날씨 빅데이터 콘테스트
산사태 예측 모델 구현

A2W

김홍범, 곽희원, 이원권, 최디도

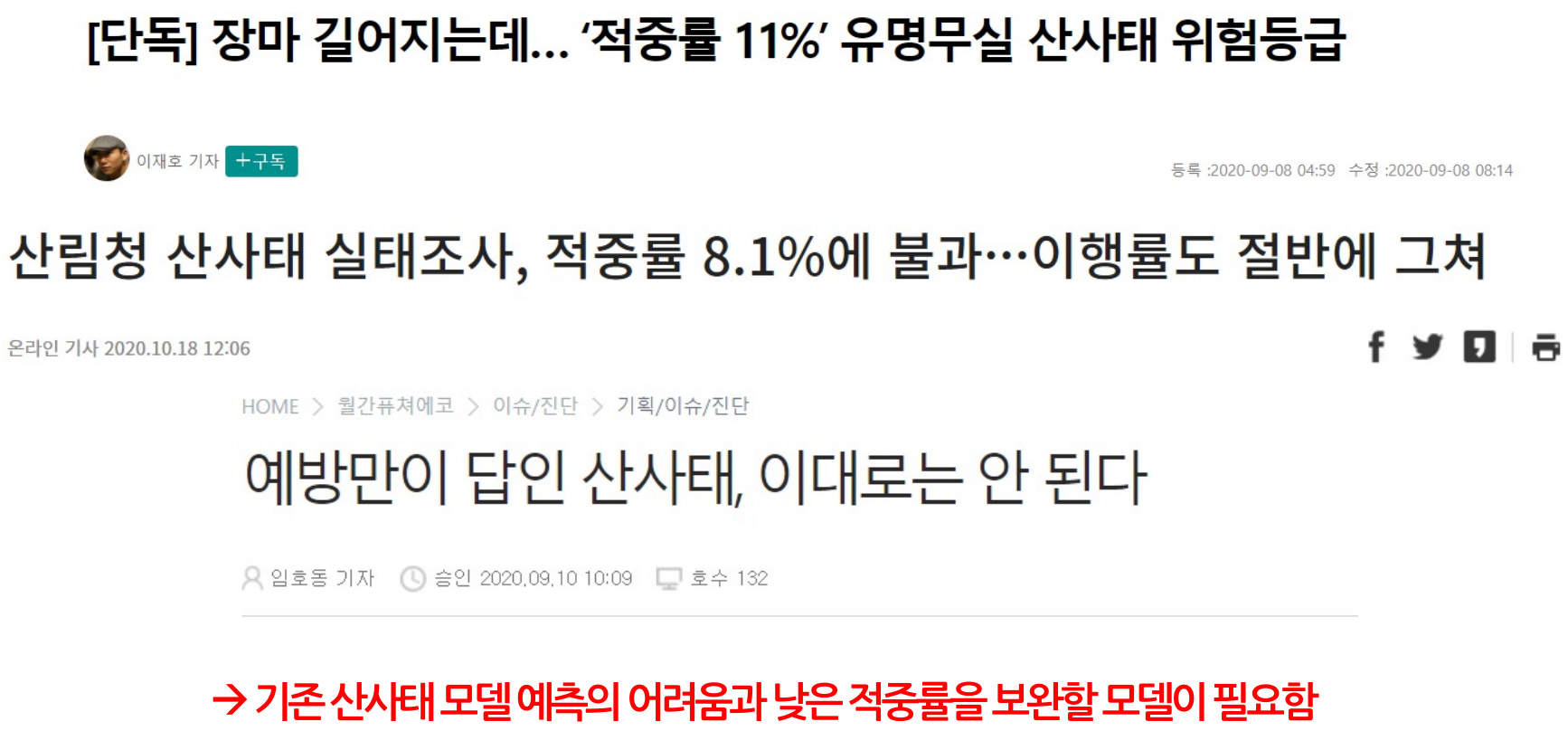
CONTENTS

배경

배경

01. 목적 : 산사태 예측 모델을 구축하게 된 이유

적중하기 어려운 산사태에 대해, 보다 정확한 모델 구축의 필요성



[단독] 장마 길어지는데... '적중률 11%' 유명무실 산사태 위험등급

이재호 기자 +구독

등록 :2020-09-08 04:59 수정 :2020-09-08 08:14

산림청 산사태 실태조사, 적중률 8.1%에 불과…이행률도 절반에 그쳐

온라인 기사 2020.10.18 12:06

HOME > 월간퓨쳐에코 > 이슈/진단 > 기획/이슈/진단

예방만이 답인 산사태, 이대로는 안 된다

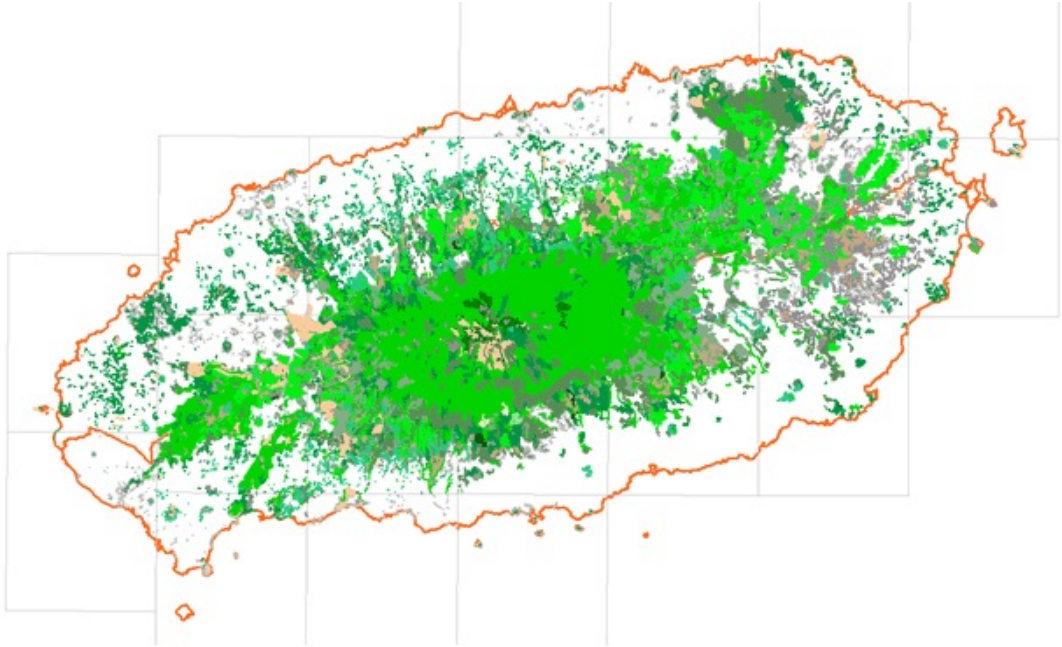
임호동 기자 승인 2020.09.10 10:09 호수 132

→ 기존 산사태 모델 예측의 어려움과 낮은 적중률을 보완할 모델이 필요함

배경

02. 임상도 정의

우리나라의 산림이 어떻게 분포하고 있는가를 보여주는 대표적인 산림지도

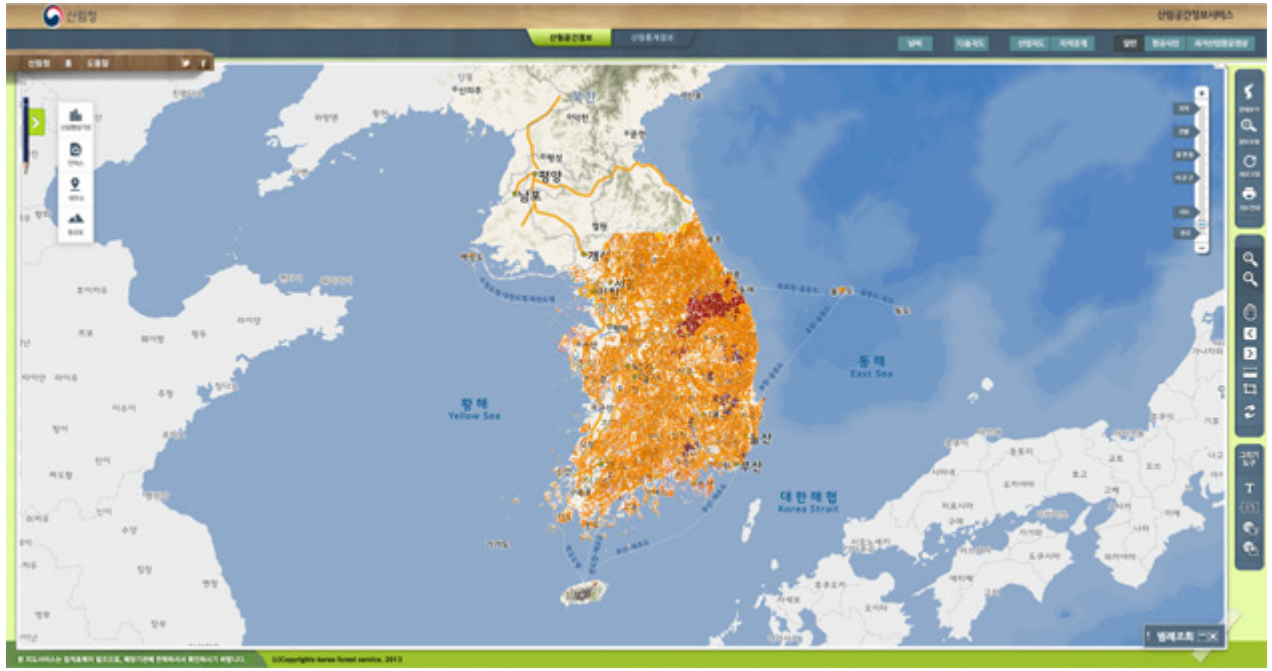


임상도 예시 (제주도)

배경

03. 토양도 정의

산림경영, 산지관리, 환경영향평가 등에 필요한 입지,토양환경에 대해 작도단위인
토양형을 구획단위로 조사 및 분석한 정보를 대축척화 하여 수치지도로 나타낸 산림주제도

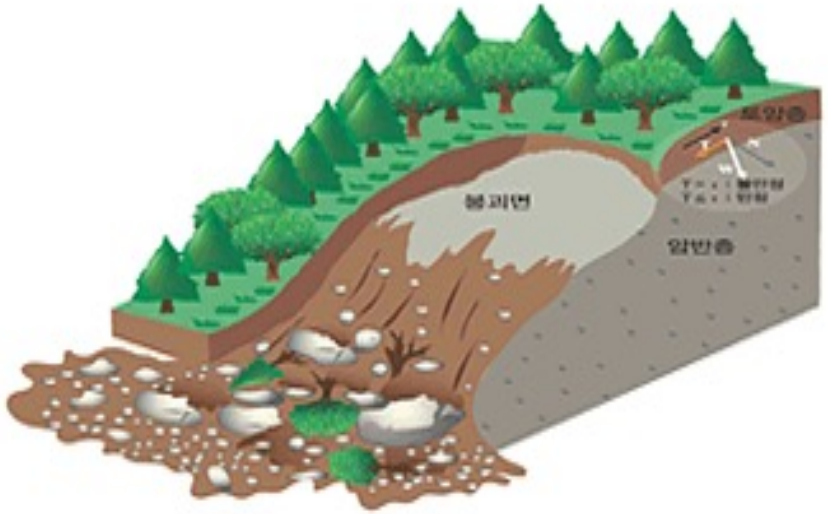


토양도 예시 (전국)

배경

04. 산사태 정의 : 사방사업법 제 2조 5호, 6호

산사태 : 자연적 또는 인위적인 원인으로 산지가 일시에 붕괴되는 것
토석류 : 산지 또는 계곡에서 토석, 나무 등이 물과 섞여 빠른 속도로 유출되는 것



산사태 유형 A

산사태 유형 B(토석류)

배경

05. 날씨 데이터 정의 : ASOS 및 AWS

ASOS : 전국 96개소의 종관기상관측장비로 지방청, 지청, 기상대, 관측소 등에 설치
AWS : 전국 494개소의 방재기상관측장비로 산악지역이나 섬처럼 사람이 관측하기 어려운 곳에 설치

구분	기압	기온	습도	풍향 풍속	강수 mm		일사 일조	지면/초상/지중온도
					0.5	0.1		
ASOS	O	O	O	O	O			
AWS	O	O	O	O	O	O	O	O

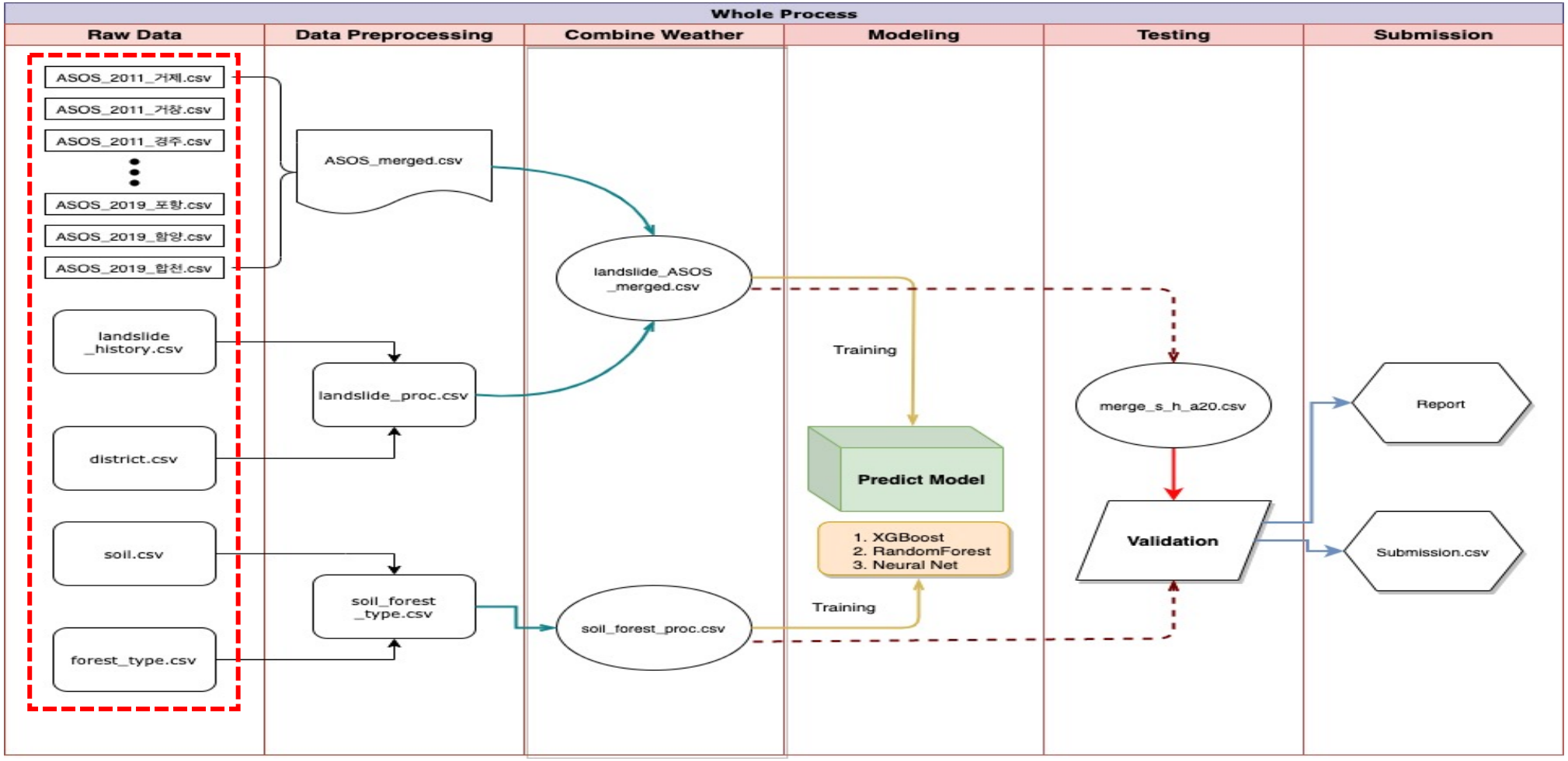
→ AWS 데이터는 고도가 다르기 때문에 경상북도, 경상남도의 큰 단위에서 항상성 있는 예측 모델을 구축하기 위해 사용하지 않음.

II

활용 데이터

활용 데이터

01. 전체 프로세스



활용 데이터

02. ASOS 데이터

ASOS_day_sd.csv

변수명	정의	설명	예시
numOfRows	한 페이지 결과 수	한 페이지당 표출되는 데이터의 수	1, 2, 3
pageNo	페이지 번호	페이지 번호를 나타내는 코드	1, 2, 3
totalCount	데이터 총 개수	데이터 개수	445
resultCode	응답메시지 코드	응답 메시지코드	00
resultMsg	응답메시지 내용	응답 메시지 설명	NORMAL SERVICE
dataType	데이터 타입	응답자료형식 (XML/JSON)	XML
stnId	지점 번호	종관기상관측 지점 번호 첨부 참조	108
stnNm	지점명	종관기상관측 지점명 첨부 참조	서울

→ 기상청02_지상(종관,ASOS)일자료_조회서비스_오픈API활용가이드 참고

활용 데이터

03. 산사태 발생 이력 데이터

landslide_history.csv

변수명	정의	설명	예시
date	산사태 발생 일자	Char, 산사태 발생 일자 코드	yyyymmdd
sd	산사태 발생 지역_시도	Char, 산사태 발생 지역 코드	경상북도/경상남도
sgg	산사태 발생 지역_시군구	Char, 산사태 발생 지역 코드	포항시
umd	산사태 발생 지역_읍면동	Char, 산사태 발생 지역 코드	흥해읍
sum_cnt	산사태 발생 횟수의 합	Int, 산사태 발생 횟수 코드	1
sum_hpa	산사태 발생 면적의 합	Num, 산사태 발생 면적 코드	7

활용 데이터

04. 행정동 데이터

district.csv

변수명	정의	설명	예시
OBJECT_ID	객체 식별자	Num	순번으로 생략 가능
BASE_DATE	기준년월일	Char	20200630
ADM_DR_CD	행정동코드	Char	637개
ADM_DR_NM	행정동명	Char	
GEOM	공간정보	Multipolygon, geopandas 사용	

활용 데이터

05. 임상도 데이터

forest_type.csv

변수명	정의	설명	예시
STORUNST_CD	입목존재 코드	숲에 나무가 어느 정도 있는지를 나타내는 지표	0, 1, 2
FROR_CD	임종 코드	인공림의 유무 판단하는 코드	0, 1, 2
FRTP_CD	임상 코드	어떤 나무가 주로 숲에 분포	0, 1, 2, 3, 4
KOFTR_GROUP_CD	수종그룹 코드	어떤 나무 종류 인지 나타내는 변수	10~99
DMCLS_CD	경급 코드	나무가 얼마나 큰지	0, 1, 2, 3
AGCLS_CD	영급 코드	나무 얼마나 오래 됐는지	1~9
DNST_CD	밀도 코드	교목(키가 8미터이상인 나무)중 수관면적비율로 판단하는 변수	A, B, C
HEIGT_CD	임분고 코드	임분고의 높이	0, 2, ··, 38, 40

활용 데이터

06. 토양도 데이터

soil.csv

변수명	정의	설명	예시
OBJ_ID	객체ID	객체 구분 숫자입니다.	1,2,3, …
ARA_XCRD	지역X좌표	조사 지역의 X 좌표	280272.57
ARA_YCRD	지역Y좌표	조사 지역의 Y 좌표	266104.66
PRRCK_LARG	모암대 코드	암석의 종류(모암대 기준)	0, 1, 2
PRRCK_MDDL	모암중 코드	암석의 종류(모암중 기준)	11 ~ 34
LOCTN_ALTT	입지표고	조사 지역의 고도	799.2
LOCTN_GRDN	입지경사도	조사 지역의 경사각	21.8
EIGHT_AGL	8방위각도	경사의 8방위각도	-1, 0~369

활용 데이터

06. 토양도 데이터

soil.csv

변수명	정의	설명	예시
CLZN_CD	기후대코드	지역의 기후대(온대, 난대)	1, 2, 3, 4
TPGRP_TPCD	지형구분코드	지역의 지형(산정, 산복, 계곡)	1~12, 99
PRDN_FOM_C	사면형태코드	경사면 형태 코드	3
SLANT_TYP	경사형코드	경사형(상승, 평행, 하강)	1, 2, 3
SLDPT_TPCD	토심구분코드	땅의 깊이(30cm, 60cm 기준)	10, 20, 30
SCSTX_CD	토성코드	흙의 성질(사양토, 양토, 등등)	1~11, 99
SLTP_CD	토양형코드	지영의 토양형(갈색건조산림, 거주지)	1~29, 82~99
LDMARK_STN	지형지물표준코드	지형 지물을 구분해주는 코드	L102

활용 데이터

06. 토양도 데이터

soil.csv

변수명	정의	설명	예시
CLZN_CD	기후대코드	지역의 기후대(온대, 난대)	1, 2, 3, 4
TPGRP_TPCD	지형구분코드	지역의 지형(산정, 산복, 계곡)	1~12, 99
PRDN_FOM_C	사면형태코드	경사면 형태 코드	3
SLANT_TYP	경사형코드	경사형(상승, 평행, 하강)	1, 2, 3
SLDPT_TPCD	토심구분코드	땅의 깊이(30cm, 60cm 기준)	10, 20, 30
SCSTX_CD	토성코드	흙의 성질(사양토, 양토, 등등)	1~11, 99
SLTP_CD	토양형코드	지영의 토양형(갈색건조산림, 거주지)	1~29, 82~99
LDMARK_STN	지형지물표준코드	지형 지물을 구분해주는 코드	L102

활용 데이터

07. 최종 데이터

Merge_final_real.csv

날짜	행정동	온도	…	강수량	임종	…	임상	토성	…	경사도	…	산사태
2011-01-01	울릉군 울릉읍	0.8	…	0	2	…	2	0	…	0	…	0
2016-10-05	울릉군 울릉읍	16.9	…	104.5	2	…	2	0	…	0	…	1
…	…	…	…	…	…	…	…	…	…	…	…	…
2019-10-03	영양군 일월면	16.9	…	21.9	2	…	2	4	…	20	…	1

- 행정동 변수 : 2개
- ASOS 날씨 데이터 : 14개
- 임상도 데이터 : 12개
- 토양도 데이터 : 7개

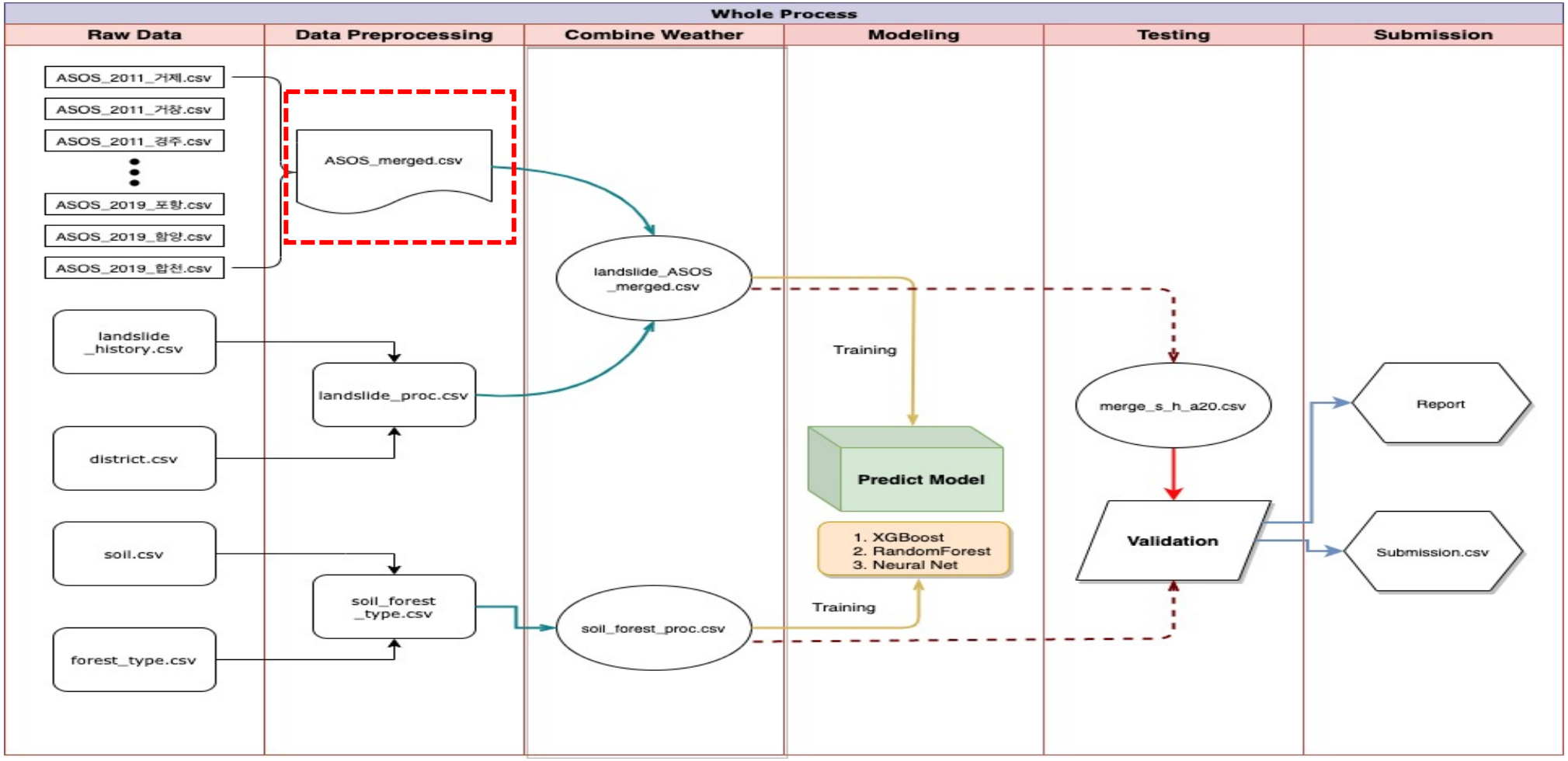
  병합 후 약 10개의 ROW, 35개의 Feature

III

데이터 전처리

데이터 전처리

01. ASOS 데이터



데이터 전처리

01. ASOS 데이터

ASOS_merged.csv

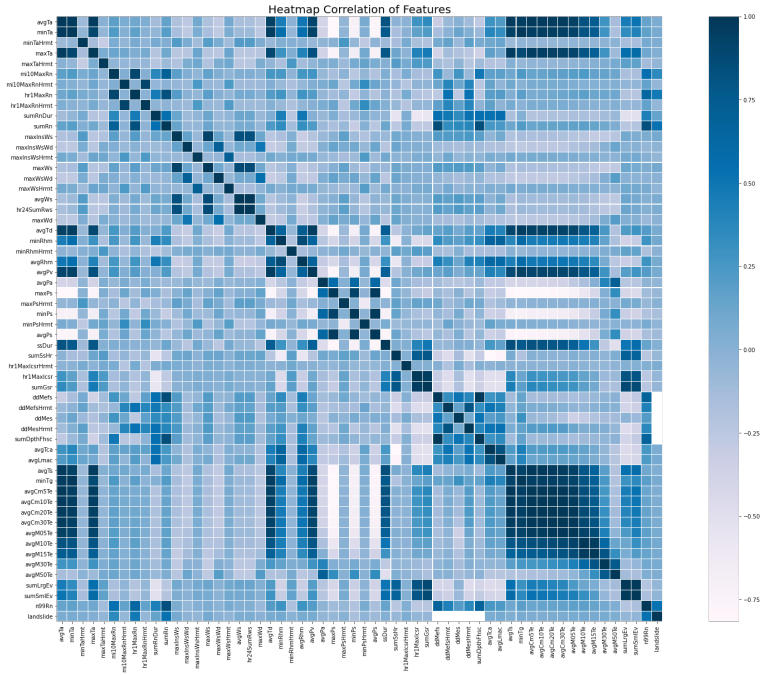
ASOS_2011_sd.csv

+

...

+

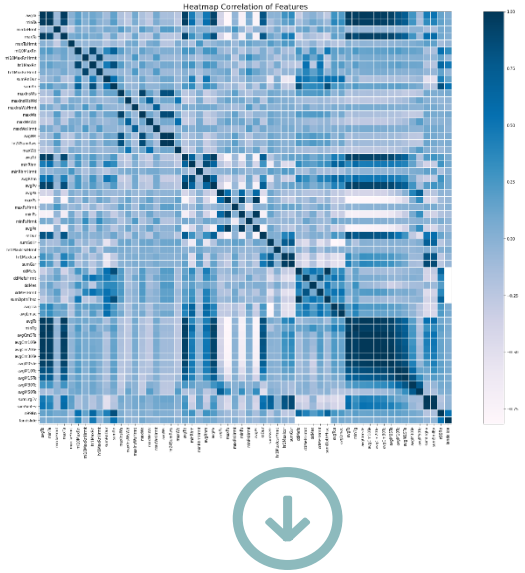
ASOS_2019_sd.csv

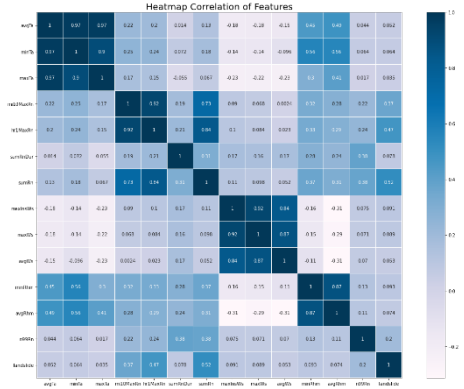


데이터 전처리

01. ASOS 데이터

ASOS_merged.csv





- ASOS의 경우 불필요한 변수가 너무 많아 Heatmap을 작성하여 실제 산사태 발생 여부에 관련 없는 변수는 과감히 제거

A. 시각 데이터 (Hrmt), 일조, 기압, 운량, 지중 온도 , 안개, 이슬점 등의 경우 너무 낮은 상관 계수를 지녔기에 산사태 발생에 큰 영향이 없다고 판단하여 제거

B. 적설량 데이터의 경우 상관 계수가 NaN으로 산사태와 눈은 상관이 없으므로 적설 관련 데이터 삭제

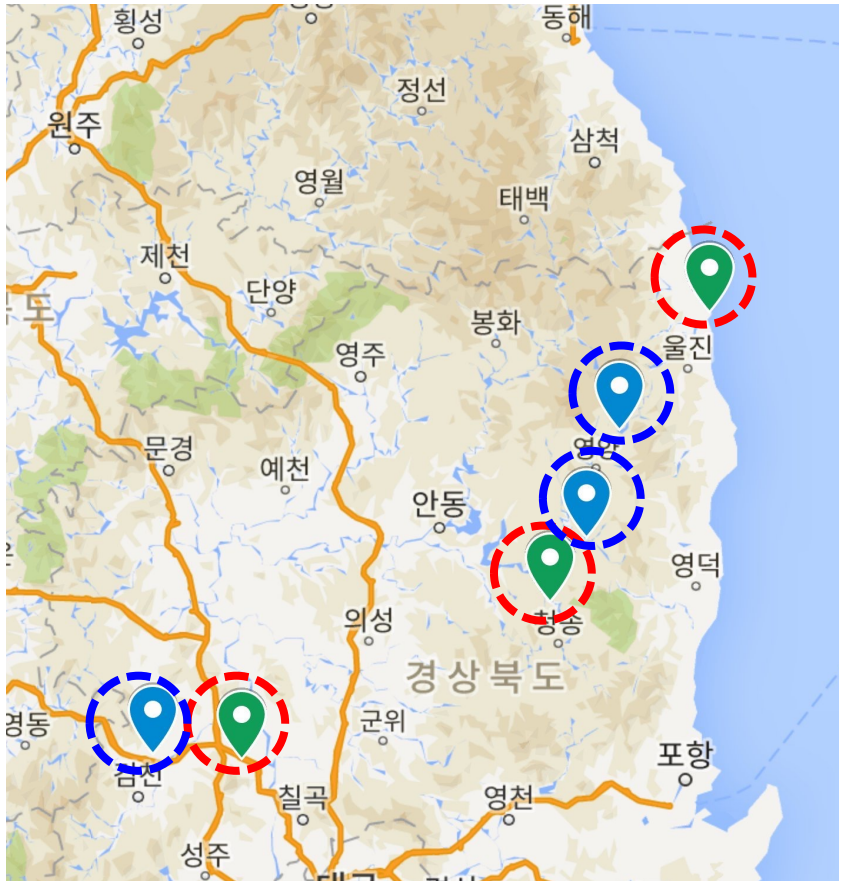
C. 온도의 경우 평균, 최저, 최고 기온을 제외한 지면 온도나 초상온도의 경우 내용적으로 중복된다고 판단하여 삭제

D. 강수량 결측치는 0으로 처리, 온도 결측치는 중간 값을 이용하여 보간

데이터 전처리

01. ASOS 데이터

ASOS_merged.csv

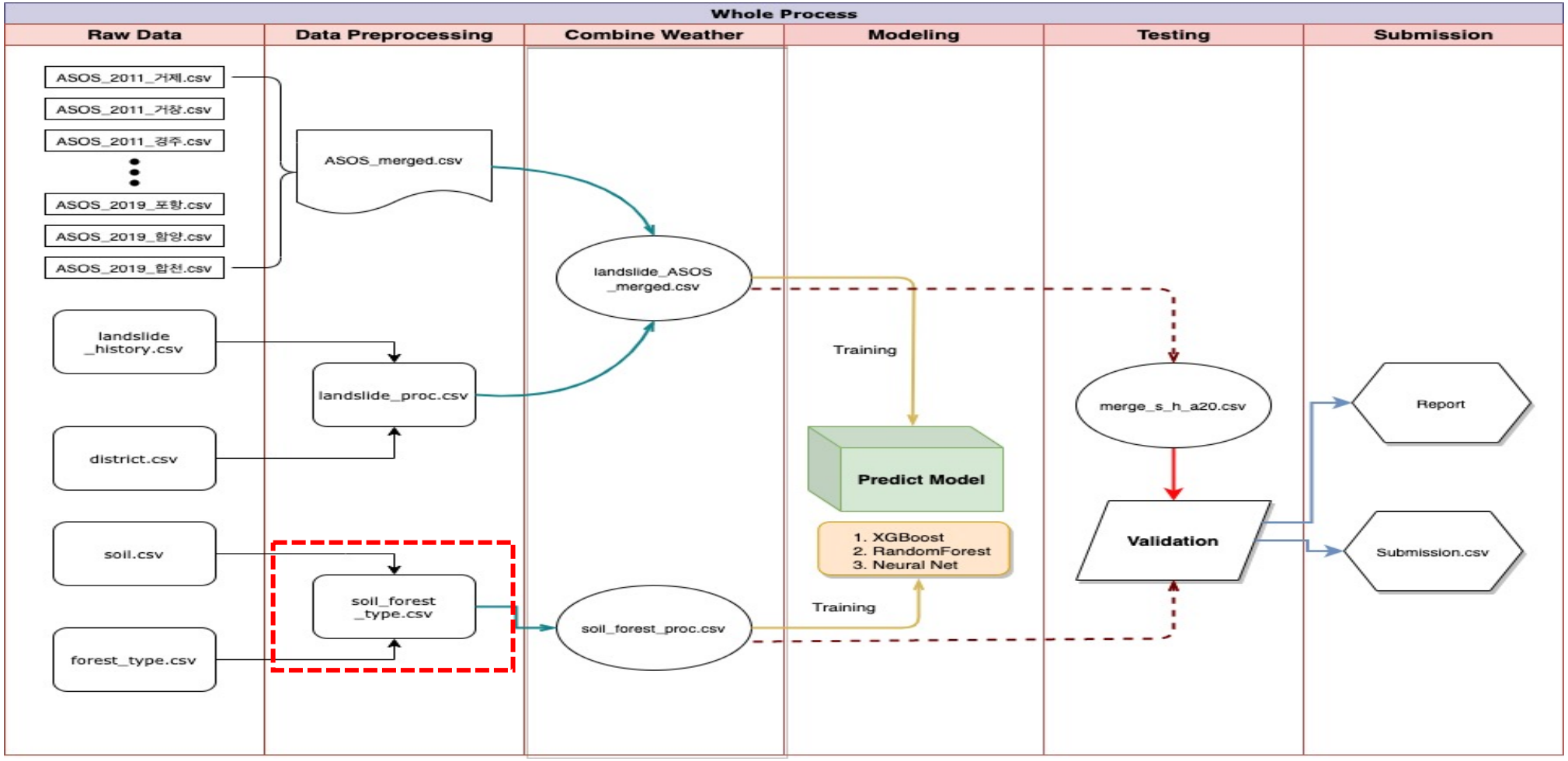


- 지도의 파란색 포인트들은 산사태가 발생했던 지역들로 각각 영양 석보면(우측 상단), 영양 수비면(우측 중단), 김천시(좌측 하단)
- 파란색 포인트 도시 내에는 ASOS 측정소가 존재하지 않는 관계로 산사태 발생 지역 읍면동 기준 가장 가까운 측정소의 날씨와 매칭
- 김천시의 경우 가장 가까운 구미 기상대의 날씨와 매칭하고 영양군의 남부 지역은 청송 기상관측소, 그리고 북부 지역은 울진 기상대와 매칭(빨간색 포인트로 매칭)

→ ASOS 데이터가 존재하지 않는 지역별 예외 장소 가까운 곳으로 매칭

데이터 전처리

02. 토양도 데이터 + 임상도 데이터



데이터 전처리

02. 토양도 데이터 + 임상도 데이터

soil_forest_type.csv

- 임상도, 토양도의 경우 범주형 데이터의 비중이 높아 HEATMAP 분석이 힘듦

A. 객관적으로 불필요한 변수 제거 : 공간정보 변수 : "geometry", "area", 병합시 추가된 변수 : "index_left"

B. 불필요한 변수 제거
행정동 구분 코드랑 의미가 같은 " BASE_DATE " , " ADM_DR_CD " 제거
산사태 발생횟수와 의미가 같은 'sum_cnt' 제거
기타 특이사항, 지형지물표준코드, 맵라벨 코드 분석 시 의미가 없으므로 제거

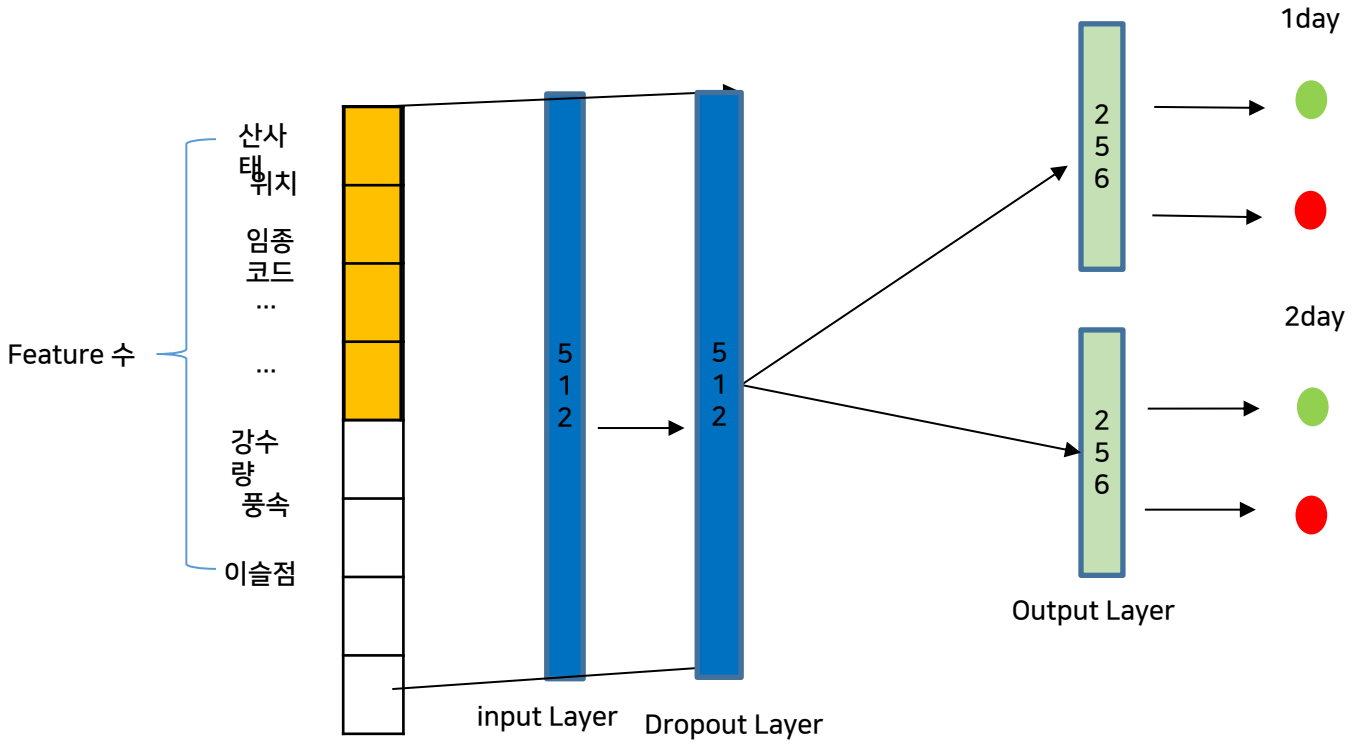
C. 결측치 처리
나무가 없는 지역은 임상정보가 없으므로 임의적으로 0값 부여

IV

분석 기법 및 결과

분석 기법 및 결과

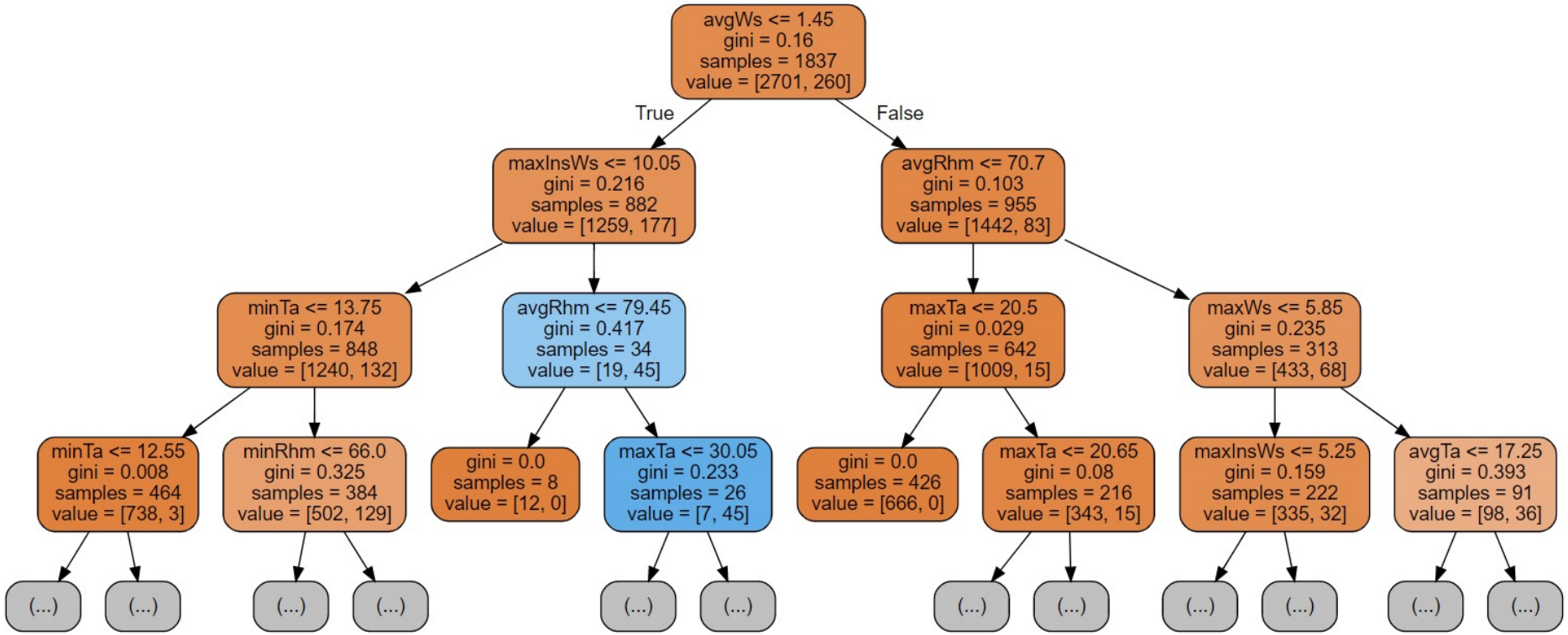
01. DNN(Deep Neural Network)



- 입력층과 출력층 사이에 여러 개의 은닉층(Hidden Layer)로 이루어진 인공신경망
- 입력 변수들 간의 비선형 조합 가능
- 활성화 함수 → 입력층 : 'relu', 출력층 : ' sigmoid', -> loss : BinaryCrossentropy
- Positive value의 비중이 0.02%에 불과한 unbalancing 데이터이므로 resamping을 진행
- 추후, 과적합 방지를 위해 재교육도 실시

분석 기법 및 결과

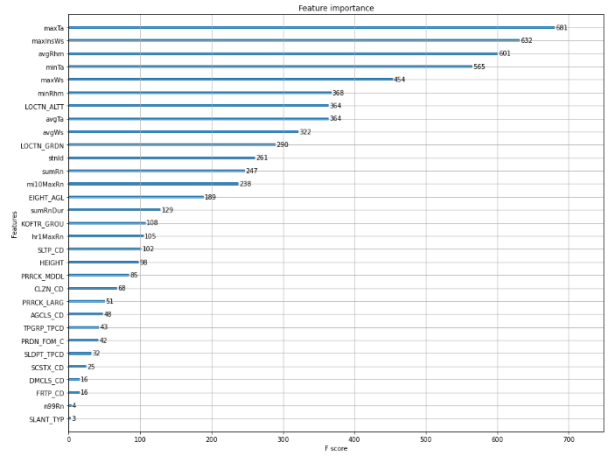
02. RF(Random Forest)

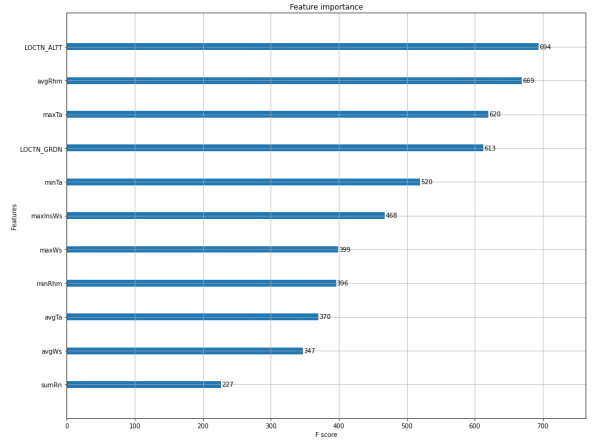


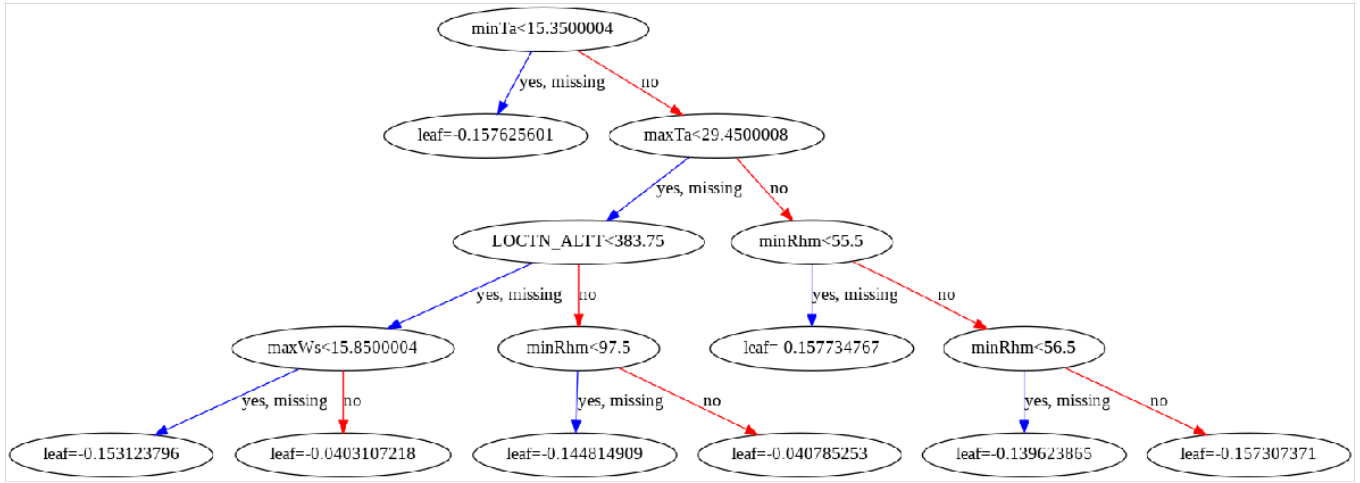
- 지도 분류 학습 알고리즘으로 Decision Tree 기반의 앙상블 학습 방법의 일종
- Overfitting 해결을 위하여 Max_Depth 설정
- GridSearchCV 를 이용하여 복수의 하이퍼 파라미터 최적화 (Train set : Test set = 75 : 25)

분석 기법 및 결과

03. XGBoost(eXtreme Gradient Boosting)







- Gradient Boosting 계열 알고리즘으로 학습 속도가 느리고 과적합 이슈를 보완
- Regression, Classification 문제를 모두 지원하며 성능과 자원 효율이 좋음
- GridSearchCV 를 이용하여 복수의 하이퍼 파라미터 최적화 (Train set : Test set = 75 : 25)
- XGBoost의 Feature Importance를 통해 전체 변수 중 중요 변수만 이용하여 새롭게 학습

분석 기법 및 결과

04. 결과

모델	Under sampling	Accuracy	CSI
DNN	1: 30	70.26%	2.24%
RandomForest	1:10	95.1%	0%
RandomForest	1:20	94.78%	1.57%
XGBoost	1:5	87.68%	4.03%
XGBoost	1:10	93.27%	4.64%
XGBoost	1:15	90.81%	4.3%
XGBoost	1:20	93.07%	3.24%

최종 모델의 평균 ACC는 93.27%, CSI는 4.64%

V

활용 방안과 효과

활용 방안과 효과

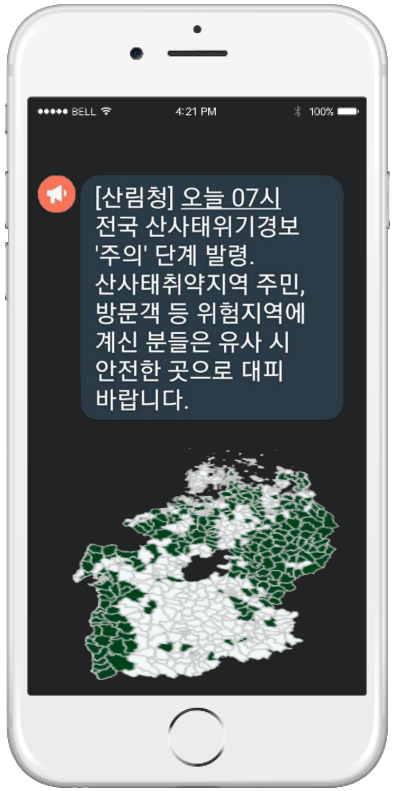
01. 활용 방안

산사태 발생 예측 가능성을 GIS로 시각화 →데이터베이스 구축

활용 방안과 효과

02. 기대효과

기존 산림청 산사태 재난 경보 문자를 GIS 데이터와 함께 제공



→ 기존 재난 경보 문자보다 훨씬 더 시각적인 경고 메시지로 인식할 수 있음

→ 지역을 구체적으로 확인할 수 있어, 예방 효과 UP

VI

관련 논문

관련 논문

01. 논문 목록

1. Dou, J., Yunus, A. P., & Merghadi, A., et all. (2021). “A Comparative Study of Deep Learning and Conventional Neural Network for Evaluating Landslide Susceptibility Using Landslide Initiation Zones.” Understanding and Reducing Landslide Disaster Risk, 215–223.

2. 최선규(Choi, Sun-Gy), 석재욱(Suk, Jae-Wook), and 정향선(Jeong, Hyang-Seon). "토사비탈면 붕괴에 대한 계측관리기준 제안: 변위를 기준으로." 한국방재학회논문집 21.1 (2021): 251-260.

3. Rahimi, S., Wood, C.M. & Bernhardt-Barry, M. “The MHVSR technique as a rapid, cost-effective, and noninvasive method for landslide investigation: case studies of Sand Gap and Ozark, AR, USA.” Landslides (2021). https://doi.org/10.1007/s10346-021-01677-7

4. Merghadi, Abdelaziz et al. “Machine Learning Methods for Landslide Susceptibility Studies: A Comparative Overview of Algorithm Performance.” Earth-Science Reviews 1 Aug. 2020. Earth-Science Reviews. Web.

5. 남경훈(Nam, Kounghoon), 김만일(Kim, Man-il), 권오일 (Kwon, Oil), 왕파우 (Wang, Fawu), and 정교철(Jeong, Gyo-cheol). "AutoML을 이용한 산사태 예측 및 변수 중요도 산정." 지질공학 30.3 (2020): 315-325.

6. 알-마문(Al-mamun), 장동호(Jang, Dong-ho),and 박종철(Park, Jongchul). "산사태 분포 예측을 위한 로지스틱, 베이지안, Maxent의 비교." 한국지형학회지 24.2 (2017): 91-101.

7. 마호섭(Ma, Ho Seop), 강원석(Kang, Won Seok), and 이성재(Lee, Sung Jae). "지역산림환경을 기반으로 한 산사태 발생 위험성의 예측 및 평가." 한국산림과학회지 103.2 (2014): 233-239.

8. 김호걸(Kim, Hogul), 이동근(Lee, Dong Kun), 모용원(Mo, Yongwon), 길승호(Kil, Sungho), 박찬(Park, Chan), and 이수재(Lee, Soojae). "MaxEnt 모형을 이용한 기후변화에 따른 산사태 발생가능성 예측." 환경영향평가 22.1 (2013): 39-50.

관련 논문

01. 논문 목록

9. 이승우(Lee, Seung-Woo), 김기홍(Kim, Gi-Hong), 윤차영(Yune, Chan-Young), 유한중(Ryu, Han-Joong), and 홍성재(Hong, Seong-Jae). "데이터베이스 구축을 통한 산사태 위험도 예측식 개발." 한국지반공학회논문집 28.4 (2012): 23-39.

10. 김기홍(Kim, Gi Hong), 윤찬영(Yune, Chan Young), 이환길(Lee, Hwan Gil), and 황제선(Hwang, Jae Seon). "GIS를 이용한 인제 산사태발생지역의 토석류 분석." 한국측량학회지 29.1 (2011): 47-53.

11. 윤홍식(Yun, Hong Sik),이동하(Lee, Dong Ha), and 서용철(Suh, Yong Cheol). "GIS 기법 및 발생자료 분석을 이용한 산사태 위험지도 작성." 한국지리정보학회지 12.4 (2009): 59-73.

12. 조명희(Jo, Myung Hee), and 조윤원(Jo, Yun Won). "기상과 지형자료를 통합한 산사태 위험지 예측 기법 개발 -울진지역을 대상으로." 한국지리정보학회지 12.2 (2009): 1-10.

13. 마호섭(Ma, Ho Seop), and 정원옥(Jeong, Won Ok). "우리나라 국립공원지역의 산사태 발생특성 분석." 한국산림과학회지 96.6 (2007): 611-619.

14. 이진덕(Lee, Jin Deok), 연상호(Yeon, Sang Ho), 김성길(Kim, Sung Kll), and 이호찬(Lee, Ho Chan). "산사태의 발생가능지 예측을 위한 GIS의 적용." 한국지리정보학회지 5.1 (2002): 38-47.

15. Cruden, David M., and David J. Varnes. “Landslide Types and Processes.” Special Report - National Research Council, Transportation Research Board 247 (1996): 36–75. Print.

동료 평가

구분	이원권	김홍범	곽희원	최디도
논문 조사 및 배경 지식 조사	25%	25%	25%	25%
데이터 전처리	25%	25%	25%	25%
모델링	25%	25%	25%	25%
결과 정리 및 분석	25%	25%	25%	25%
아이디어 제시	25%	25%	25%	25%

Thank you

감사합니다.