고리즘으로써 그래디언트 부스팅(Gradient Boosting)을 개량한 알고리즘이다 (Zheng 등, 2019). 그래디언트 부스팅은 기울기 하강법을 이용하여 연속되는 트리가 이전 트리의 예측 오차를 기울기 하강법으로 보완하여 순차적으로 모형을 생성해 나가면서 예측력을 높이는 기계학습 기법이다. 제대로 예측되지 못한 개체들에 집중하여 새로운 예측모델을 만드는 반복과

Table 1: Derivative variable listk

Period	Weather factor	Descriptive statistics
For 1 days (From noon the day before to midnight on the day before)	Temperature	Min, Max, difference(Max−Min), Mean, Sum, Standard Deviation
	Humidity	
	Wind speed	
	Cloudiness	
	Precipitation	
For 3 days (From 3 days ago to time of occurence)	Precipitation	Mean, Sum
For 7 days (From 7 days ago to time of occurence)	Precipitation	Mean, Sum
For 10 days (From 10 days ago to time of occurence)	Precipitation	Mean, Sum

정을 거치며 여러 약한 학습기의 조합으로 강한 학습기를 생성한다. XGB 모형은 그래디언트 부스팅 모형을 병렬학습이 지원되도록 분산처리하여 빠르게 변수를 발견하는 기법이다. 메모리에 맞지 않는 데이터 처리를 위한 병렬연산을 지원하여 빠르며, 매우 큰 모델에 사용가능하며, 과적합이 잘 일어나지 않는 장점이 있다. 하지만 파라미터의 개수가 많아 복잡하다는 단점이 있다. 본 논문에서는 그 중 트리 허용 깊이(max depth)와 학습률(learning rate), 2개의 파라미터를 조정하여 예측력이 우수하도록 모형을 적합하였다.

3. 데이터 및 자료 분석

3.1. 데이터 및 분석 방법

본 연구 주제인 서리는 지표 부근 공기의 이슬점 온도가 어는점 이하일 때 발생하므로, 기온의 연속성을 고려하여 매 정오를 기준으로 24시간(기준일 정오부터 다음날 정오까지) 서리발생여부를 예측하였다. 2000년 1월부터 2021년 4월까지 데이터에서 사계절의 영향으로 서리가 발생하기 시작하는 10월부터 4월까지 데이터만 추출했으며, 기준년도 10월부터 다음년도 4월에 서리가 한번도 발생하지 않은 지점의 데이터는 해당 기간 제거했다. 제거하는 과정에서 대한민국 총 53개 지점의 서리발생이력데이터를 수집하였고, 해당 지점의 시간별 기상요소(기온, 풍속, 습도, 운량, 강수량, 일사량)를 수집하여 기준일 정오부터 다음날 정오까지 평균, 총합, 최대, 최소, 분산값을 계산하여 서리발생시점과 시간 간격을 일치시켰다. 추가로 분석의 정확성을 향상시키고, 유의미한 정보를 얻기 위해 원시데이터로부터 여러 변수의 조합 · 조정을 통해 Table 1과 같이 파생변수를 생성하였다. Rajaei 등 (2015)에 따르면 모든 토양 유형의 열전도율은 토양 입도 분포 곡선, 수분 함량 및 건조 밀도에 따라 달라질 수 있기때문에 토지상황을 반영하기 위해 예측일 포함 3일, 7일, 10일동안의 강수량의 평균과 총합을 계산했다. 이때 변수명 뒤에 특정 숫자가 없으면 서리를 예측하는 날의 24시간 날씨 통계값을, 뒤에 숫자 1이 붙으면 전날 24시간의 통계값을, 3이나 7, 10은 예측일을 포함한 3일동안의 통계값을 의미한다.

데이터에서 무작위로 70%를 추출하여 교차검증을 통해 훈련 및 검증자료(training and validation data)로 사용했고, 나머지 30%의 데이터는 테스트데이터(test data)로 모형의 성능을 평가하였다. 훈련 및 테스트 데이터에서 서리 발생과 미발생의 비율은 Table 2와 같으며, 훈련 및 테스트 데이터에서 모두 서리발생과 미발생 비율은 1:3 정도이다.

본 논문에서 서리발생 예측을 위해 통계 프로그래밍 언어인 R과 Python을 모두 사용하였으며, 모형 적합 및 예측을 위해 주로 Python의 sklearn 패키지를 사용하였다.

Table 2: Count of frost and non-frost occurrence in data set

Data set	Ocuurence status	Count
Train & Validation set	Frost (O)	21769
	Frost (X)	71387
Total Train & Validation set		93156
Test set	Frost (O)	9290
	Frost (X)	30635
Total Test set		39925

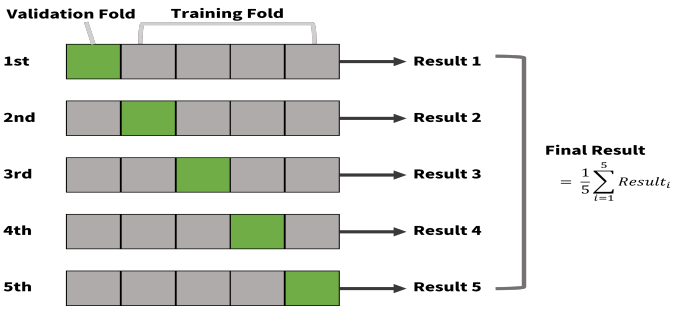

Figure 1: *K-FOLD process.*

Table 3: Confusion matrix

		Predicted	
		Frost (O)	Frost (X)
Actual	Frost (O)	True Positives (TP)	False Negatives (FN)
	Frost (X)	False Positives (FP)	Ture Negatives (TN)

3.2. 모형 적합 결과

훈련 데이터에서 과적합과 과소적합을 방지하도록 아래 Figure 1과 같이 K-Fold 교차검증을 사용해 5개 Fold로 분할하여 정확도(Acc)와 AUC(Area Under the Curve)를 계산했다. 평균 Acc와 AUC를 참조하여 하이퍼 파라미터를 선택했다. 정확도는 예측결과와 실제 관측결과를 Table 3과 같이 혼동행렬로 표현했을때, 아래의 식과 같이 정의 되며, AUC는 모든 임계값에서 분류 모델의 성능을 보여주는 ROC(Receiver Operating Characteristic)그래프 곡선 아래영역이다. 정확도는 전체의 경우에서 얼마나 잘 맞춰는지를 의미하며, AUC는 값이 클수록 서리를 구별하는 모델의 성능이 좋음을 의미한다.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{3.1}$$

모형별 다양한 하이퍼 파라미터로 검증한 결과는 Table 4와 같다. 모형별로 비슷한 수준의 성능을 보였다. 모형별로 살펴보면, SVM 모형은 하이퍼 파라미터 C에 대한 영향보다 커널에 대한 영향을 크게 받는것으로 나타났다. 커널이 'rbf'이고, C가 3인 모형의 정확도와 AUC가 각각 0.89, 0.931로 가장 높았다. 랜덤포레스트 모형은 트리의 갯수보다는 트리 깊이에 대한 영향을 크게 받았다. 트리 깊이가 7인 경우 AUC는 동일하나, 정확도 면에서 트리갯수가 100인 경우가 정확도가 0.892로 조금 더 높아 최종 파라미터로 선정하였다. MLP 모형은 최대 반복횟수보다는 은닉층의 개수에 대한 영향이 큰 것으로 보여진다. 정확도는 반복최대횟수가 1000, 은닉층의 개수가 7일때, 0.895로 가장 높은 수치를 보였다. XGB 모형은 하이퍼 파라미터에 상관없이

Table 4: Model validation results

Model	Parameter		AUC	ACCURACY
SVM	kernel : rbf	C : 1	0.929	0.889
	kernel : rbf	**C : 3**	**0.931**	**0.89**
	kernel : linear	C : 1	0.913	0.873
	kernel : linear	C : 3	0.911	0.872
Random Forest	n estimator : 100	max depth : 3	0.922	0.865
	n estimator : 1000	max depth : 3	0.921	0.866
	n estimator : 100	**max depth : 7**	**0.936**	**0.892**
	n estimatorr : 1000	max depth : 7	0.936	0.891
MLP	max iter : 100	hidden layer : 3	0.909	0.863
	max iter : 1000	hidden layer : 3	0.909	0.863
	max iter : 100	hidden layer : 7	0.936	0.894
	max iter : 1000	**hidden layer : 7**	**0.937**	**0.895**
XGB	max depth : 3	learning rate : 0.1	0.938	0.895
	max depth : 7	**learning rate : 0.1**	**0.944**	**0.901**
	max depth : 3	learning rate : 1	0.938	0.896
	max depth : 7	learning rate : 1	0.924	0.883

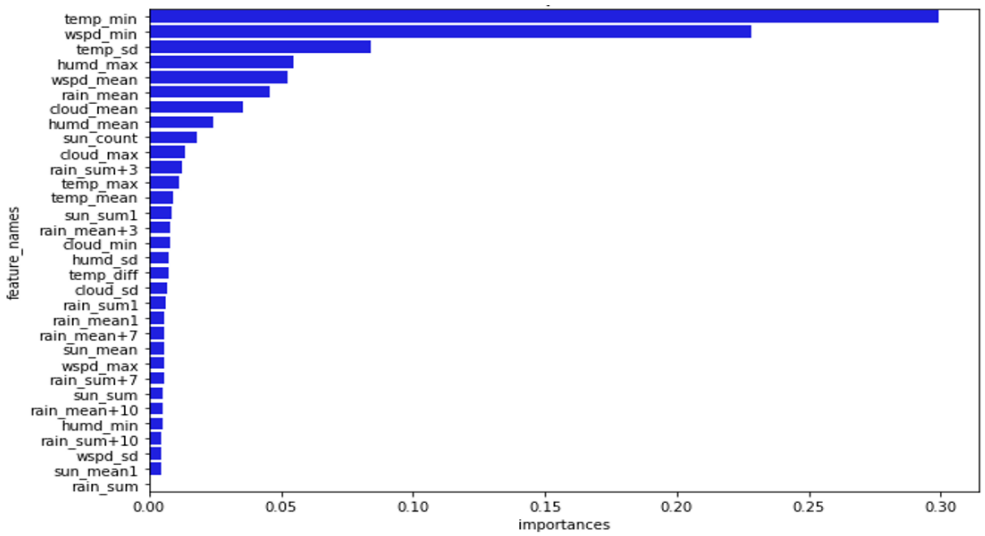

Figure 2: *Variable importance in.*

비슷하게 좋은 성능을 보였다. 하지만 트리 깊이 7, 학습률 0.1로 최적화했을때, 정확도와 AUC가 0.901, 0.944로 모형들보다 가장 높은 수치를 보였다.

가장 좋은 성능을 보였던 XGB 모델로 변수 중요도 우선순위는 다음 Figure 2와 같다. 선형상관관계와 비교해봤을때, 최저기온은 XGB에서도 가장 중요도가 높은 변수임을 보였다. 그 다음으로는 최소풍속, 기온 표준편차, 최대습도, 평균풍속, 평균 강수량, 평균운량, 평균 습도, 총 일사시간, 최대운량, 3일 누적 강수량 이었다. 선형상관관계 분석에서 기온 관련 변수만 상관계수가 높았던 것에 비해, XGB에서는 다양한 기상요소가 중요한 변수로 선정됨을 볼 수 있다.

Table 5: Result of forecasting forest

Model	XGB	SVM	RF	MLP
Accuracy	**0.904**	0.897	0.895	0.894
Precision	0.824	0.832	**0.836**	0.827
Recall	**0.748**	0.698	0.680	0.689
F1	**0.784**	0.759	0.750	0.752
CSI	**0.645**	0.612	0.600	0.602

3.3. 모형 성능 평가

본 논문에서 모형의 성능을 파악하기 위하여 정확도(Accuracy)와 정밀도(Precision), 재현율(Recall), F1, 임계성공지수(CSI)와 같은 다양한 성능평가지표를 사용했다. 정확도는 검증에서 사용한 개념과 같으며, 정밀도, 재현율, F1, 임계성공지수는 다음과 같이 정의된다.

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP}+\text{FP}}, \\ \text{Recall} &= \frac{\text{TP}}{\text{TP}+\text{FN}}, \\ \text{F1} &= \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}, \\ \text{CSI} &= \frac{\text{TP}}{\text{TP}+\text{FP}+\text{FN}}. \end{aligned} \tag{3.2}$$

정밀도는 서리로 예측된 것들 중 실제로도 서리인 경우의 비율이며, 재현율은 실제 서리가 관측되었을 때, 예측이 서리가 발생할 것으로 된 경우의 비율이다. F1 점수는 정밀도와 재현율의 조화평균으로 주로 분류 클래스 간의 데이터가 불균형이 심각할때 주로 사용한다. 임계성공지수는 예측과 관측에서 서리 발생에 관련된 모든 경우의 총합에서 옳은 예측의 비율을 의미하는데, 틀림(False Alarms)과 놓침(Misses) 모든 경우에 민감해서 기상학적 빈도수를 같는 경우에 실질적인 정보를 제공할 수 있다. 정확도, 정밀도, 재현율, F1, 임계성공지수가 1에 가까울수록 모형의 예측 성능이 우수하다.

본 논문에서는 교차검증을 이용해 훈련 데이터로 하이퍼파라미터 값을 선정하였고, 적합한 모형으로 테스트 데이터 값들을 예측하여 모형의 성능을 비교하고자 한다. 모형에 따른 평가지표 비교결과는 다음 Table 5와 같다. SVM, 랜덤포레스트, MLP, XGB 모든 모형의 독립변수는 기상요소(풍향, 풍속, 습도, 기온, 일사) 파생변수로 모두 동일하게 일치시켜 예측하였다. 예측 결과 정확도면에서는 XGB가 0.904로 가장 높은 점수를 가지고 있으나, 모든 모델이 약 0.9 정도로 비슷한 점수를 유지하고 있다. 정밀도는 랜덤포레스트가 0.836으로 제일 높았으며, 재현율에서는 XGB가 타 모형에 비해 상당히 높은 수치를 보여줬다. 하지만 XGB 모형이 정밀도와 재현율의 조화평균인 F1과 CSI에도 가장 높은 점수를 보였다. 또한 테스트 데이터 예측결과를 가지고 계산한 ROC 그래프인 Figure 3을 살펴보면 모든 모델들이 서리를 구별하는데 있어 안정적인 것으로 보여지나, 그 중에서도 XGB의 AUC가 0.948로 가장 높았다. XGB 모형이 정확도, 재현율, F1, CSI, AUC 여러 지표에서 높은 점수를 가지면서 가장 안정적인 모형으로 나타난다.

4. 결론

본 논문에서는 최근 기후 문제로 잦은 피해를 야기하는 서리발생 예측을 하고자 하였다. 서리 발생에 가장 기본적인 기상요소만을 이용하여 다양한 모델(SVM, 랜덤포레스트, MLP, XGB)을 적용하여 예측결과를 비교해 보았다. 예측결과 XGB 모형의 정확도가 0.904, CSI 0.645로 가장 좋은 성능을 보여주었다.

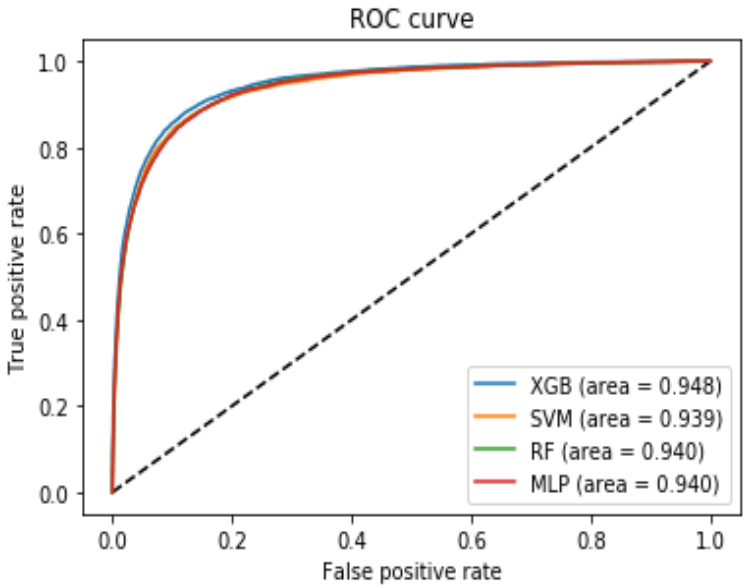

Figure 3: *ROC Curves of models.*

본 연구에서는 오직 기상변수로 모델을 비교하고, 가장 성능이 좋았던 XGB에서 서리발생에 주요한 변수를 확인했다. 추후 해당 연구를 확장하여 서리예측 정확도 향상을 위하여 서리 예측에 중요한 다른 변수들을 모색하는 등 다각적인 접근이 필요할 것으로 예상된다. 또한 전처리 과정에서 서리 발생과 서리 미발생의 빈도가 불균형한것을 고려하여 SMOTE(Synthetic Minority Over-sampling Technique)나 ADASYN(Adaptive Synthetic Sampling) 등을 통해 학습률을 높여 정확도 비교 연구도 미래 서리발생 예측 정확도 향상에 보탬이 될 것으로 사료된다. 더 나아가 모델 운영에 있어서 독립변수들간의 관계를 살펴 주성분분석 등을 진행하거나 중요도가 높은 특정 변수들만을 추출하여 모델을 단순화하는 방법도 서리발생 예측 알고리즘 운영에 유의미한 접근이 될 것으로 예상된다.

References

Cao Z, Han H, Gu B, and Ren N (2009). A novel prediction model of frost growth on cold surface based on support vector machine. *Applied Thermal Engineering*, **29**(11–12), 2320–2326.

Castaneda-Miranda A and Castano VM (2017). Smart frost control in greenhouses by neural networks models. *Computers and Electronics in Agriculture*, **137**, 102–114.

Diedrichs AL, Bromberg F, Dujovne D, Brun-Laguna K, and Watteyne T (2018). Prediction of frost events using machine learning and IoT sensing devices, *IEEE Internet of Things Journal*, **5**(6), 4589–4597.

Ding L, Noborio K, and Shibuya K (2019). Frost forecast using machine learning-from association to causality, *Procedia Computer Science*, **159**, 1001–1010.

Ghielmi L and Eccel E (2006). Descriptive models and artificial neural networks for spring frost prediction in an agricultural mountain area. *Computers and Electronics in Agriculture*, **54**(2), 101–114.

Halil RA??O and Demirci M (2019). Predicting the turkish stock market bist 30 index using deep learning. *International Journal of Engineering Research and Development*, **11**(1), 253–265.

Lee H, Chun JA, Han HH, and Kim S (2016). Prediction of frost occurrences using statistical modeling approaches. *Advances in Meteorology*.

Lee YB and Ro ST (2002). Frost formation on a vertical plate in simultaneously developing flow. *Experimental Thermal and Fluid Science*, **26**(8), 939–945.

Rajaei P and Baladi GY (2015). Frost depth: general prediction model. *Transportation Research Record*, **2510**(1), 74–80.

Rozante JR, Gutierrez ER, da Silva Dias PL, de Almeida Fernandes A, Alvim DS, and Silva VM (2020). Development of an index for frost prediction: Technique and validation. *Meteorological Applications*, **27**(1), e1807.

Sallis P, Jarur M, and Trujillo M (2008, November). Frost prediction characteristics and classification using computational neural networks, *In International Conference on Neural Information Processing*, 1211–1220.

Tamura Y, Ding L, Noborio K, and Shibuya K (2020, December). Frost prediction for vineyard using machine learning. *In 2020 Joint 11th International Conference on Soft Computing and Intelligent Systems and 21st International Symposium on Advanced Intelligent Systems (SCIS-ISIS)*, 1–4. IEEE.

Wassan S, Xi C, Jhanjhi NZ, and Binte-Imran L (2021). Effect of frost on plants, leaves, and forecast of frost events using convolutional neural networks. *International Journal of Distributed Sensor Networks*, **17**(10), 15501477211053777.

Zendehboudi A and Li X (2017). Robust predictive models for estimating frost deposition on horizontal and parallel surfaces, *International Journal of Refrigeration*, **80**, 225–237.

Zheng H and Wu Y (2019). A xgboost model with weather similarity analysis and feature engineering for short-term wind power forecasting. *Applied Sciences*, **9**(15), 3019.

Received April 30, 2022; Revised May 19, 2022; Accepted May 26, 2022

머신러닝을 사용한 서리 예측 연구

김효정[a], 김삼용[1,a]

[a]중앙대학교 응용통계학과

요 약

서리는 표면 근처의 공기의 이슬점 온도가 빙점 이하일 때 수증기가 승화, 응축되어 땅이나 물체에 얼게 되는 작은 얼음 결정체이다. 서리가 내리면 농작물이 직접 피해를 입는다. 농작물이 낮은 온도에 접촉하면 조직이 얼어서 세포막이나 엽록체가 딱딱해지고 파괴되거나 건조한 세포가 죽습니다. 2020년 7월, 세계 최대 커피 생산국인 브라질 미나스제라이스 주에 갑작스러운 영하의 날씨와 서리가 내려 지역 커피 나무의 약 30%가 피해를 입었다. 이로 인해 피해로 커피값이 크게 올랐고, 피해가 심각한 농가는 농작물이 회복되기까지 3년이 걸리기 때문에 2024년에야 커피를 생산할 수 있다. 본 논문에서는 심한 서리가 내리는 것을 방지하기 위해 기상청이 제공하는 서리 발생 데이터와 기상관측 데이터를 이용해 서리를 예측하려고 했다. 관측 지점의 고도 및 풍속, 온도, 습도, 강수량, 흐림 등의 기상 요인을 반영하여 모델을 구축하였다. XGB, SVM, Random Forest, MLP 모델을 사용하여 다양한 하이퍼 파라미터를 학습 데이터로 적용하여 각 모델에 가장 적합한 모델을 선택하였다. 마지막으로, 결과는 테스트 데이터에서 정확도(acc)와 중요 성공 지수(CSI)로 평가되었다. XGB는 90.4%의 acc와 64.4%의 CSI로 다른 모델에 비해 최고의 모델이었고, SVM은 89.7%의 acc와 61.2%의 CSI로 그 뒤를 이었다. 랜덤 포레스트와 MLP는 약 89%의 acc와 약 60%의 CSI로 비슷한 성능을 보였다.

주요용어: 서리 예측, 머신러닝, XGB, SVM, Random Forest, MLP

이 논문은 2018년도 대학원생지원장학금의 지원에 의해 작성되었음.

[1]교신저자: (06974) 서울시 동작구 흑석로 84, 중앙대학교 통계학과. E-mail: sahm@cau.ac.kr