한국빅데이터학회
제3권 제2호, 2018, pp. 71-78

머신러닝과 샘플링을 이용한 강원도 지역 산불발생예측모형 개발

Development of a Gangwon Province Forest Fire Prediction Model using Machine Learning and Sampling

채경재† · 이유리 · 조용주 · 박지현

인하대학교 통계학과 석사과정

요 약

본 연구는 산불 발생 예측 모형의 정확도를 높이기 위해 머신러닝 기법을 적용한 연구이다. 산불 피해 면적이 가장 큰 강원도를 중심으로 2003년부터 2016년까지 총 14년의 산불 자료를 이용하였다. 기상자료의 오차를 줄이기 위해 강원도를 9개의 구역으로 나누어 각 구역 관측소의 기상자료를 이용하였다. 9개의 구역으로 나누어 각 구역의 산불 예측 모형을 만들게 되면 산불이 발생한 날(majority)과 산불이 발생하지 않은 날(minority)의 비율 차이가 큰 불균형 문제가 발생한다. 불균형 문제에서는 모델의 성능이 떨어지는 현상이 발생할 수 있다. 이를 해결하기 위해 여러 샘플링 방법을 적용하였다. 또한 모델의 정확도를 높이기 위해 캐나다 산불 기상 지수(FWI)의 5가지 지수를 파생변수로 사용하였다. 모델링 방법은 통계적 방법인 로지스틱 회귀분석 방법과 머신러닝 방법인 random forest와 xgboost 방법을 사용하였다. 각 구역의 최종모델의 선택기준을 정확도, 민감도, 특이도를 고려하여 정했으며, 9개 구역의 예측 결과는 산불이 발생한 104건 중 80건의 발생 예측에 성공하였으며 산불이 발생하지 않은 9758건 중 7426건의 발생하지 않음을 예측했다. 전체의 정확도는 76.1%였다.

■ 중심어 : 캐나다 산불 기상 지수(FWI), 머신러닝 모델, 샘플링, 불균형 데이터

Abstract

The study is based on machine learning techniques to increase the accuracy of the forest fire predictive model. It used 14 years of data from 2003 to 2016 in Gang-won-do where forest fire were the most frequent. To reduce weather data errors, Gang-won-do was divided into nine areas and weather data from each region was used. However, dividing the forest fire forecast model into nine zones would make a large difference between the date of occurrence and the date of not occurring . Imbalance issues can degrade model performance. To address this, several sampling methods were applied. To increase the accuracy of the model, five indices in the Canadian Frost Fire Weather Index (FWI) were used as derived variable. The modeling method used statistical methods for logistic regression and machine learning methods for random forest and xgboost. The selection criteria for each zone's final model were set in consideration of accuracy, sensitivity and specificity, and the prediction of the nine zones resulted in 80 of the 104 fires that occurred, and 7426 of the 9758 non-fires. Overall accuracy was 76.1%.

■ Keyword : Forest fire Weather Index(FWI), Machine learning model, sampling, imbalanced data

2018년 12월 01일 접수; 2018년 12월 02일 수정본 접수; 2018년 12월 31일 게재 확정.
† 교신저자 22171471@inha.edu

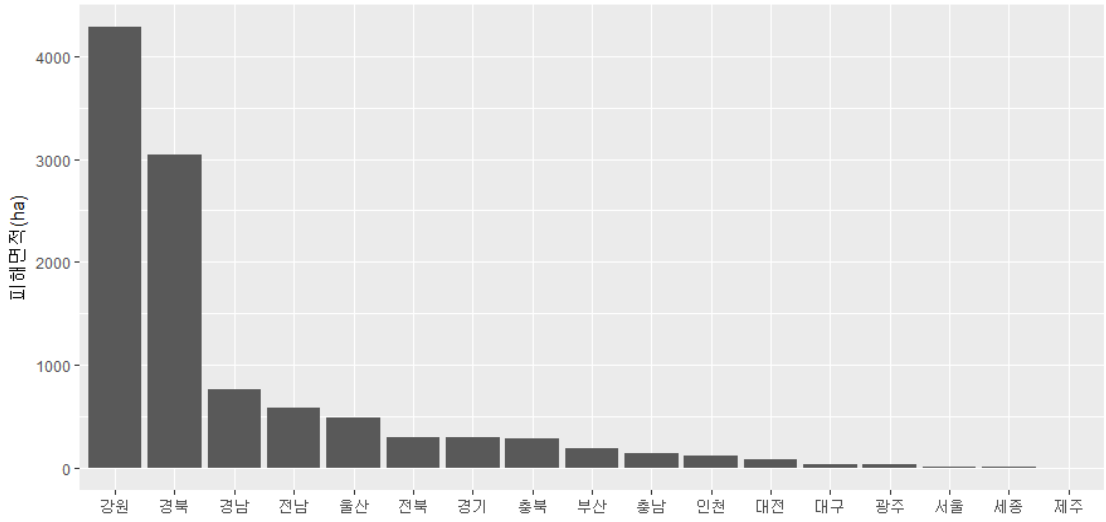
Ⅰ. 서 론

UN이 기후 변화 등 지구환경 위기극복을 위해 2011년을 ‘세계 산림의 해’로 정할만큼 국내외로 산림의 중요성에 대한 인식이 커지고 있다. 우리나라의 산림은 6·25전쟁과 전쟁 후 사회 혼란기에 심각하게 훼손되었다. 이후 정부의 적극적인 산림 조성 사업 계획으로 2015년 기준 산림 비율이 63.2%로 OECD 국가 중 4위에 해당한다.

산림 보존을 위한 노력에도 불구하고 산림의 파괴를 일으키는 산불은 2017년 692건, 피해 면적 1,480ha로 10년 사이 가장 큰 피해를 기록했다. 산불의 예방은 중요한 문제이며 산불이 발생할 가능성이 높은 시기에 인력과 자원을 집중적으로 배치한다면 효율적으로 운영할 수 있다. 산불 발생 원인파악 및 예측을 위한 연구는 2007년 캐나다 산불 기상지수를 이용한 산불발생확률모형 개발 연구, 2017년 한국의 산불 발생과 기상인자와의 관계 분석에 관한 연구 등이 있었다.

최근 4차 산업의 시대에는 빅데이터와 머신러닝을 활용한 예측모형의 정확도 향상이 가능해졌으며 많은 곳에서 활용되고 있다. 기상 관측소만 해도 전국 96개소의 종관기상관측장비(ASOS)와 494개소의 방재기상관측장비(AWS)가 있다. 본 연구에서는 최대한의 데이터 사용과 머신러닝 기법을 활용하여 산불의 예측 모형을 개발하였다.

본 연구의 산불 예측 모형은 일별 기상자료를 사용하였으며, 기상자료의 오차를 줄이기 위해 강원도 지역을 하나의 관측소 기상자료로 사용하는 것이 아니라 9개의 구역으로 나누어서 각 구역 관측소의 기상자료를 사용하였다. 하지만 구역을 나누어서 예측 모형을 만들면, 산불이 발생하는 날과 발생하지 않는 날의 비율이 차이가 큰 불균형 문제가 발생한다. 이러한 경우 일반적으로 예측모형의 성능이 떨어진다. 불균형 문제를 해결하기 위해 주로 사용되는 샘플링 방법을 사용하였다. 또한 통계적 기법인 로지스틱 회귀분석과 머신러닝 기법을 활용하여 예측 모형을 개발했다.

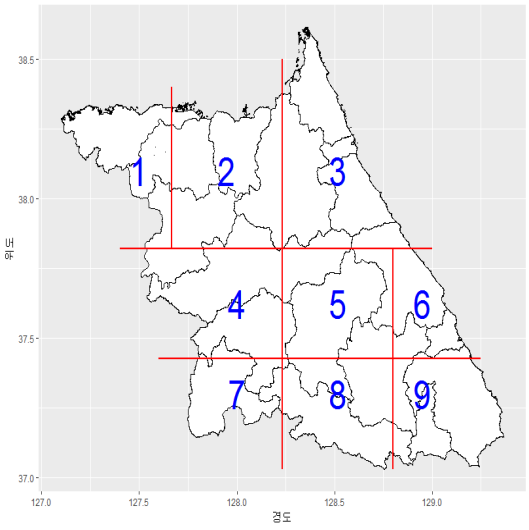


〈그림 1〉 시도별 피해면적

Ⅱ. 분석 데이터

2.1 연구 대상지

본 연구에서는 그림1과 같이 산불이 가장 많이 발생하는 강원도의 산불 예측모형을 개발했으며, 기상의 오차를 줄이기 위해 그림 2와 같이 9개 구역으로 나누어 각 구역에 맞는 기상자료를 사용하였다.



〈그림 2〉 강원도 9개 지역 세분화

2.2 산불 발생자료, 기상자료

산불 발생 데이터는 산림청에서 제공하는 전국 산불발생 통계자료로써 2003년부터 2016년까지 총 14년 산불 자료를 이용했다. 이 기간과 동일한 강원도 지역의 각 구역을 대표할 수 있는 기상자료를 구하기 위해 강릉, 대관령, 동해, 북강릉, 북춘천, 속초, 영월, 원주, 인제, 정선군, 철원, 춘천, 태백, 홍천의 종관기상관측 기상 자료를 기상청에서 수집하였고, 구역에 중복된 기상자료는 평균으로 사용하였다. 수집한 기상자료는 기온, 강수, 풍속, 습도 등이 있다.

2.3 캐나다 산불 기상지수

캐나다 산불 기상지수(Forest fire Weather Index, FWI)는 현재 캐나다에서 사용하고 있는 캐나다 산불위험 평가 시스템(Canadian Forest Fire Danger Rating System, CFFDRS)의 구성 요소로 캐나다 전 지역에서 현재 산불예방 및 진화 등에 활용되고 있다.(Van Wagner, 1987) FWI 시스템에서는 5가지의 지수를 사용하는데 평균기온, 평균 습도, 평균풍속, 강수량을 이용하여 계산한다. 미세 연료 지수(Fine Fuel Moisture Code, FFMC), 부식층지수(Duff moisture code, DMC), 가뭄지수(Drought code, DC)를 1차적으로 계산한 후 이들의 조합으로 ISI(Initial Spread Idex), BUI(Build Up Index)지수를 구한다. 기상데이터를 활용하여 Natural Resources Canada에서 제공하는 코드를 이용하여 5가지의 지수를 구하고 변수로 사용하였다.

Ⅲ. 분석 방법

통계 분석 툴로는 R 3.5.1 버전을 사용하였으며 2003년부터 2016년까지의 데이터를 임의(random)로 7:3의 비율로 train셋과 test셋으로 나누어 train셋으로 예측 모형을 만들고 test셋으로 모형의 성능을 검증했다. Random forest와 xgboost 방법은 임의성(random)을 포함하고 있는 모델로써 seed를 1234로 고정하고 모델을 만들었다. 각 모델의 예측 값은 0~1 사이의 값으로 나오는데 ROC curve의 AUC(Area under the curve)가 최대가 되는 값을 cut off로 정해 그 이상 되는 값을 1이라고 예측했다. 모델 성능에 대한 평가 지수로써 각 모형의 정확도, 민감도, 특이도를 구하였다.

정확도(Accuracy)는 전체 중 예측에 성공한 비율이다. 민감도(Sensitivity)는 실제 발생한 산불을 산불이 발생한다고 예측하는 비율이다. 특

이도(Specificity)는 실제 발생하지 않은 산불을 발생하지 않는다고 예측하는 비율이다. 3가지의 값이 모두 높은 모델이 좋은 예측 모형이라고 할 수 있다.

3.1 앙상블 모델

앙상블 모델이란 여러 개의 모형으로 나온 예측/분류 결과를 종합하여 최종적인 의사결정에 활용하는 방법으로 일반적으로 단일 모형보다 예측력이 높다. 앙상블 모델에는 부스팅(boosting), 배깅(bagging)이 있다.

배깅은 원 자료의 샘플링을 통해 여러 개의 모형을 만든 뒤 각각 모형의 결과를 투표(voting) 또는 평균(average)으로 결합하여 예측하는 방법으로 예측 값의 Variance을 감소시켜준다.

Schapire(1990)에 의해 소개된 부스팅은 기계학습 알고리즘으로서 오분류된 관측 값을 더 많이 사용함으로써 분류를 하기 힘든 관측 값에 대해 분류를 잘하도록 만든 방법으로 예측 값의 Bias를 감소시켜준다.

3.2 Random forest

Random forest는 의사결정트리의 앙상블 형태이다. 의사결정트리는 한번 분기 때마다 변수 영역을 구분하는 모델로 타겟 변수가 연속형/범주형에 상관없이 쓸 수 있으며, 모형을 설명하기 쉬운 장점이 있다. 하지만 수평 또는 수직 분할로 잘 나누어지지 않는 구조의 데이터가 있는 곳에서는 성능이 떨어진다. 또한 한 번에 하나의 변수를 처리하기 때문에 변수 간의 상호 작용을 잡지 못한다. 이러한 단점을 해결하기 위해 나온 방법이 random forest이다. 데이터를 샘플링 하여 여러 개의 트리 모형을 만든 뒤 각 트리들의 결과를 투표 또는 평균 내는 방식으로 배깅 방법과 유사하다. 그러나 각 모형에서 데이터의 샘플링뿐만 아니라 변수를 임의(random)로 선택함으로써 몇 개의 변수만 사용하므로 트리의 다중공선성 문제를 보완해준다.

3.3 XGBoost

Tianqi Chen과 Carlos Guestrin이 소개된 XGBoost방법은 eXtreme Gradient Boosting의 약자로 부스팅 방법 중 하나이다. XGBoost는 나무를 생성할 때 오분류된 관측 값을 다음 모델에 더 많이 사용하는 방식으로 오분류된 관측 값에 대해 성능을 더 향상시키는 방법으로 훈련하는 연속적인 훈련 알고리즘이다. 또한 학습하는 동안 모든 CPU 코어들을 사용하는 parallel computing으로 속도가 빠른 장점이 있고 Python, R 등 다양한 언어를 지원하기 때문에 유용성이 높다.

Ⅳ. 샘플링 방법

각 구역의 train셋의 산불의 발생일과 발생하지 않은 날의 차이는 표 2와 같이 크게 차이가 났다. 모델링 결과는 표 1과 같고 몇몇 구역의

〈표 1〉 구역별 산불의 발생일과 발생하지 않은 날의 수

지역번호	non-occurred	occurred
1	4338	30
2	4258	97
3	4301	74
4	4324	76
5	4398	14
6	4349	52
7	4326	41
8	4336	46
9	4357	32

〈표 2〉 기존 데이터 분석 결과

지역번호	기존 데이터								
	로지스틱 회귀분석			Random forest			xgboost		
	accur[1]	sens[2]	spec[3]	accur	sens	spec	accur	sens	spec
1	0.68	1.00	0.68	0.87	0.58	0.87	0.81	0.83	0.81
2	0.74	0.78	0.74	0.75	0.74	0.75	0.76	0.78	0.75
3	0.66	0.91	0.65	0.32	1.00	0.31	0.79	0.64	0.79
4	0.78	0.81	0.78	0.51	0.94	0.50	0.83	0.81	0.83
5	0.74	0.75	0.74	0.01	1.00	0.00	0.79	0.50	0.79
6	0.63	0.90	0.62	0.83	0.50	0.84	0.59	0.80	0.59
7	0.45	1.00	0.44	0.69	0.57	0.69	0.64	0.57	0.64
8	0.95	0.75	0.95	0.90	0.38	0.91	0.89	0.63	0.89
9	0.84	0.78	0.84	0.01	1.00	0.00	0.83	0.89	0.83

[1]accur : accuracy [2]sens : sensitivity [3]spec : specificity

결과에 대해서 민감도나 특이도 중 한쪽으로 쏠리는 결과가 나왔다. 이러한 불균형문제를 해결하기 위해서 언더/오버 샘플링 및 SMOTE 방법을 사용하였다.

4.1 언더샘플링 기법

언더샘플링 기법은 반응변수의 클래스 중 많은 범주의 클래스(major class)를 무작위로 제거해 데이터의 불균형 문제를 해결하는 방식이다. 언더샘플링 기법은 데이터의 양을 제거함으로써 모형구축 속도를 줄일 수 있지만, 정보가 손실되는 단점을 가지고 있다.

4.2 오버샘플링 기법

오버샘플링 기법은 반응변수의 클래스 중 적은 범주의 클래스(minor class)를 무작위로 복제해 데이터의 불균형 문제를 해결하는 방식이다. 오버샘플링 기법은 데이터의 양이 증가함으로 인해 모형구축 속도가 증가하고, 소수 범주를 복제하기 때문에 overfitting의 문제가 발생 할 수 있다.

4.3 SMOTE 방법

SMOTE (Synthetic Minority Over-Sampling Technique) 기법은 오버샘플링의 단점인 over-fitting을 보완하기 위한 방법이다.

반응변수의 클래스 중 적은 범주의 클래스(minor class) 중 무작위로 하나를 선택한 후, 이 데이터의 k개의 근접 이웃을 찾는다. 그리고 선택된 하나의 샘플과 k개의 이웃의 차를 구한 다음 이 차이에 0 ~ 1 사이의 임의의 값을 곱하여 기존 샘플에 더한 후 훈련 데이터에 추가한다. 이 과정을 반복한다.

SMOTE 알고리즘은 적은 범주의 클래스(minor class)의 자료를 늘린다는 점에서 오버샘플링과 유사하지만, 같은 자료를 복제하는 것이 아니라 기존의 자료를 적절히 조합하여 새로운 샘플을 만듦으로써 오버샘플링의 단점인 over-fitting을 보완한다고 알려져 있다.

〈표 3〉 로지스틱 회귀분석 결과

지역번호	로지스틱 회귀분석								
	언더샘플링			오버샘플링			SMOTE		
	accur	sens	spec	accur	sens	spec	accur	sens	spec
1	0.65	1.00	0.64	0.71	1.00	0.71	0.67	0.92	0.67
2	0.60	0.83	0.59	0.55	0.91	0.54	0.72	0.78	0.72
3	0.61	1.00	0.61	0.64	0.91	0.64	0.61	0.91	0.61
4	0.78	0.75	0.78	0.66	0.88	0.66	0.65	0.94	0.65
5	0.91	0.75	0.91	0.81	0.63	0.81	0.81	0.50	0.81
6	0.64	0.90	0.63	0.66	0.80	0.66	0.55	0.90	0.55
7	0.94	0.43	0.94	0.88	0.43	0.88	0.37	1.00	0.36
8	0.93	0.75	0.93	0.93	0.75	0.93	0.87	0.75	0.87
9	0.82	0.78	0.82	0.58	1.00	0.58	0.65	0.89	0.65

〈표 4〉 Random forest 분석 결과

지역번호	Random forest								
	언더샘플링			오버샘플링			SMOTE		
	accur	sens	spec	accur	sens	spec	accur	sens	spec
1	0.85	0.75	0.85	0.89	0.50	0.90	0.63	0.83	0.63
2	0.81	0.74	0.81	0.77	0.61	0.77	0.64	0.78	0.64
3	0.48	1.00	0.48	0.54	0.82	0.54	0.59	1.00	0.58
4	0.74	0.88	0.74	0.55	0.94	0.54	0.77	0.81	0.77
5	0.65	0.75	0.65	0.92	0.50	0.93	0.61	1.00	0.60
6	0.60	0.90	0.60	0.68	0.70	0.68	0.72	0.70	0.72
7	0.49	0.86	0.49	0.01	1.00	0.00	0.64	0.86	0.64
8	0.88	0.63	0.88	0.81	0.63	0.81	0.79	0.75	0.79
9	0.68	0.67	0.68	0.64	0.78	0.64	0.74	0.67	0.74

〈표 5〉 xgboost 분석 결과

지역번호	xgboost								
	언더샘플링			오버샘플링			SMOTE		
	accur	sens	spec	accur	sens	spec	accur	sens	spec
1	0.77	0.92	0.76	0.60	1.00	0.59	0.67	1.00	0.67
2	0.75	0.78	0.75	0.78	0.74	0.78	0.75	0.70	0.75
3	0.72	0.82	0.71	0.49	1.00	0.48	0.62	0.91	0.62
4	0.70	0.81	0.70	0.63	0.94	0.62	0.72	0.81	0.72
5	0.58	0.88	0.58	0.63	0.88	0.63	0.65	1.00	0.65
6	0.45	1.00	0.44	0.93	0.60	0.93	0.53	0.80	0.53
7	0.17	1.00	0.17	0.73	0.57	0.74	0.51	0.86	0.51
8	0.72	0.75	0.72	0.78	0.63	0.78	0.93	0.63	0.93
9	0.57	1.00	0.57	0.75	0.78	0.75	0.75	0.78	0.75

〈표 6〉 최종 분석결과

지역번호	사용모델	샘플링	accur	sens	spec
1	XGboost	사용안함	0.81	0.83	0.81
2	Random forest	언더샘플링	0.81	0.74	0.81
3	XGboost	언더샘플링	0.72	0.82	0.71
4	로지스틱회귀분석	사용안함	0.83	0.81	0.83
5	Random forest	언더샘플링	0.91	0.75	0.91
6	Random forest	SMOTE	0.72	0.70	0.72
7	Random forest	SMOTE	0.64	0.86	0.64
8	로지스틱회귀분석	사용안함	0.95	0.75	0.95
9	로지스틱회귀분석	사용안함	0.75	0.78	0.75

〈표 7〉 총 산불발생 예측 결과표

observed	predicted		Accuracy (%)
	Yes	No	
Yes	80	24	76.9%
No	2332	7426	76.1%
		Overall	76.1%

Ⅴ. 연구결과

각 샘플링 방법과 모형의 장·단점으로 구역에 맞는 방법은 다르게 나타났다. 최종 모형으로써 민감도와 특이도의 값이 0.7이상이면서 정확도가 가장 높은 모형을 최종모형으로 선택하였으며, 7번 구역에서는 기준을 낮추어 특이도 민감도의 기준을 0.6으로 잡고 최종 모델을 결정하였다.

Ⅵ. 결 론

민감도와 특이도는 하나의 값이 높아지면 다른 값이 낮아지는 trade-off 관계에 있다. 따라서 민감도와 특이도의 비율은 모델을 사용하는 연구자의 판단으로 선택하게 된다. 본 논문에서는 민감도와 특이도가 각각 0.6, 0.7 이상이면서 정확도가 가장 높은 모델을 최종 모델로 표 8 과 같이 선택하였다. 최종모델로 9개 구역의 test셋의 예측 결과를 표 7과 정리하였다. 산불이 발생한 104건 중 80건의 발생 예측에 성공하였으며 산불이 발생하지 않은 9758건 중 7426건의 발생하지 않음을 예측했다. 전체의 정확도는 76.1%였다.

참 고 문 헌

[1] 박흥석, 이시영, 채희문, 이우균 (2009) **현캐나다 산불 기상지수를 이용한 산불 발생 확률모형 개발**, 한국방재학회논문집, 제9권, 제3, pp. 95~100.

[2] 이병두, 유계선, 김선용, 김경하 (2012) **로지스틱 회귀모형을 이용한 산불발생확률모형 개발**, 한국임학회지, Vol. 101, No. 1, pp. 1-6.

[3] Amiro, B.D., Logan, K.A., Wotton, B.M., Flanigan, M.D., Todd, J.B., Stocks, B.J. and Martell, D.L. (2004) *Fire Weather index system components for large fires in the Canadian boreal forest.* International Journal of Wildland Fire, Vol 13, pp. 391-400.

[4] Breiman, L. (2001) *Random Forests. Machine Learning*, Vol. 45, No. 1, pp. 5-32.

[5] Chawla, N. V., Bowyer, K. W., Hall, L. O.,

& Kegelmeyer, W. P.(2002). *SMOTE: synth-etic minority over-sampling technique.*

[6] Freund, Y. and Schapire, R. (1996), *Experim-ents with a new boosting algorithm, Machine Learning : Proceedings of the Thirteenth International Conf erence*, San Francisco, USA, 148-156.

[7] Gareth J, Daniela W, Trevor H, Robert T (2015), *"An Introduction to Statistical Learning withAppli cations in R"*, Springer, NewYork.

[8] XGBoost (2016) https://xgboost.readthedocs-.io/en/latest/

저 자 소 개

채 경 재(Kyoung-jae Chae)

- 2017 : 인하대학교 통계학과 (학사)
- 2019 : 인하대학교 통계학과 (석사)
- 관심분야 : Big Data Analytics, Data Mining, Finance data

이 유 리(Yu-Ri Lee)

- 2017 : 인하대학교 통계학과 (학사)
- 2019 : 인하대학교 통계학과 (석사)
- 관심분야 : Big Data Analytics, Data Mining, Recommender System

조 용 주(yong-ju cho)

- 2017 : 인하대학교 산업공학과(학사)
- 2019 : 인하대학교 통계학과 (석사)
- 관심분야 : Big Data Analytics, Data Mining, Machine Learning

박 지 현(Ji-Hyun Park)

- 2017년 : 인하대학교 통계학과(학사)
- 2019년 : 인하대학교 통계학과(석사)
- 관심분야 : Big Data Analytics, Data Mining, AI