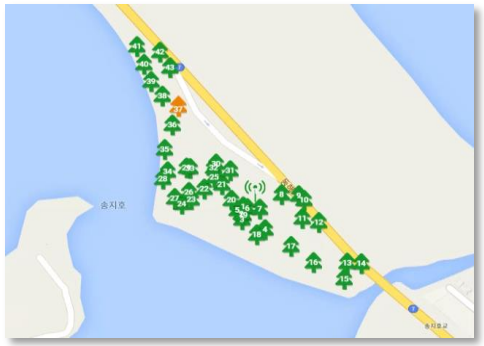
○○아카데미(훈련기관명)

이상목 탐지 분석

목차

01 프로젝트 개요

프로젝트 주제 선정 배경



고성시 내의 센서가 설치된 43그루의 나무 중 전염병이 걸린 나무 검출

상,중,하단,토양 4곳의 센서 수분량 활용

- 수목 전염병 중 하나인 재선충이 나무의 수분 공급로를 차단하고 수목의 생장을 방해하는 문제 발생
- 나무에 센서를 설치하여 수분량 데이터를 수집하고
- 수집된 데이터의 분석을 통해 재선충을 조기에 발견해서 전염병 확산을 예방할 수 있음
- 교육 과정에서 다루는 데이터는 대부분 제한된 범위의 정제된 데이터로 실제 업무 환경에서 발생하는 복잡한 문제 상황을 반영하는데 한계가 있음
- 실제 수목 관리 현장에서 사용하고 있는 데이터 분석을 통해 현장감 있는 데이터 규모와 난이도를 경험하고 학습한 내용들을 적용하고 활용하는 훈련 가능

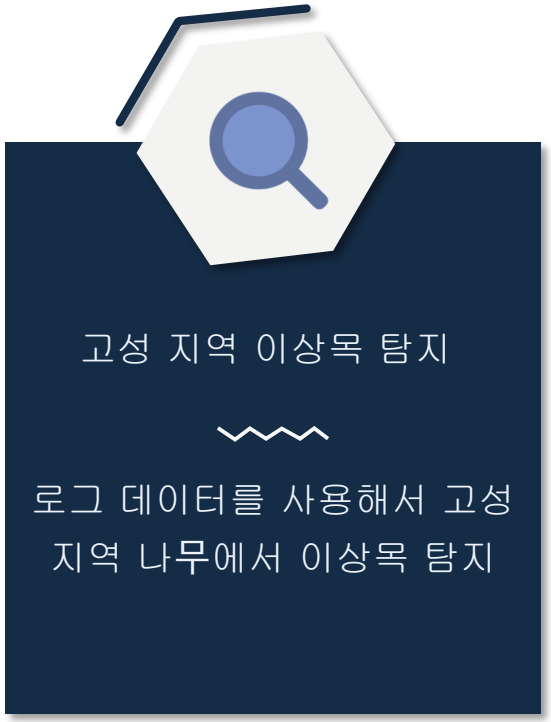
01 프로젝트 개요

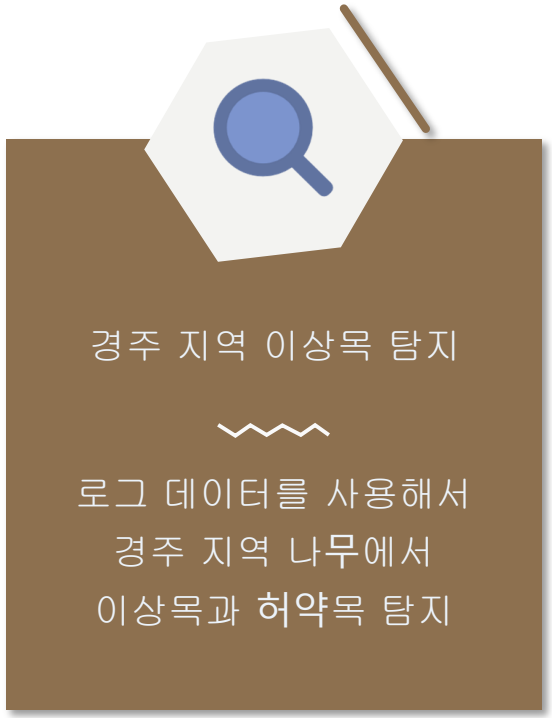
프로젝트 요구 사항

- 강원도 고성 지역의 43그루 나무에 부착된 센서로부터 발생하는 로그 데이터 사용
- 탐색적 분석 및 다양한 통계적 검정 기법을 활용해서 43그루의 나무 중 1개의 이상목 탐지
- 경상북도 경주 지역의 11그루의 나무에 부착된 센서로부터 발생하는 로그 데이터 사용
- 다양한 데이터 분석 기법을 통해 11그루의 나무 중 이상목 3개, 허약목 2개 정상목 6개를 탐지
- 센서로부터 발생하는 로그 데이터에는 결측 데이터가 포함되어 있으므로 분석이 가능하도록 결측 데이터 처리
- 분석의 완성도를 높일 수 있는 다양한 데이터 전처리 기법을 적용해서 데이터 정제

구현 기능

요구사항을 기반으로 프로젝트 수행을 통해 다음과 같은 문제를 해결한다





01 프로젝트 개요

개발 환경

1 **DBMS**

MySQL 8.0.22

MySQLWorkbench 8.0.22

Exerd 3.3.5

2 **Development Environment**

Miniconda 4.9.2

VisualStudioCode 1.5.2

3 **형상관리도구**

Git 2.31.1

Github

4 **협업도구**

Gitmind

Notion

5 **클라우드 컴퓨팅 환경**

AWS EC2, ECS, S3, RDS

적용 기술

데이터 적재 및 추출
- PyMySQL1.0.2

데이터 전처리
- Pandas1.2.3
- Numpy1.19.5

데이터 수집
- Selenium3.141.0,
- scipy1.6.2,
- Requests2.25.1 ,
- Beautifulsoup4 4.6.0,
- lxml4.6.3

형상관리도구
- Git, Github

생각정리도구
- Gitmind

활용가능언어
- SQL, Python, HTML

AWS
- EC2
- ECS
- RDS

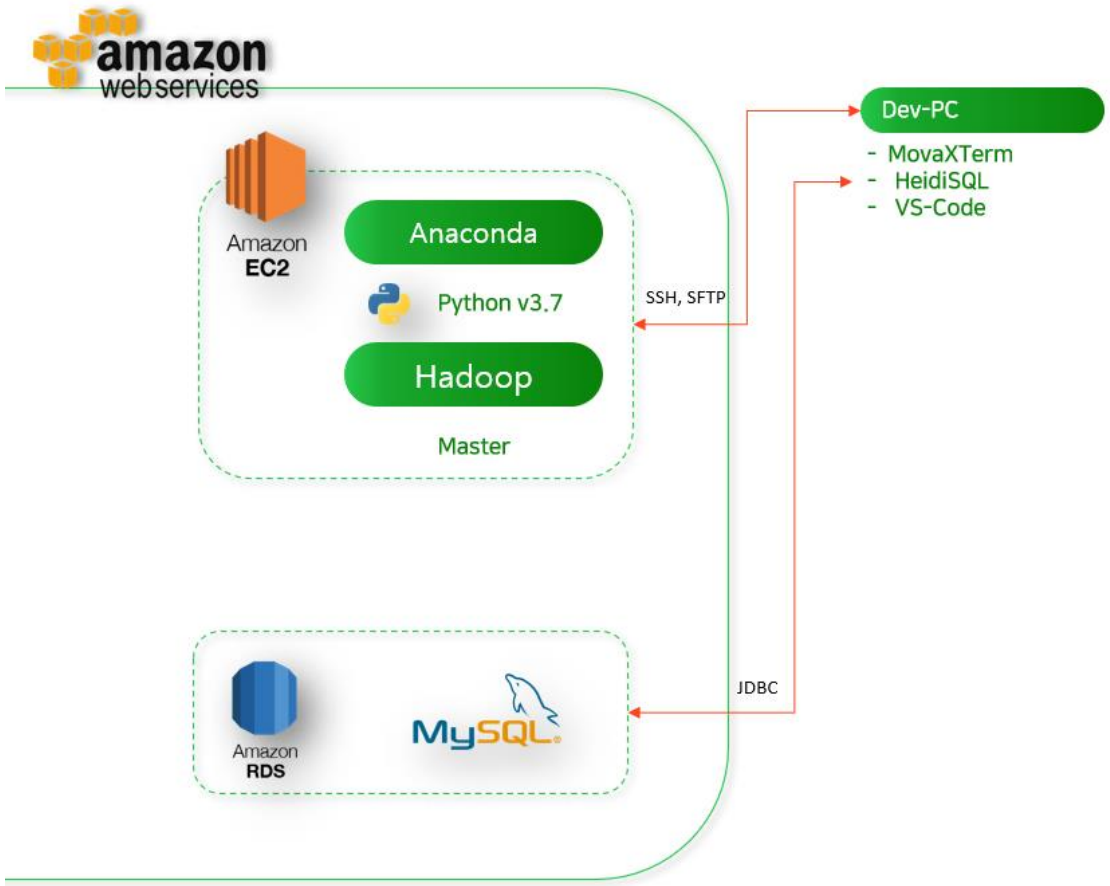
ML/DL
- Scikit learn0.24.1
- Tensorflow2.4.1
- Keras2.4.3

통계
- Satus
- Statsmodels
- Scipy

시각화
- Seaborn
- matpolotlib

01 프로젝트 개요

시스템 아키텍처



- 모든 시스템 리소스는 아마존 클라우드 서비스 환경에 구축
- 파이썬은 Anaconda 가상 파이썬 환경 사용
- 원격 서버 접속은 Visual Studio Code의 Remote Development 플러그인으로 접속해서 사용
- 데이터베이스는 아마존 RDS 인스턴스에 MySQL로 구성

02 프로젝트 팀 구성 및 역할

팀 구성원별 역할

훈련생	역할
AAA	▪ 웹 크롤링 데이터 수집 ▪ 데이터 전처리 ▪ 데이터 분석 ▪ 예측 모델링
BBB	▪ 업무 분석 ▪ 웹 크롤링 데이터 수집 ▪ 데이터 전처리 ▪ 데이터 분석 ▪ 예측 모델링
CCC	▪ 웹 크롤링 데이터 수집 ▪ 데이터 전처리 ▪ 데이터 분석 ▪ 예측 모델링

02 프로젝트 팀 구성 및 역할

팀 구성원별 역할

훈련생	역할
DDD	▪ 웹 크롤링 데이터 수집 ▪ 데이터 전처리 ▪ 데이터 분석 ▪ 예측 모델링

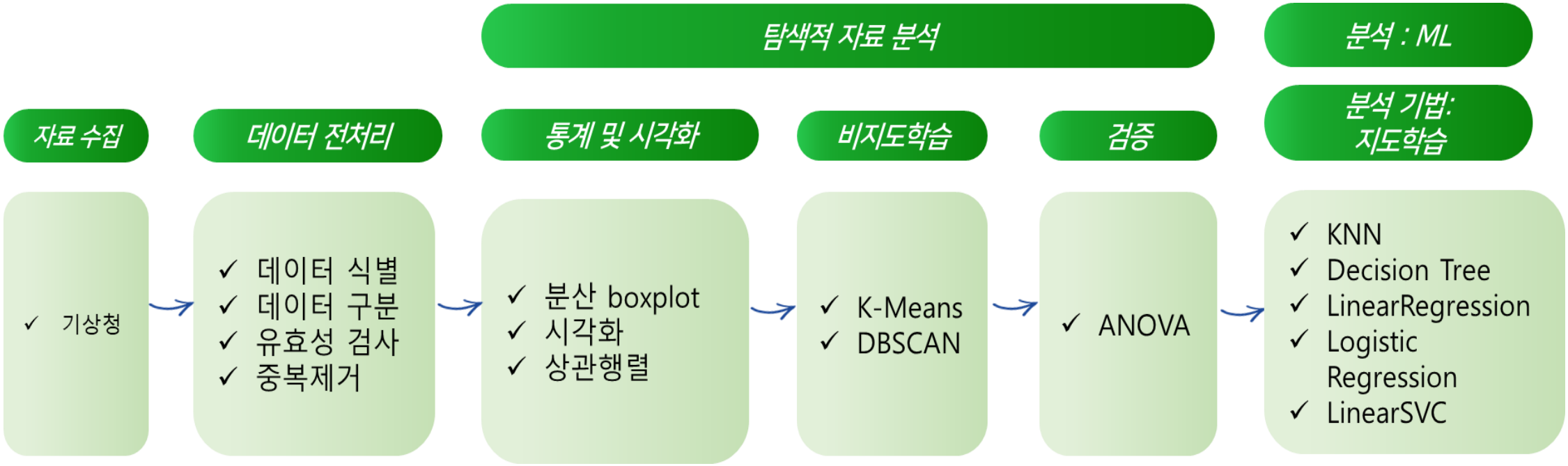
03 프로젝트 수행 절차 및 방법

프로젝트 수행 일정

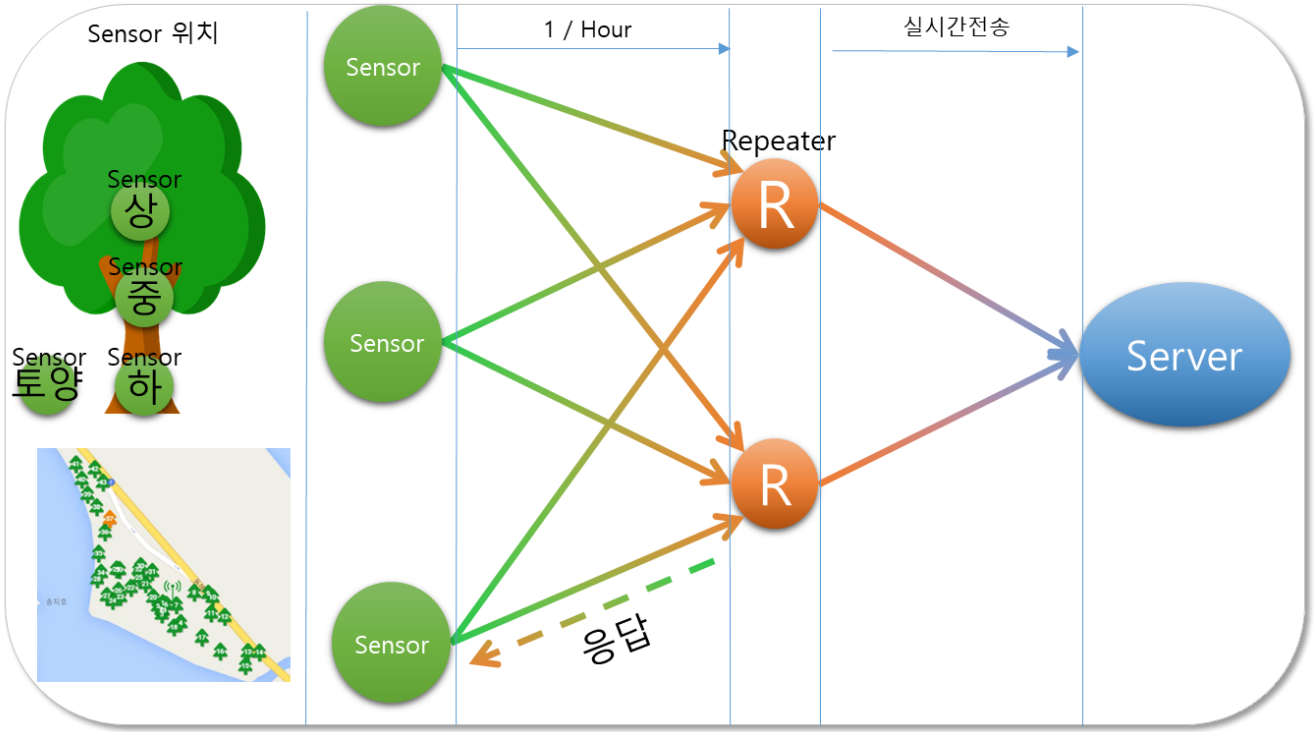
		구분	담당	과제내용	4월											
					19	20	21	22	23	24	25	26	27	28	29	30
					월	화	수	목	금	토	일	월	화	수	목	금
2차 프로젝트	1차 고성 (3일)	업무분석		데이터구조 및 수집환경												
		데이터수집		기상청자료_크롤링												
				Step01.센서자료 DB Upload												
				기상청자료_API												
				기상청자료_CSV DB Upload												
		전처리		Step02-03.데이터구분 (Parsing) 및 전처리검증(SQL)												
				Step04-05.유효성검사(정규표현식&Nan처리) 및 전처리검증(SQL)												
				Step06.중복데이터제거 및 전처리검증(SQL)												
		데이터분석		1차_데이터분석												
				1차_데이터분석												
				1차_데이터분석												
		머신러닝		KNN												
	2차 고성 (1개월)	업무분석		데이터구조 및 과제분석												
		전처리		2차_데이터 전처리												
		데이터분석		2차_데이터분석												
				2차_데이터분석												
				2차_데이터분석												
		머신러닝		LinearRegression												
	3차 경주 (2년)	전처리		3차_데이터 전처리												
		데이터분석		2차_데이터분석												
				2차_데이터분석												
				2차_데이터분석												
		머신러닝		DecisionTreeClassifier, K-Means												
				LogisticRegression, DBSCAN												
				LinearSVC												
		발표준비		PT 준비												

04 프로젝트 수행 결과

프로젝트 수행 절차



활용 데이터 정의



수목 센서 로그 데이터 사용

- 고성
  - ✓ 송지호 일대 43 그루
  - ✓ 2021.02 ~ 2021.04
- 경주
  - ✓ 양남면 일대 11 그루
  - ✓ 2018.08 ~ 2020.12
- 1시간에 한 번씩 센서로부터 데이터가 전송됨
- 전송이 실패하면 3번까지 재시도를 통해 데이터 전달
- 이후에는 데이터 폐기

활용 데이터 정의

```
2021-04-17 00:55:44 msg : [KGSS],[210223],[0],Data,[3967],155456,36,7,12376,36160,[2 FE54BA34,075830,-99,-99,3557,3646,3553,4085,5.6*2
```

데이터 형식

- TCP 네트워크 프로토콜 기반으로 전송되는 로그 데이터
- 전체 데이터 구성 항목 중 분석에는 뒷부분 9 개의 속성 사용

data	의미
2021-03-06 00:01:05	서버전송시간
[KGSS], [KGSM],[KGSG]	중계기이름
[210223]	펌웨어버전
[0]	중계기 번호
Data, RUN	중계기가 Alive상태
[2379]	중계기전송데이터 seq번호
145530	중계기시간
34	중계기 내 온도
13	중계기 내 습도
12372	인입전압
36180	사용전류
[2]	수신안테나
FE54BAC0	센서ID
065752	센서시간
-99	센서측정온도
-99	센서측정습도
3592	상단수분치
3553	중단수분치
3554	하단수분치
4095	토양수분치
5.6	센서전압
*0	

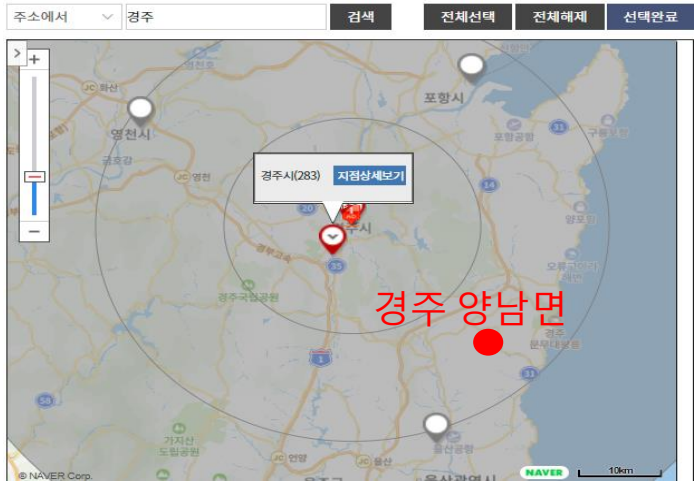
04 프로젝트 수행 결과

활용 데이터 정의

- 강수
- 습도
- 온도
- 일조/일사

기상 관측 데이터 (ASOS) 사용

- 주변 환경 정보를 반영하기 위해 센서의 측정 위치 인근 관측소의 기상 관측 데이터 사용
- 웹 크롤링을 통해 데이터 수집



활용 데이터 정의

강수 | 온도

습도 | 일조/일사

단기예보 - 파일셋

> 동네예보/단기예보, 양남면, 3시간기온

> 동네예보/단기예보, 양남면, 하늘상태

> 동네예보/단기예보, 양남면, 6시간강수량

> 동네예보/단기예보, 양남면, 습도

동네 예보 데이터 사용

- 주변 환경 정보를 반영하기 위해 센서의 측정 위치 인근 관측소의 단기 예보 데이터 사용
- 웹 크롤링을 통해 데이터 수집

04 프로젝트 수행 결과

데이터 전처리

- 데이터 특성, 정규표현식 등을 사용해서 수목 센서 로그 데이터의 결측치를 제거하는 전처리 수행

데이터 결측 검사(로그 텍스트 " , " 구분자 체크)

id	trans_dt	seq	log_msg
0	2021-03-06 00:00:00	0	2021-03-06 00:00:02 msg : [KGSG],[210218],[2],Data,[3732],145420,22,19,12376,36350,[3]FE5BDC62,2,2939,-99,-99,0000,..
1	2021-03-06 00:00:00	1	2021-03-06 00:00:06 msg : [KGSS],[210223],[0],RUN,[2376],145432,34,12,12372,36230

	seq	log_msg	cnt
:00:00	10	2021-03-06 00:01:05 msg : [KGSS],[210223],[0...	18
:00:00	45	2021-03-06 00:04:25 msg : [KGSM],[210218],[...	18

✓ 구분자 개수 체크

컬럼별 Pattern 검사

✓ 정규표현식을 통한 컬럼별 패턴체크

```
##################################################
# Pattern check
##################################################
for column_name in list_columns :
    if column_name == "trans_dtm_rep_id" :
        pattern = "^[0-9]{4}-[0-9]{2}-[0-9]{2} [0-9]{2}:[0-9]{2}:[0-9]{2} msg : \[[A-Z]{4}\]$"
    elif column_name == "rep_data_type" :
        pattern = "^Data$"
    elif column_name == "rep_data_seq" :
        pattern = "^\[[0-9]+\]$"
    elif column_name == "rep_tm" :
        pattern = "^[0-9]{6}$"
    elif column_name in ["rep_tmpr", "rep_humi"] :
        pattern = "^[-0-9]+$"
    elif column_name in ["rep_volt", "rep_elec"] :
        pattern = "^[0-9]+$"
    elif column_name == "rep_antno_ssr_id" :
        pattern = "^\[[0-9]\][A-Z0-9]+$"
    elif column_name == "ssr_tm" :
        pattern = "^[0-9]{6}$"
    elif column_name in ["ssr_tmpr", "ssr_humi"] :
        pattern = "^[-0-9]+$"
    elif column_name == ["ssr_high", "ssr_midd", "ssr_undr", "ssr_land"] :
        pattern = "^[0-9]$"
    elif column_name == "ssr_volt_chksum" :
        pattern = "^[0-9]{1,2}.[0-9]{1,2}\*[0-9]+$"
```

중복제거

데이터 중복 조건

- ✓ 센서ID = :sensor_id
- ✓ 센서시간 = :sensor_tm
- ✓ 전송일시 〉= :tr_dtm −1시간
- ✓ 전송일시 〈= :tr_dtm +1시간

✓ SQL Query를 통한 중복제거 검증

```
1   -- 중복 제거된 개수 확인 --
2 • use Forest;
3 • show tables;
4 • select count(1) as cnt from TBL_2ND_DAT_S01_PCK_OK_VAL
5   where dat_id not in (select dat_id from TBL_2ND_DAT_S01_PCK_OK_VAL_DUP);
6
```

Result Grid

	cnt
▶	67

04 프로젝트 수행 결과

데이터 전처리

- 기상 데이터의 결측치는 선형보간을 적용해서 의미 있는 값으로 대체
- 각각의 컬럼으로 구분된 날짜 및 시간 데이터를 하나의 컬럼 데이터로 통합하는 전처리 수행

관측 데이터 결측 처리

✓ 기상 데이터의 결측처리를 위해 선형보간을 통해 결측값을 대체

날짜	강수	온도	습도	…
2020.07.01 06:00	1.2	11	70.1	…
2020.07.01 07:00	Null	Null	65.8	…
2020.07.01 08:00	0	Null	54.3	…
2020.07.01 09:00	0	16	54.2	…

✓ 결측값 선형보간

날짜	강수	…	…	…
2020.07.01 06:00	1.2	11	70.1	…
2020.07.01 07:00	0.6	12.7	65.8	…
2020.07.01 08:00	0	14.3	54.3	…
2020.07.01 09:00	0	16	54.2	

※ 선형 보간법(線型補間法, linear interpolation)은 끝점의 값이 주어졌을 때
그 사이에 위치한 값을 추정하기 위하여 직선 거리에 따라 선형적으로 계산하는 방법.

동네예보 데이터 시간 처리

✓ 다운로드한 강수, 온도, 습도 등을 하나의 데이터로 병합
✓ 로그 데이터와 병합하기 위해 날짜, 시간을 로그의 시간대와 일치화
✓ 이후 1시간대로 변환하여 기상데이터와 마찬가지로 선형보간 처리.

day	hour	forecast	강수	…
1	200	4	1.2	
1	200	7	7.1	
1	200	10	0	
…	…	…	…	

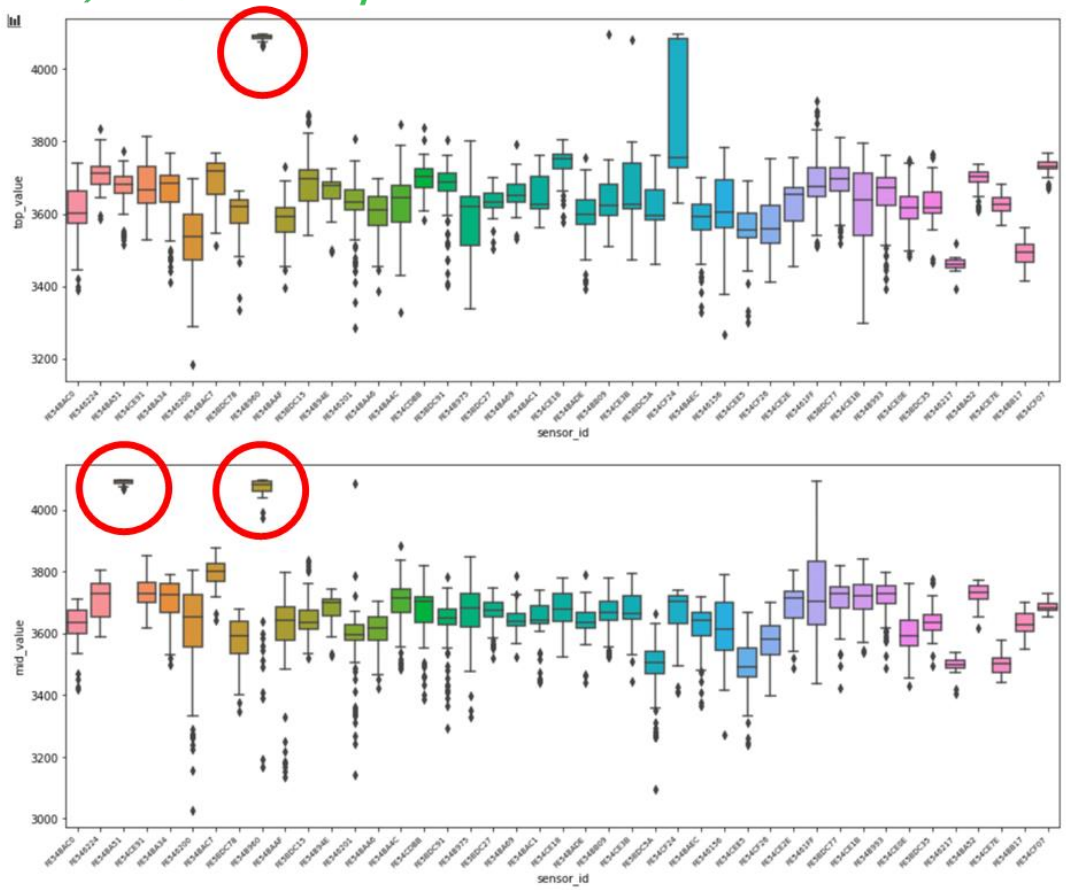
✓ 예보시간 동일 처리

날짜	강수	온도	…	습도	…
2020.07.01 06:00	1.2	11	…	70.1	…
2020.07.01 09:00	7.1	12	…	65.8	…
2020.07.01 13:00	0	17	…	54.3	…
…	…	…	…	…	…

04 프로젝트 수행 결과

탐색적 분석

1) 전체 센서 boxplot

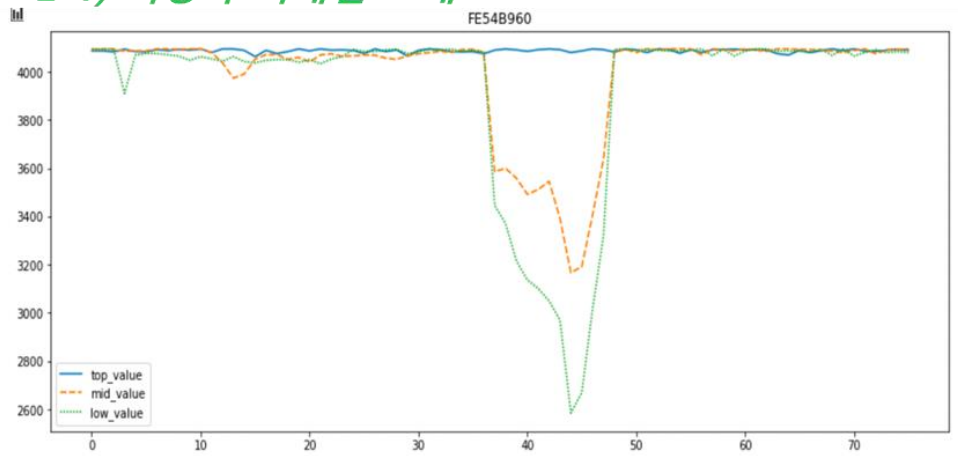


- 전체 나무의 수분 관련 데이터에 대해 boxplot 시각화 수행
- 각 수목 데이터의 전체적인 분포 파악
- 육안으로 관찰되는 이상 분포 데이터 확인

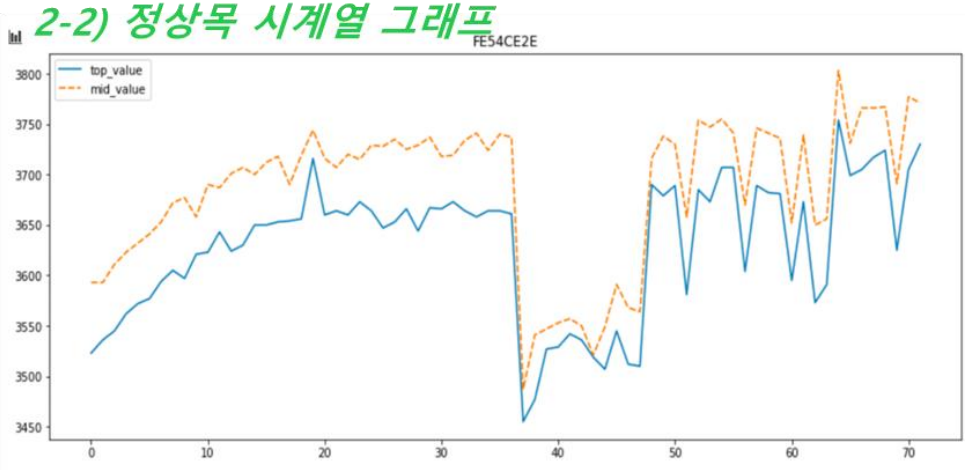
04 프로젝트 수행 결과

탐색적 분석

2-1) 이상목 시계열 그래프



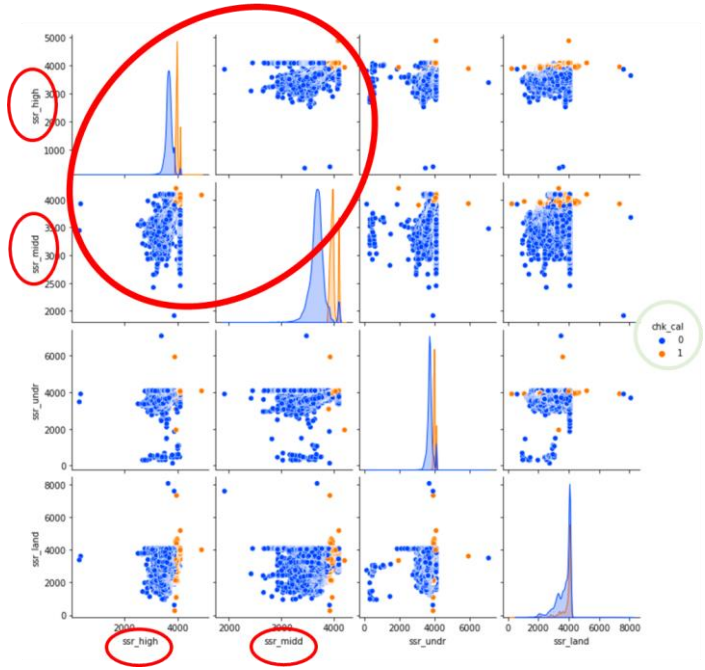
2-2) 정상목 시계열 그래프



- 시각화를 통해 추정한 이상목과 정상목에 대해 각각 시계열 선형 그래프로 특성 시각화
- 이상목에서 특이한 데이터 흐름이 나타나는 것 확인
- 수분 데이터의 측정값이 높고 수분 패턴이 고르지 않은 나무가 이상목
- 최종적으로 센서 아이디 FE54B960의 나무를 이상목으로 추정

04 프로젝트 수행 결과

탐색적 분석

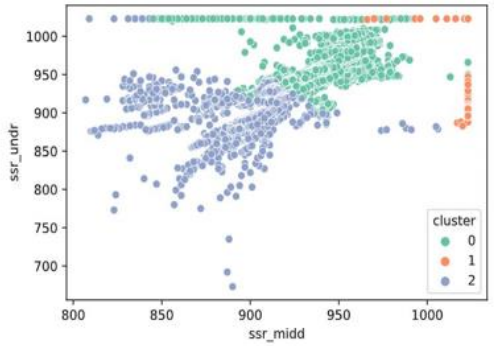


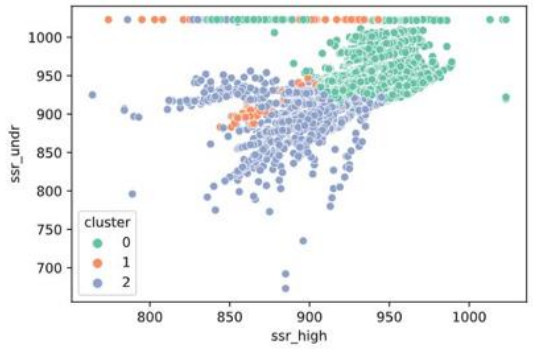
- 이상목 분류 후 산포도와 커널밀도추정 그래프로 시각화 시도
- 추가로 상관계수를 구하고 히트맵으로 시각화
- 중단 수분량과 상단 수분량 사이의 상관관계가 높은 것으로 파악됨
- 이 기준을 다음 단계의 이상목 검출 기준으로 활용 예정

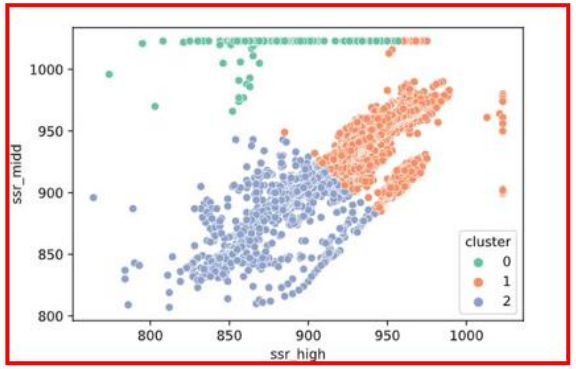
04 프로젝트 수행 결과

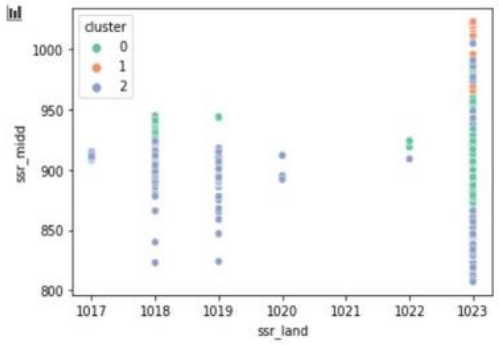
탐색적 분석

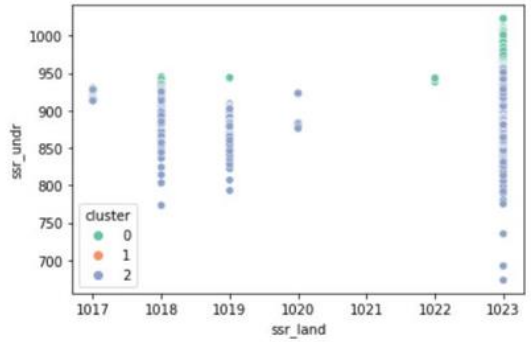
- 경주 지역 수목에서 이상목과 허약목을 찾기 위해
- 2년치의 데이터에 대해 이전 분석에서 중요한 기준이었던 중단 수분량과 상단 수분량으로 군집화 수행 (KMeans)

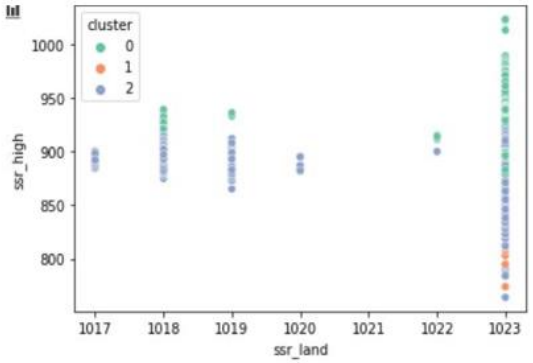






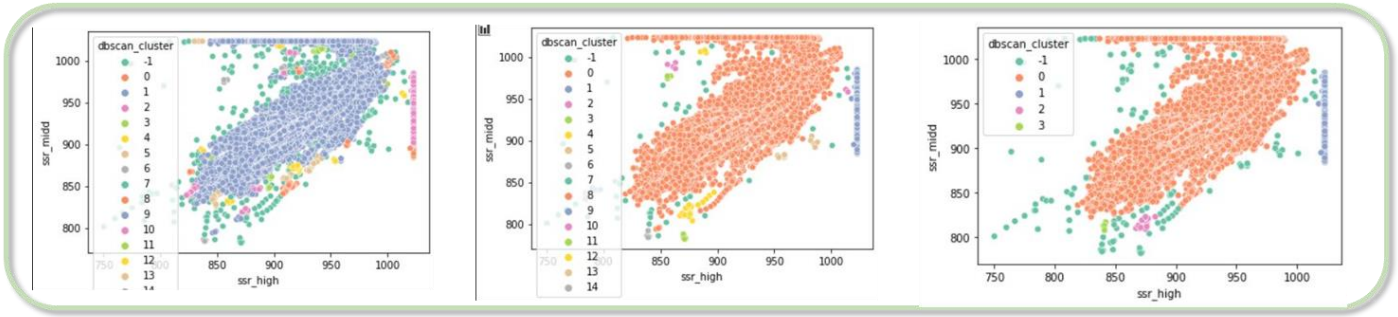




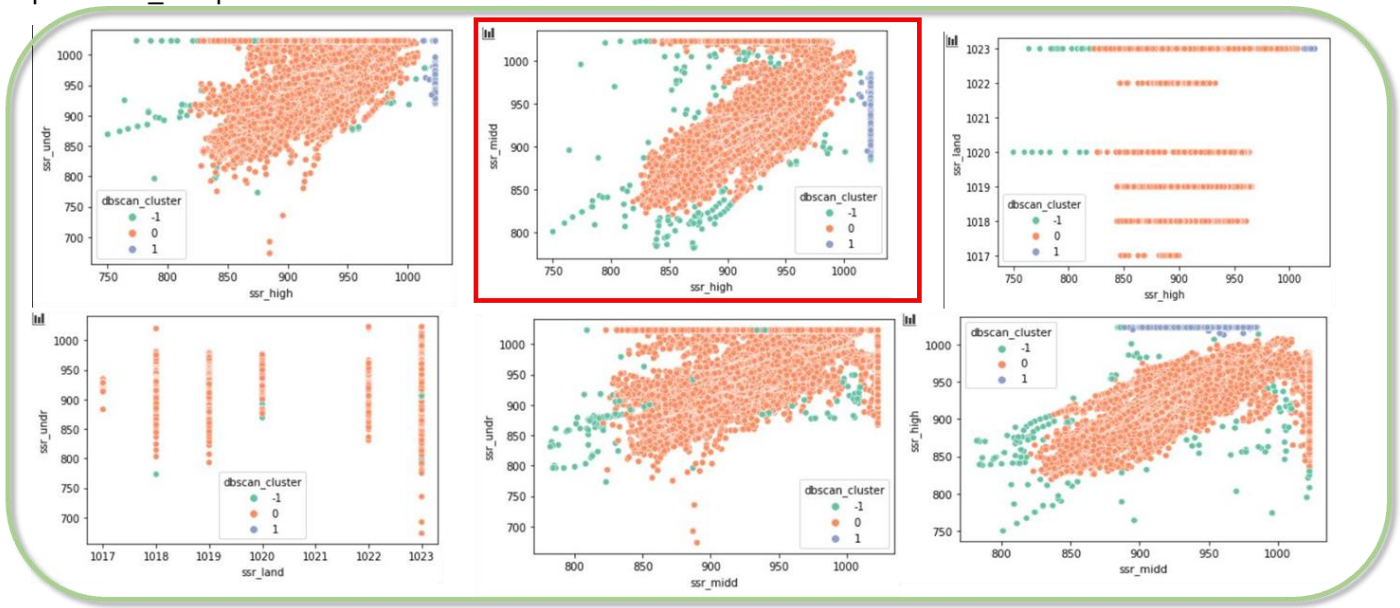


04 프로젝트 수행 결과

탐색적 분석



Eps=8 min_samples = 13

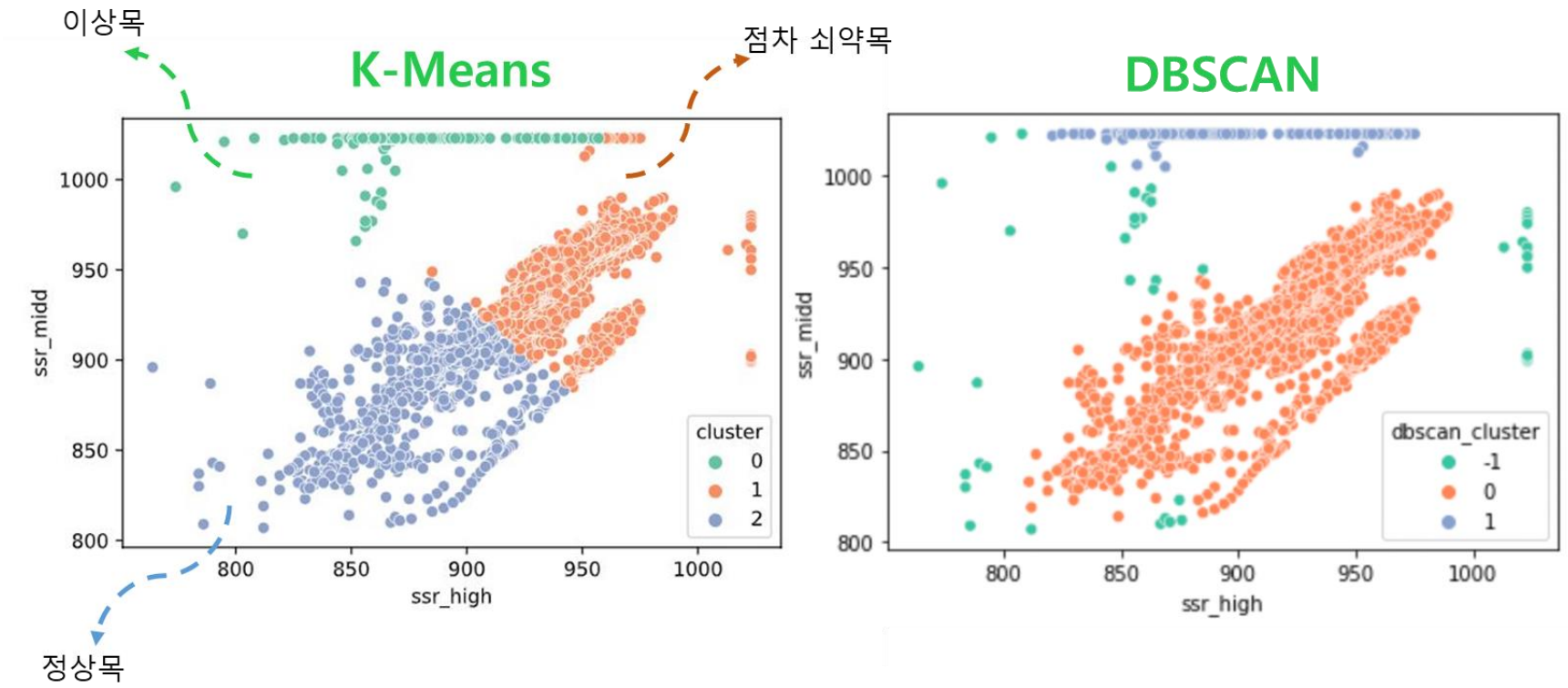


- 경주 지역 수목에서 이상목과 허약목을 찾기 위해
- 2년치의 데이터에 대해 이전 분석에서 중요한 기준이었던 중단 수분량과 상단 수분량으로 군집화 수행 (DBScan)

04 프로젝트 수행 결과

탐색적 분석

- 분석 결과 두 군집 알고리즘 중에서 수목 로그 데이터에는 Kmeans가 더 나은 군집 결과를 내는 것으로 확인됨



04 프로젝트 수행 결과

탐색적 분석

✓ ***정상목, 쇠약목, 이상목*** 세 그룹 간의 차이 검정

	Source	ddof1	ddof2	F	p-unc	np2
0	tree_status	2	17646.521477	7864.794696	0.0	0.223952

➔ p-value < 0.05 이므로 집단간의 차이가 있다.

✓ ***정상목 6그루*** 안에서 각 나무 간의 차이가 있는지

	Source	ddof1	ddof2	F	p-unc	np2
0	ssr_id	5	11202.810034	5637.899561	0.0	0.439184

➔ p-value < 0.05 이므로 집단간의 차이가 있다.

- 전체 나무를 대상으로 수분량에 대한 분산분석 결과 의미 있는 차이가 있는 것으로 확인됨
- 그러나 정상목 6그루를 대상으로 수분량에 대한 분산분석 결과도 의미 있는 차이가 있는 것으로 나타남
- 단순 수분량 분석으로 이상목 검출이 정상적으로 수행되기 어렵다는 사실 확인

탐색적 분석

- 도메인 전문가 (프로젝트 멘토)와 문제 상황 논의 후 데이터에서 이상목을 판별하는 기준에 대해 확인
- 제시된 이상목 판별 기준으로 각 데이터에 타겟 값을 할당

✓ 가설-1 : **(M-Val / H-Val * 100 < 85) and (Undr_Val / Midd_Val * 100 < 85)** 이면 이상목 아니면 정상목

✓ "85" 기준값 산출식

1. 측정값의 역수 (9000 - high, 9000 - midd, 9000 - undr)
2. 이상목의 투과율 구하기
3. 상/중/하 별 Min(rat)값 구하기 상 : 75, 중 : 93, 하 : 86
4. 상/중/하 Min(rat)의 평균 = 84.6 >> 반올림 85

04 프로젝트 수행 결과

예측 모델링

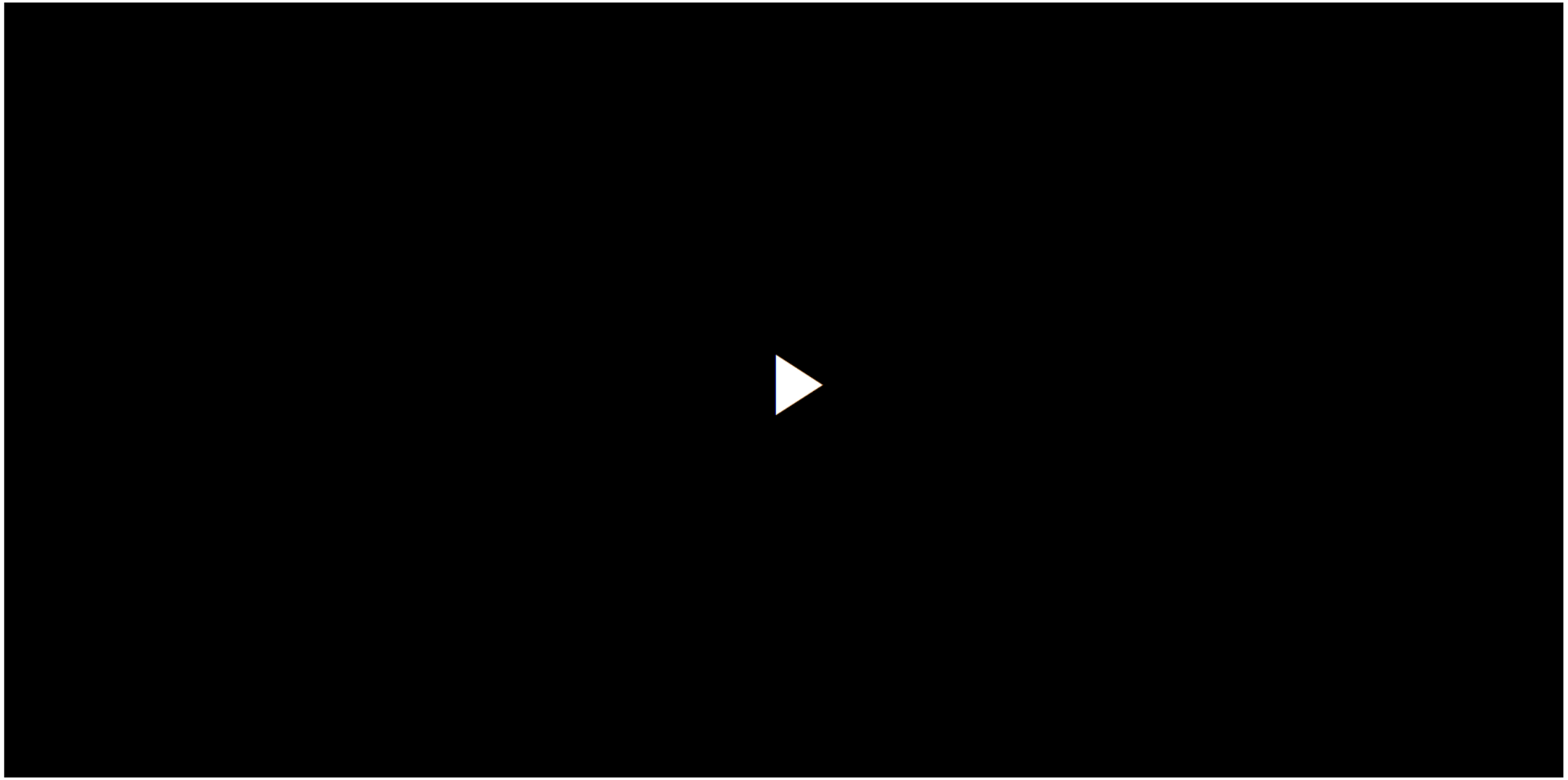
- 가설을 기반으로 할당된 타겟 값을 사용해서 머신러닝 알고리즘으로 학습하고 예측 정확도 도출

Model	score(X_train2, y_train)	score(X_test2, y_test)
KNN	0.9995972402199419	0.9995797436436226
DecisionTreeClassifier	1.0	1.0
LogisticRegression	1.0	0.9999474679554529
LinearSVC	0.9989493223128918	0.9990018911536037

04 프로젝트 수행 결과

결과 제시 5. 시연 동영상

※ 별도 첨부 가능



05 자체 평가의견

개선 사항 및 주요 고려 사항

- 이상치를 제거한 후의 데이터의 양이 충분하지 않아서 과적합 발생 가능성이 있을 것으로 예상됨
- 시계열 데이터에 적합한 예측 모델을 만들고 훈련 하는 방법에 대한 학습 필요
- 수분량의 정상 범위를 추정하기 위한 수분량 예측 모델을 먼저 개발하는 것도 고려해볼 필요가 있음
- 모델링도 중요하지만 정확한 데이터의 전처리가 매우 중요한 데이터 분석 과정임을 확인
- 현실 데이터를 기반으로 데이터 분석을 할 때 매우 다양한 변수를 고려해야 하며 특히 도메인에 대한 높은 수준의 이해가 필요