2024 날씨 빅데이터 콘테스트

기상청

수치모델 앙상블을 활용한 강수량 예측

순서

1 분석 주제

기습 폭우로 인한 문제 해결		
농업	재해	도로
병해충 초기 방제 미흡	집, 차량 등 침수 피해	포트홀 토사 도로 유입

가설 : 수치 모델 앙상블 강수 확률 자료를 활용해
누적 강수량의 계급 구간을 예측할 수 있다.

1 분석 환경 및 데이터 설명

개요

분석 환경

통합 개발 환경(IDE)

Visual Studio Code (VScode)

Python 버전

3.8.19

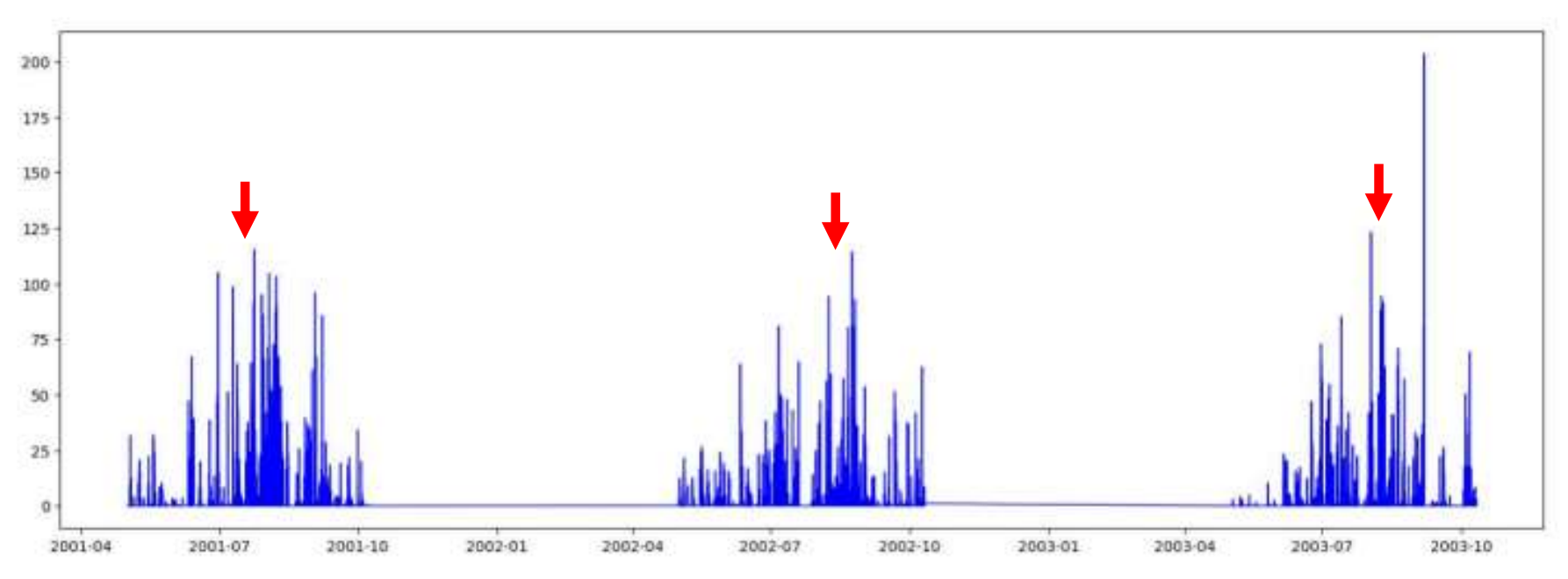
scikit-learn 버전

1.3.2

데이터 설명

변수	설명	변수	설명	변수	설명
TM_FC	기준 발표시각	V02	0.2 mm 이상 누적 확률	V07	10.0 mm 이상 누적 확률
TM_EF	예측 시간	V03	0.5 mm이상 누적 확률	V08	20.0 mm이상 누적 확률
DH	기준시각-예측 시간	V04	1.0 mm이상 누적 확률	V09	30.0 mm이상 누적 확률
STN	AWS 지점 코드	V05	2.0 mm이상 누적 확률	VV	실강수량
V01	0.1 mm 이상 누적 확률	V06	5.0 mm이상 누적 확률	class_interval	강수계급

예측시간 (ef_time) 에 따른 실강수량(vv) 확인



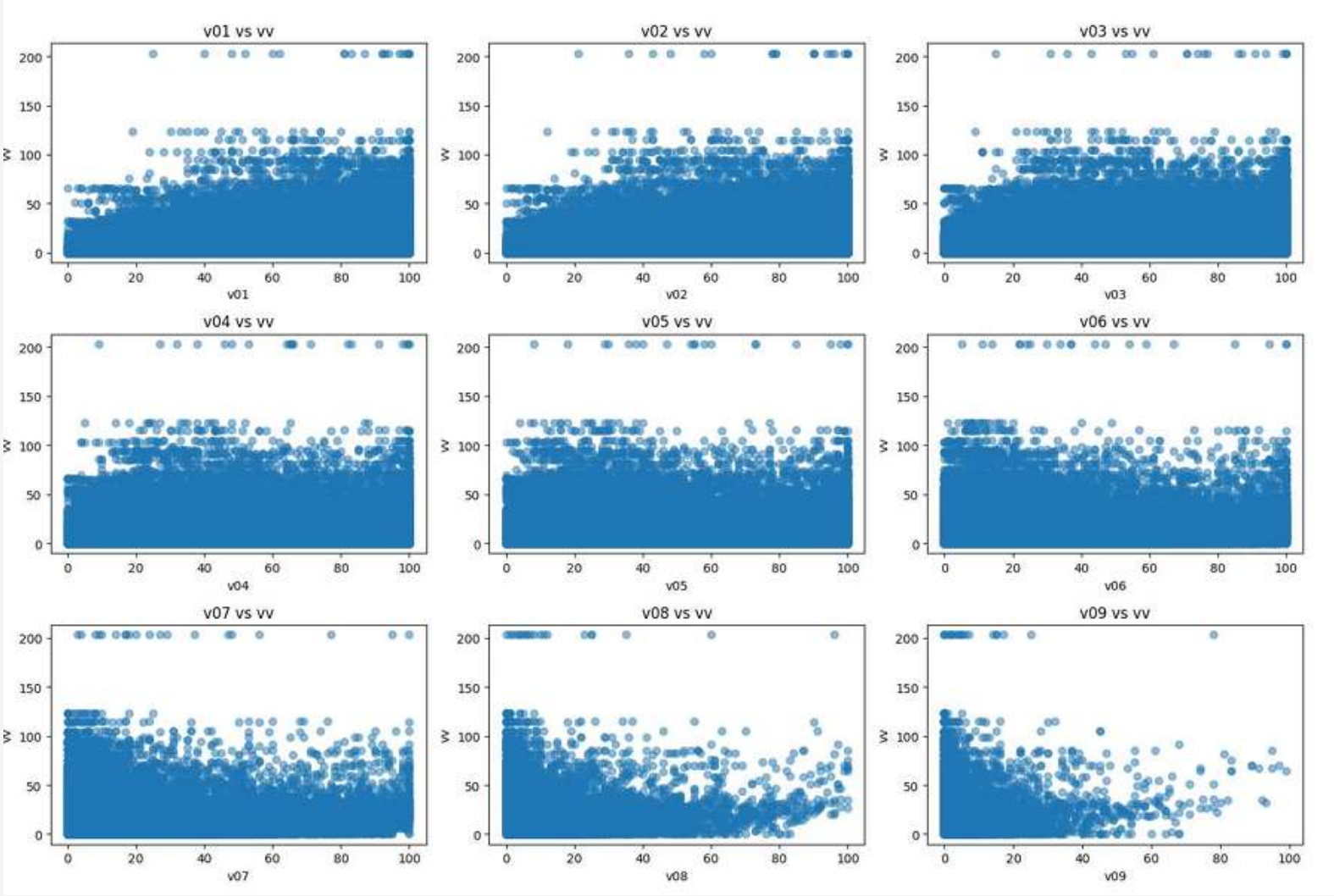
시간에 따른 주기성 → **장마철에 가장 많은 강수량 확인**

2 시각적 데이터 분석 (EDA)

각 변수와 실강수량(vv) 산점도

선형 관계가 없음

비선형 모델링



결측치, 이상치 처리

vv = -999 → 제거 후 학습 진행

vv = 203.2 12 STN (특정지역) 에서 발견 → 유지하고 학습 진행

변수 채택

fc_time (발표시각) → 변수 제외

ef_year, ef_month, ef_day, ef_hour → ef_datetime 변수 변환 후 채택

class_interval → 변수 제외, 학습 이후 vv에 맞춰 생성

변수명	변경 전	변경 후	비고
ef_year	A, B, C	2001, 2002, 2003	숫자형 변환
stn4contest	STN001, STN002 ...	1, 2 ...	숫자형 변환
season	v01 ~ v09	장마(1), 장마 아닌 기간(0)	변수 추가
mean, min, max	v01 ~ v09	그룹별 평균, 최댓값, 최솟값	변수 추가
recent		dh 가장 작은 v01 ~ v09	변수 추가
ef_hour_sin ef_hour_cos	ef_hour	sine / cosine 변환	변수 추가

```python
# 시간 관련 피처 생성 (Sine/Cosine 변환)
def add_sin_cos_features(df, col):
    df[col + '_sin'] = np.sin(2 * np.pi * df[col] / df[col].max())
    df[col + '_cos'] = np.cos(2 * np.pi * df[col] / df[col].max())
    return df
```

데이터 스케일링 - **MinMaxScaler** 채택

평가 지표

$$CSI = \frac{H}{H + F + M}$$

- **H** 강수로 예측하여 구간 예측에 성공한 값
- **F** 강수로 예측하여 구간 예측에 실패한 값
- **M** 무강수로 예측하여 구간 예측에 실패한 값
- **C** 무강수로 예측하여 구간 예측에 성공한 값 (평가에서 제외)

모델 성능 검증

3 학습 모델 선정

1 Random Forest Regressor

```
CSI: 0.07862532677313668
RandomForestRegressor(max_depth=8, max_features='sqrt')
```

2 Gradient Boosting Regressor

```
CSI: 0.0317422980978143l4
GradientBoostingRegressor(learning_rate=0.01)
```

3 MLP 신경망

```
CSI: 0.031785628067916634
XGBRegressor(base_score=None, booster=None, callbacks=None,
             colsample_bylevel=None, colsample_bynode=None,
             colsample_bytree=0.8, device=None, early_stopping_rounds=None,
             enable_categorical=False, eval_metric=None, feature_types=None,
             gamma=None, grow_policy=None, importance_type=None,
             interaction_constraints=None, learning_rate=0.01, max_bin=None,
             max_cat_threshold=None, max_cat_to_onehot=None,
             max_delta_step=None, max_depth=8, max_leaves=None,
             min_child_weight=None, missing=nan, monotone_constraints=None,
             multi_strategy=None, n_estimators=100, n_jobs=None,
             num_parallel_tree=None, random_state=None, ...)
```

+ 이외에도 Catboost, LightGBM 등 다양한 모델링 결과

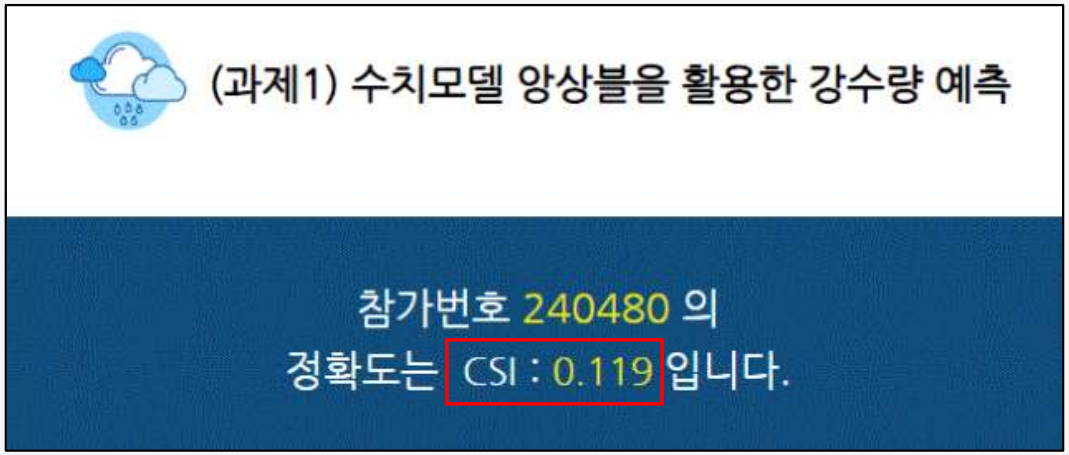
성능이 우수한

Random Forest Regressor 채택

4 CSI 결과 확인

분석 결과

```
MSE: 10.600161936801628
R2: 0.3957689028841904
CSI: 0.07862532677313668
RandomForestRegressor(max_depth=8, max_features='sqrt')
```



2개년 학습 후 예측 CSI **0.078**

3개년 학습 후 예측 CSI **0.119**

3개년이 아닌 더 많은 과거 데이터로 학습한다면
학습 검증과 예측 검증 사이의 간극을 좁힐 수 있을 것으로 기대

5 추가 검증 및 개선

회고

스케일링 부분에서 실수 발견

```
MSE: 10.600161936801628
R2: 0.3957689028841904
CSI: 0.07862532677313668
RandomForestRegressor(max_depth=8, max_features='sqrt')
```

```
MSE: 10.619806924115275
R2: 0.39464909808232784
CSI: 0.08304759690823502
RandomForestRegressor(max_depth=8, max_features='sqrt')
```

성능 개선 확인

지점 (stn4contest) 변수 별로 분리해서 학습

```
MSE: 10.619806924115275
R2: 0.39464909808232784
CSI: 0.08304759690823502
RandomForestRegressor(max_depth=8, max_features='sqrt')
```

```
MSE: 15.669329074328
R2: 0.37386875379164475
CSI: 0.10376687988628287
RandomForestRegressor(max_depth=8, max_features='sqrt')
```

성능 개선 확인

5 강수량 예측으로 인한 효과

기대 효과

CSI 점수 개선

장마 예보 개선	재해 예방 대응 강화	에너지 생산 사용 최적화	도로 교통 관리

감사합니다.