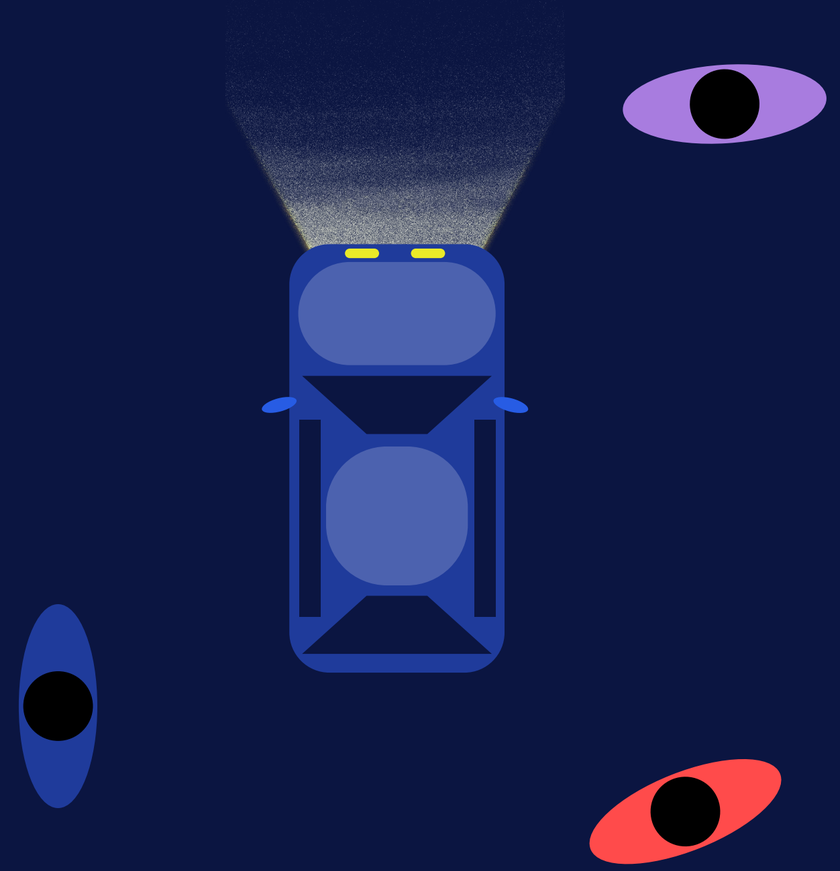
고속도로 사고 예측

2조

정다훈 강하빈 김세령 김지훈 조민서

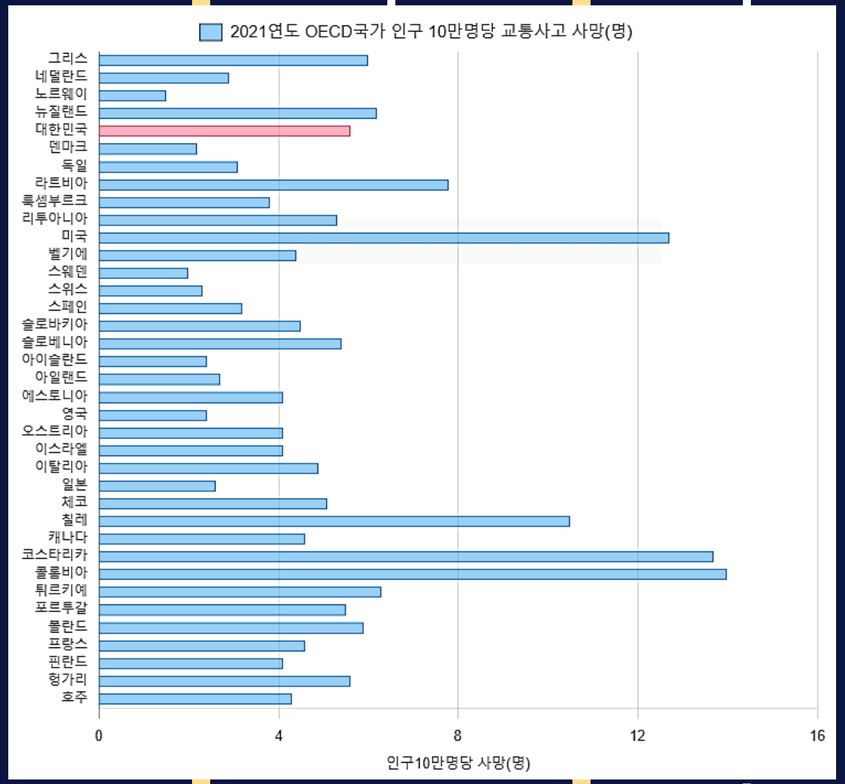
목차



01.프로젝트 개요

프로젝트 추진 배경

2023년 교통사고 198,296건,
사망자 2,551 명으로
2013년부터 11년 동안
사고가 감소하고 있는 추세



그러나

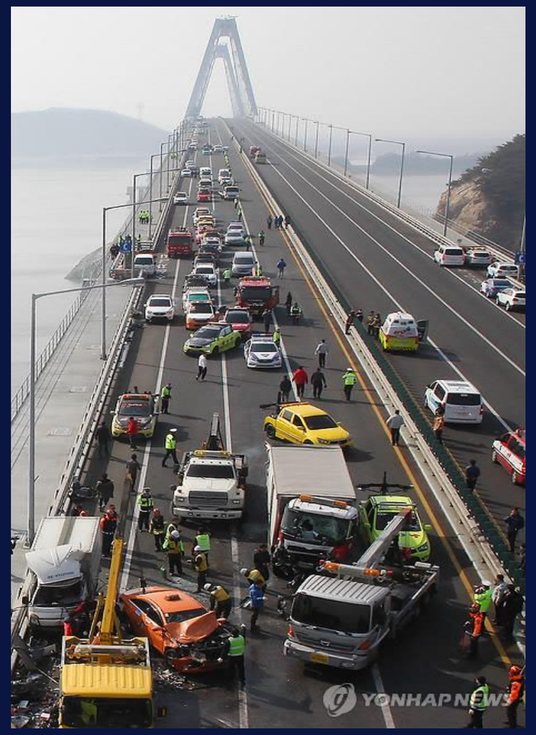
- 고속도로에서의 교통사고 건수는 여전히 증감을 반복 중임
- 치사율이 최고치를 나타냄
- 인구 10만명당 교통사고 건수 OECD 회원국 평균 27 건에 비해 우리나라는 74.2건으로 약 2.8배 높음

따라서

- 사고 심각도 영향 요인에 대한 분석이 필요함

01.프로젝트 개요

프로젝트 추진 필요성



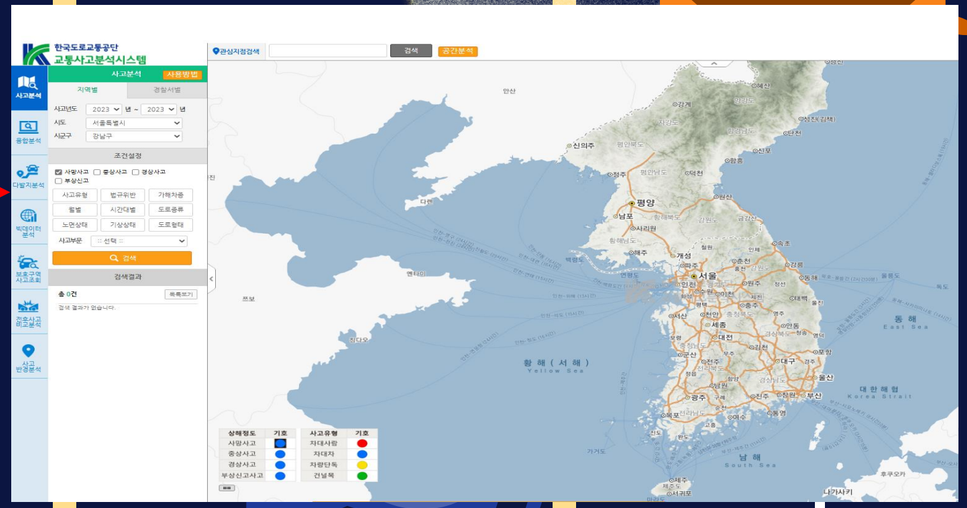
- 최근 급격한 기후변화로 인한 잦은 노면 상태의 면화, 기상 이변 등으로 인한 고속도로 내 사고 사례가 급증함

- 선행 연구로 AdaBoost, XGBoost, Random Forest 등의 기법을 사용한 교통사고 주요 원인 분석이 수행되었으나 연속적인 숫자형 데이터를 예측하는 회귀 모델과 사고 심각도와 같은 이산형 클래스를 예측하는 분류 모델을 복합적으로 적용한 연구가 미미함

02. 데이터 수집 및 전처리

데이터 수집

ASS 교통사고 분석 시스템에서 고속도로별 교통사고 데이터 수집



데이터 검색조건->
연도 : 2020~2022년도
도로 종류 : 고속국도
도로명 : 경부,광주대구,남해,서해안,영동,
중부,중부내륙,중앙,호남
사고내용 : 사망,중상,경상사고

02. 데이터 수집 및 전처리

데이터 전처리

데이터 합치기

- 도로종류별로 시트가 나누어져 있으므로 판다스를 통해 데이터 통합

널값 처리

- 사고 유형 - 단독 사고인 데이터는 피해운전자 관련 데이터가 비어 있어 해당컬럼에 대해서는 '기타'로 처리

독립변수 추가

- 사고일시 컬럼을 통해 시간대,계절,주간/야간 변수 추가
- 시군구 컬럼을 통해 지역 변수 추가 (ex. 경상북도 상면시 낙동면 -> 경상북도

종속변수 설정

- ARI (Accident Risk Index) 추가 계산식은 다음과 같다.

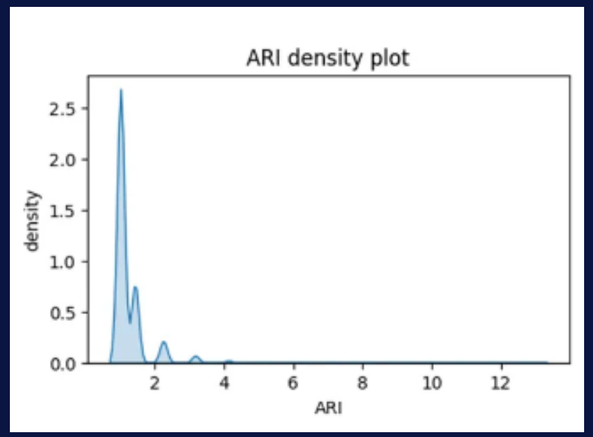
$$ARI = \sqrt{KSI^2 + 사고건수^2} \,/\, 년$$

- ARI 지수의 분포를 살펴보고 적절하게 '상', '중', '하'로 나눠서 위험도컬럼 추가

02. 데이터 수집 및 전처리

데이터 전처리

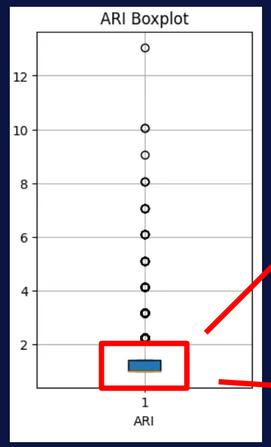
종속변수 설정

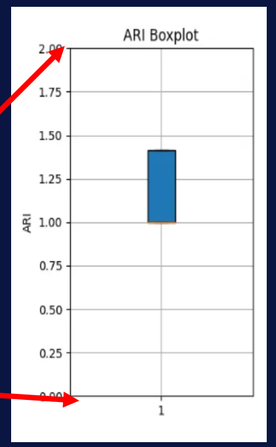


ARI	count
1.000000	4545
1.414214	1312
2.236068	364
3.162278	116
4.123106	35
5.099020	10
6.082763	7
7.071068	5
8.062258	3
10.049876	2
9.055385	1
13.038405	1

dtype: int64

ARI의 분포는 1~13사이로 나왔으며 값이 낮을수록 데이터가 많은 경향을 보임
이에 대한 위험도(상,중,하)를 구하기위한 박스플롯





02. 데이터 수집 및 전처리

데이터 전처리

독립변수 선택 과정

- 전체 독립변수(범주형) 리스트에서 카이제곱 검정을 통해 종속변수(범주형)인 위험도와 유의미한 변수들 추출

결과

variable	사고내용	피해운전자 상해정도	가해운전자 상해정도	피해운전자 차종	지역	사고유형	법규위반	시간
chi2_stat	6458.211659	3423.816462	1.181238e+03	3.484164e+02	2.313271e+02	1.950806e+02	1.640654e+02	1.574985e+02
p_value	0.000000	0.000000	1.604549e-247	9.442516e-62	7.022557e-34	4.289508e-41	1.078446e-27	4.034177e-14

피해운전자 성별	가해운전자 차종	노면상태	기상상태	시간대	요일	도로형태	계절	가해운전자 성별	주말여부
1.130312e+02	1.004443e+02	5.079760e+01	4.846871e+01	3.805739e+01	24.715434	24.394722	23.901122	14.091429	6.445243
9.061128e-21	4.556521e-16	2.870480e-08	5.095742e-07	5.444313e-09	0.016230	0.017966	0.000545	0.000871	0.039850

상관관계가 존재하는 변수(유의미한 변수)

- '위험도' 컬럼과 상관관계가 있는 독립변수

→ 카이제곱 검정의 p-value <0.05

- 유의미한 관계가 있는 변수

시간, 계절, 요일, 주말여부, 지역, 사고내용, 사고유형, 법규위반, 노면상태, 기상상태, 도로형태, 가해운전자 차종, 가해운전자 성별, 가해운전자 상해정도, 피해운전자 차종, 피해운전자 성별, 피해운전자 상해정도, 시간대

코드

```python
# 독립변수
X  = df.select_dtypes(include=['object']).columns
X = [col for col in X if col != '위험도']

# 종속 변수
y = df['위험도']

# 유의미한 변수 저장 리스트
significant_vars = []

# 각 독립 변수에 대해 카이제곱 검정 수행
for x in X:
    # 교차표 생성
    contingency_table = pd.crosstab(df[x], y)

    # 카이제곱 검정 수행
    chi2_stat, p_value, dof, expected = stats.chi2_contingency(contingency_table)

    # p-value가 유의수준보다 작으면 리스트에 추가
    alpha = 0.05
    print(f"var : {x}, P-value : {p_value}")
    if p_value < alpha:
        significant_vars.append(x)


# 결과 출력
print("유의미한 관계가 있는 변수들:")
print(significant_vars)
```

03. 모델링 및 평가

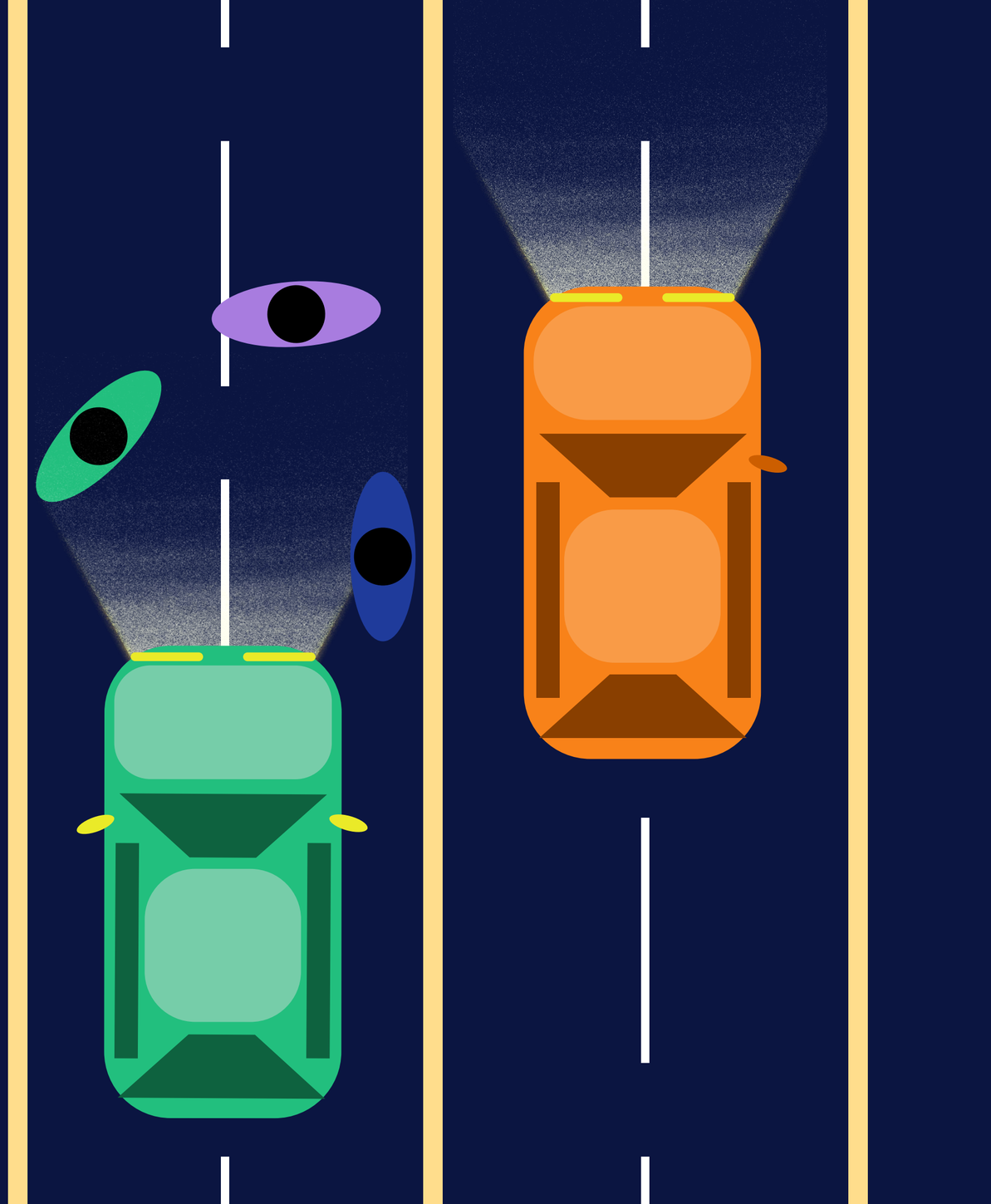
데이터 탐색 및 선택

종속 변수

- 위험도

독립 변수

- 시간, 지역, 사고내용, 법규위반, 노면상태, 기상상태, 가해운전자 차종, 가해운전자 성별, 가해운전자 상해정도, 피해운전자 차종, 피해운전자 성별, 피해운전자 상해정도

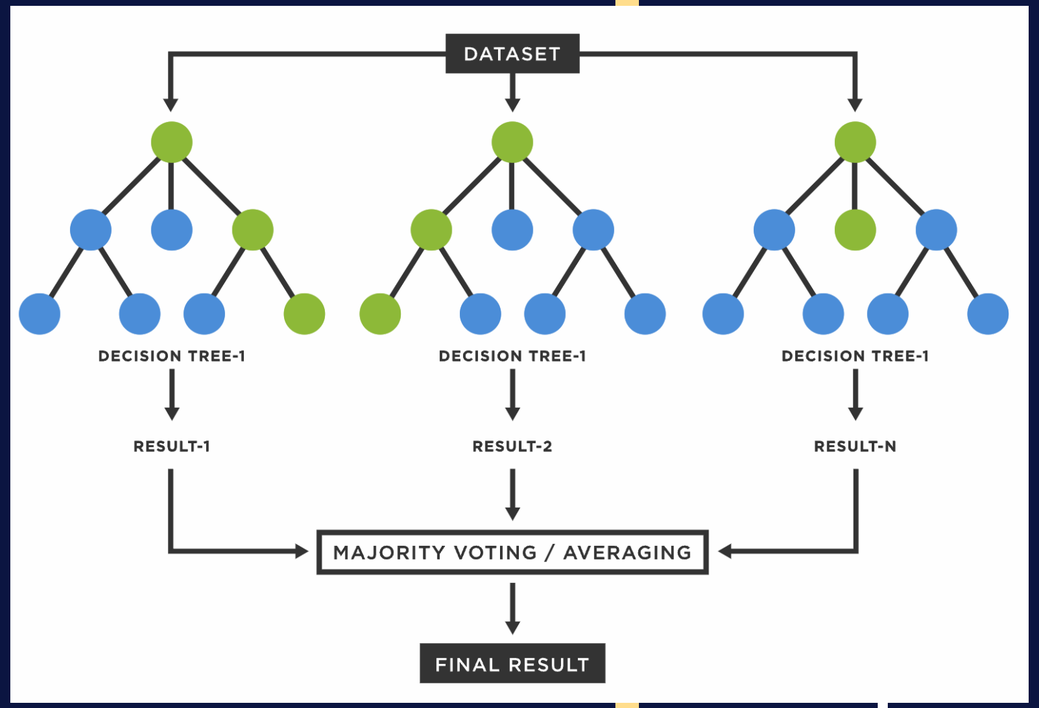


03. 모델링 및 평가

사용 모델

1. Random Forest

여러 개의 결정 트리(Decision Tree)를 결합하여
예측 성능을 향상시키는 앙상블 학습 기법
각 트리는 데이터의 일부를 랜덤하게 선택해
학습하며, 최종 예측은 각 트리의 예측을
다수결로 결정

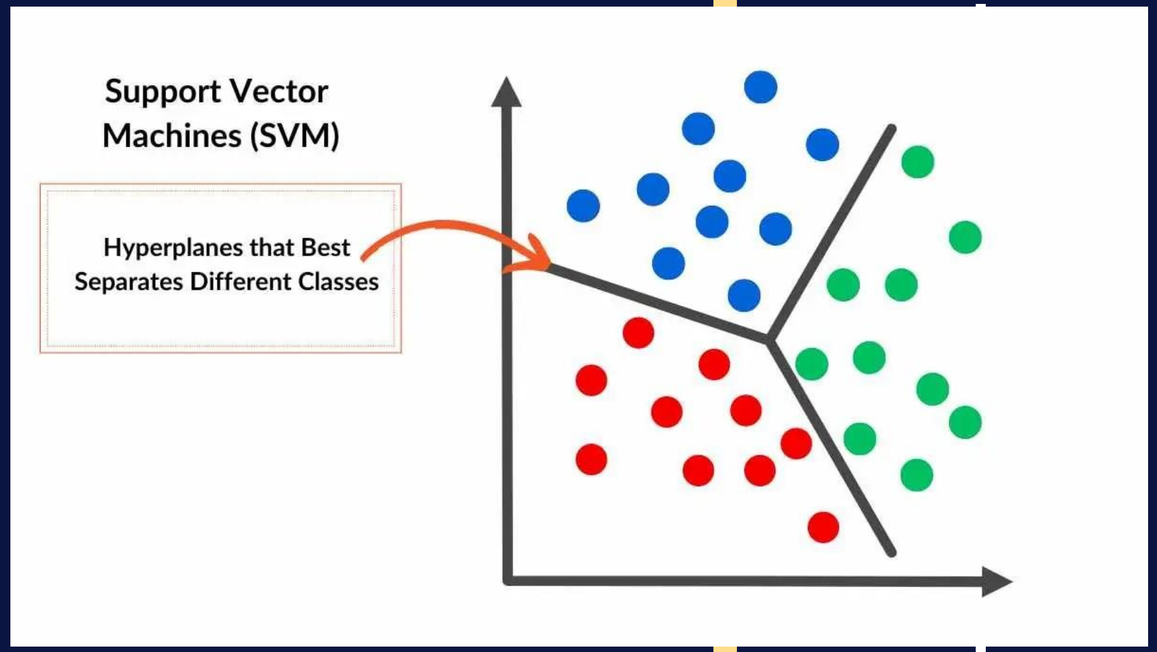


03. 모델링 및 평가

사용 모델

2. Support Vector Machine

주어진 데이터 포인트를 두 개의 클래스 사이에 최적의 경계(결정 경계)를 만드는 방식으로 분류하며, 이 경계는 가장 가까운 데이터 포인트(서포트 벡터)와의 거리를 최대화하도록 설정



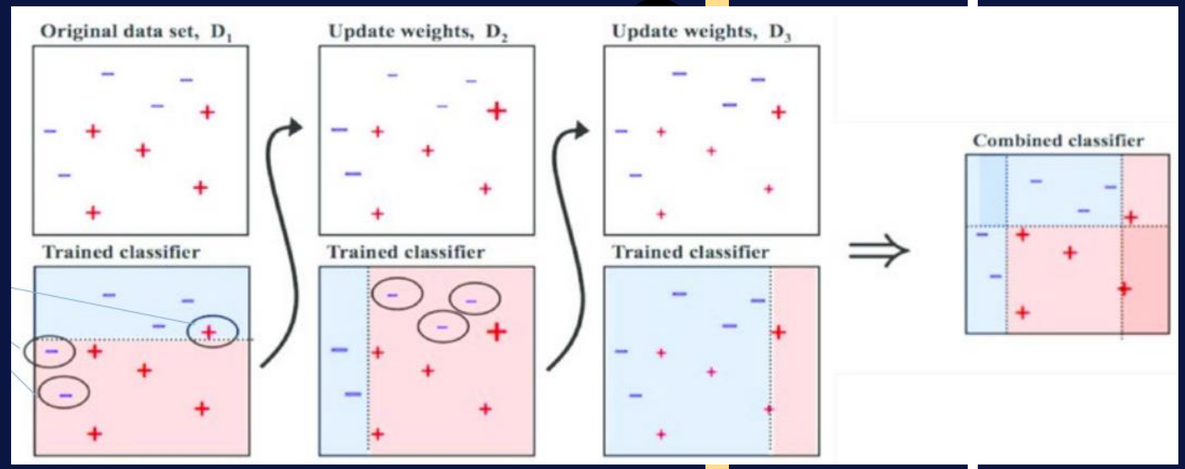
03. 모델링 및 평가

사용 모델

3. AdaBoost

여러 약한 학습자(weak learners)를 결합해 강력한 학습자를 만드는 부스팅 기법

초기 학습자가 잘못 예측한 데이터에 더 큰 가중치를 부여하며, 이후 학습자들이 이러한 오류를 수정하도록 반복적으로 학습



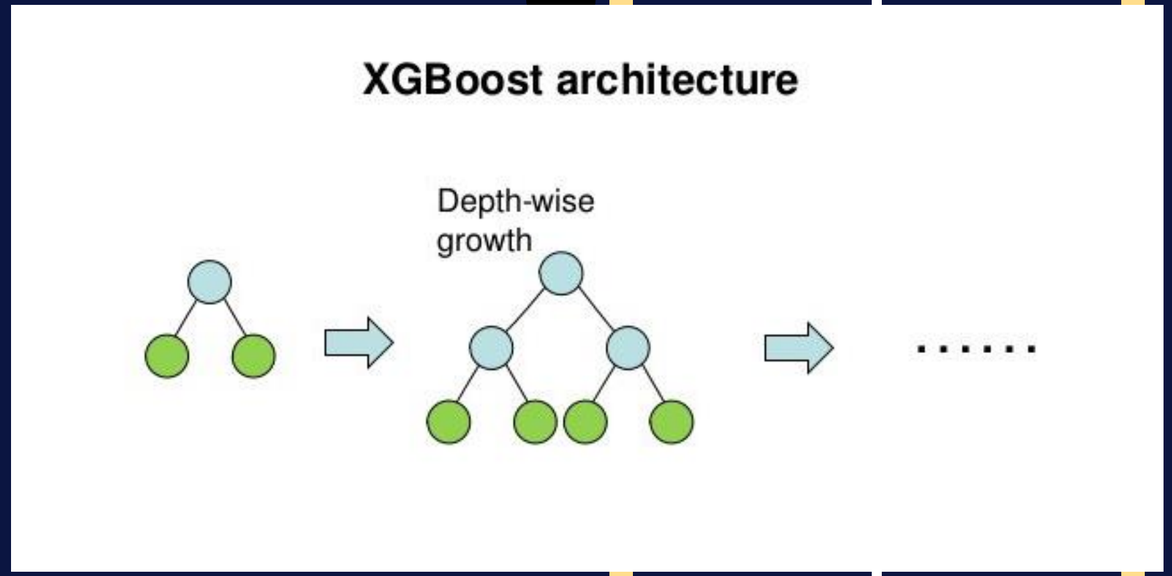
03. 모델링 및 평가

사용 모델

4. XGBoost

그래디언트 부스팅 원리에 기반하여, 이전 모델이 만든 오류를 점진적으로 보완하는 방식으로 동작

각 단계에서 잔여 오차(Residual Error)를 최소화하기 위해 새로운 결정 트리를 추가하며, 학습 속도와 효율성을 높이기 위해 병렬 처리와 특수한 트리 구조를 사용



03. 모델링 및 평가

사용 모델

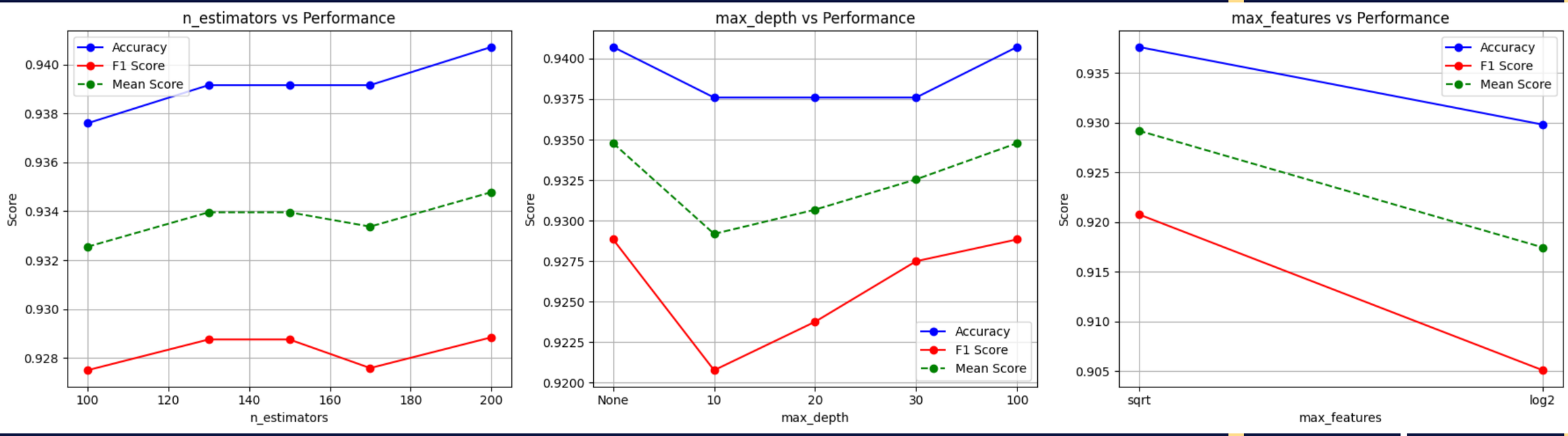
Accuracy, F1-score

	Random Forest	SVM	AdaBoost	XGBoost
Accuracy	0.940	0.920	0.925	0.929
F1-score	0.928	0.897	0.892	0.923

04. 하이퍼파라미터 수정

Random Forest

하이퍼파라미터 : n_estimators, max_depth, max_features



최고 Mean Score (n_estimators): 0.9348 at n_estimators=200

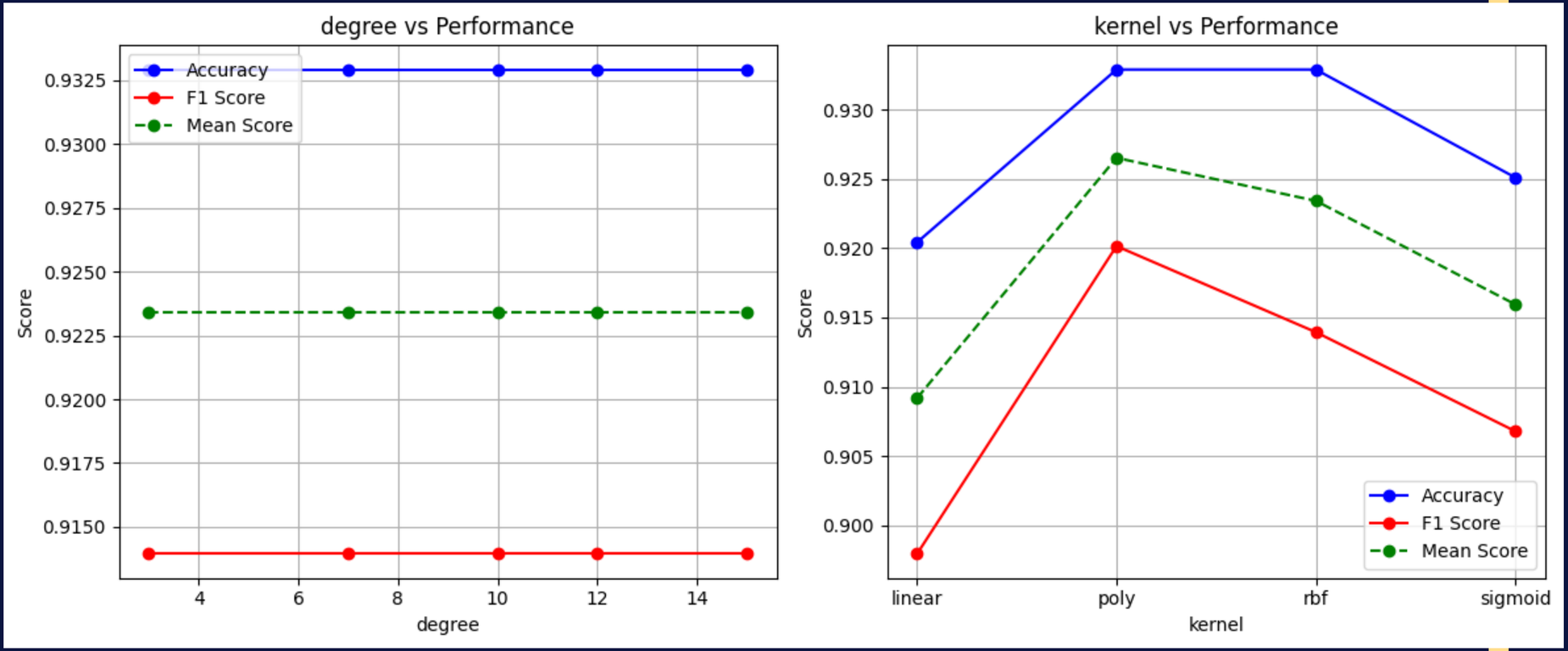
최고 Mean Score (max_depth): 0.9348 at max_depth=None

최고 Mean Score (max_features): 0.9292 at max_features=sqrt

04. 하이퍼파라미터 수정

Support Vector Machine

하이퍼파라미터 : degree, kernel



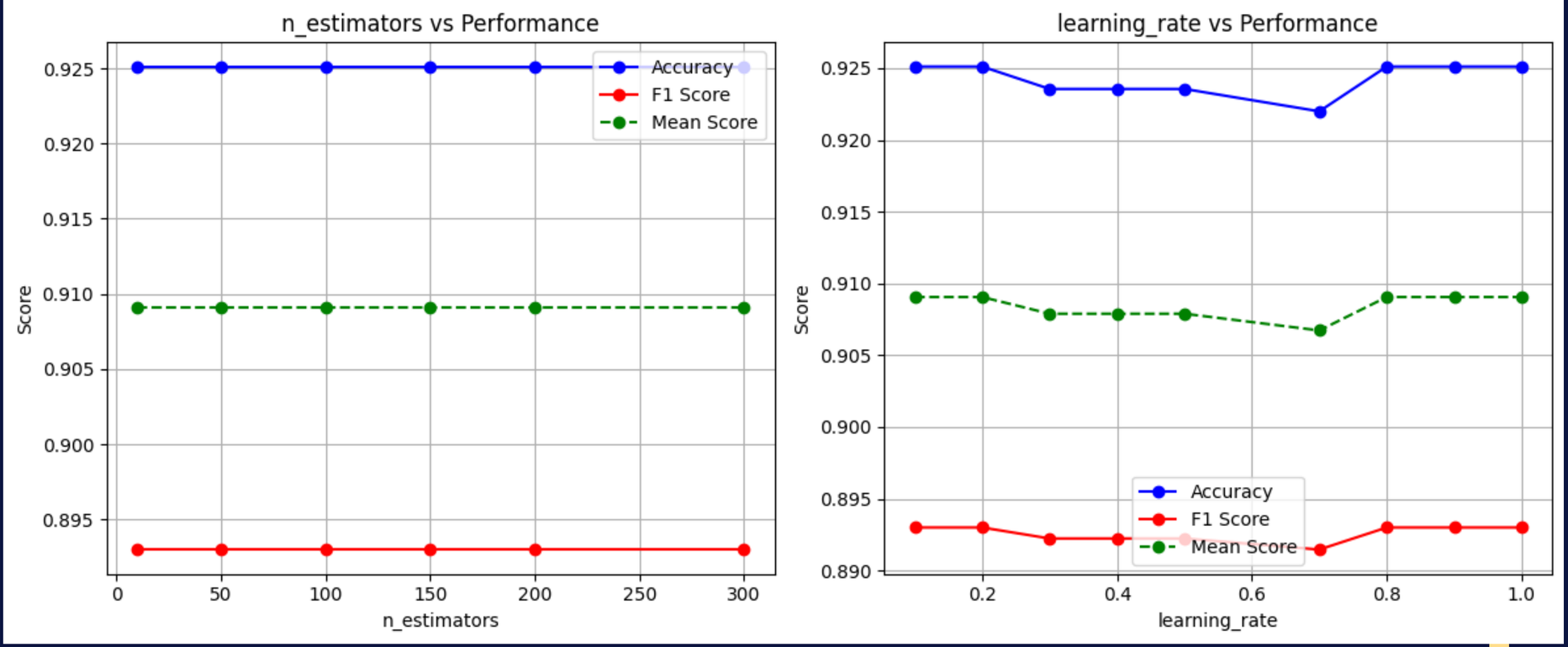
최고 Mean Score (degree): 0.9234 at degree=3

최고 Mean Score (kernel): 0.9265 at kernel=poly

04. 하이퍼파라미터 수정

AdaBoost

하이퍼파라미터 : n_estimators, learning_rate



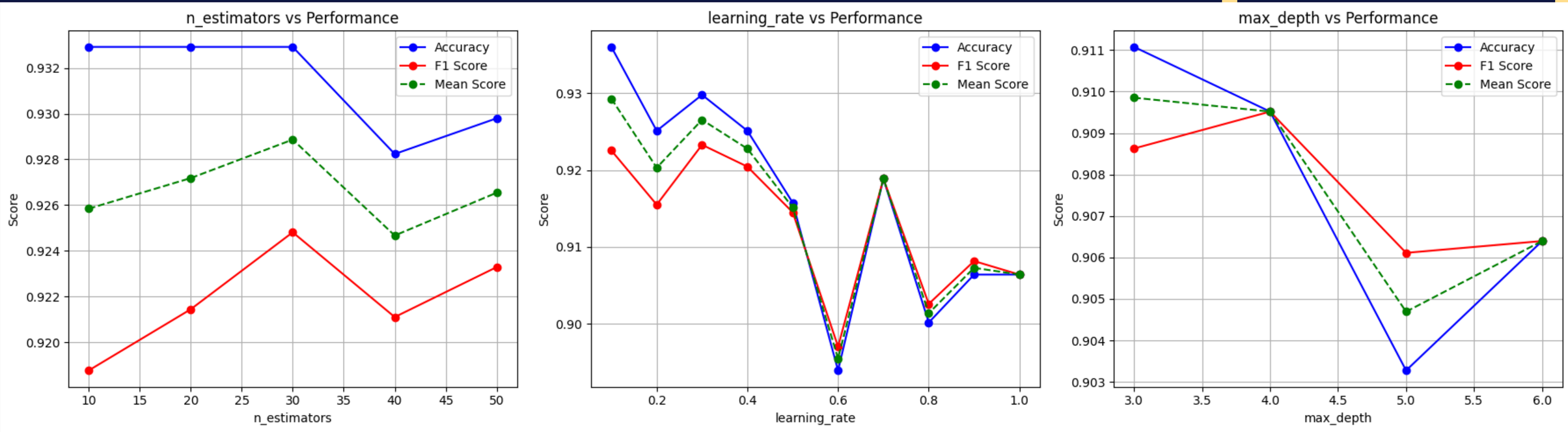
최고 Mean Score (n_estimators): 0.9091 at n_estimators=10

최고 Mean Score (learning_rate): 0.9091 at learning_rate=0.1

04. 하이퍼파라미터 수정

XGBoost

하이퍼파라미터 : n_estimators, learning_rate, max_depth



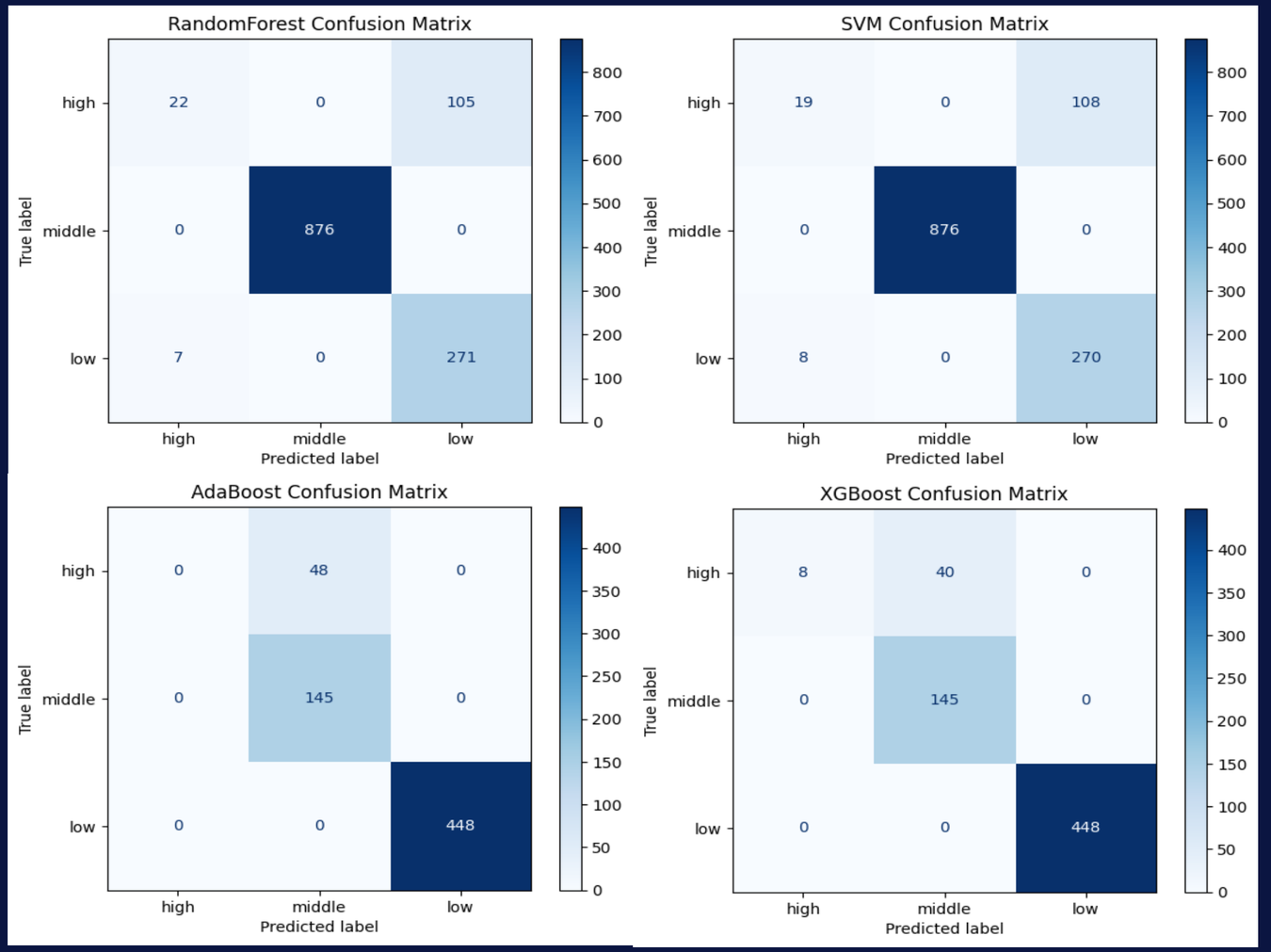
최고 Mean Score (n_estimators): 0.9288657563196047 at n_estimators=30

최고 Mean Score (learning_rate): 0.9292951107694905 at learning_rate=0.1

최고 Mean Score (max_depth): 0.9098519065039075 at max_depth=3

04. 하이퍼파라미터 수정

혼동행렬



05.결과 분석 및 한계점

결과분석

위험도 중,상의 데이터 비율 분석

	시간	계절	요일	주말여부	지역	사고내용	사고유형	법규위반	노면상태	기상상태	도로형태	가해운전자 차종	가해운전자 성별	가해운전자 상해정도	피해운전자 차종	피해운전자 성별	피해운전자 상해정도	시간대
'중' 또는 '상' 비율이 가장 높은 항목	14.00	가을	금	주중	경기도	중상사고	차대차	안전운전불이행	건조	맑음	기타	승용	남	상해없음	승용	남	중상	주간
비율	1.83	11.97	4.59	21.04	9.5	24.32	26.37	21.15	24.5	24.29	25.76	15.53	25.67	19.25	16.79	21.08	14.7	18.87
다른 종류 평균 비율	1.18	5.68	4.07	7.95	1.59	4.67	1.91	1.51	1.12	0.94	0.54	9.57	9.33	1.95	1.22	1.75	2.04	10.12
비율 차이	0.65	6.29	0.53	13.09	8.11	19.65	25.06	19.85	23.37	23.35	25.22	12.16	22.34	17.3	15.57	20.23	12.66	8.75

- 중,상의 데이터의 비율이 가장 높은 카테고리를 각 변수별로 분석해보았고 그 결과

  a. 계절은 가을 , 법규위반내용은 안전운전불이행, 노면상태는 건조, 기상상태는 맑음 이었을 때 다른 조건의 평균보다 높은 경향을 나타냄
  b. 위험도가 '중'또는 '상'의 데이터에서 가해운전자의 상해정도는 상해없음이 제일 높았지만 피해운전자의 상해정도는 중상이 제일 높음

05.결과 분석 및 한계점

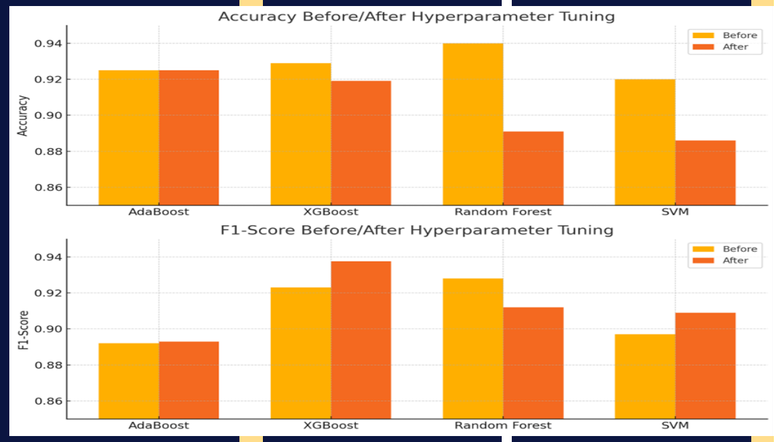
한계점

- 데이터상에서 나오지 않는 예를 들어, 차량의 상태나 운전자의 운전실력, 또는 그날 도로의 상태(차가 많이 밀린다던가 공사중이라던가 등)에 따른 변수들 역시 사고발생에 영향을 미칠 수 있기에 이것들 역시 고려해보아야 함 만약 온도, 습도 등의 컬럼이 추가된다면 지역별 날씨특성과 위험도의 관계까지 분석할 수 있을 것으로 예상됨

- 하이퍼파라미터 조정을 각각의 변수별로 최적의 값이 나오는 변수값을 설정하고 최종적으로 설정했지만 그것이 꼭 최상의 정확도가 될 수 없고 최고의 정확도를 이루는 특정 조합이 존재하므로 특정 조합을 찾아내는 방향으로 개선할 필요가 있음

- 3년간의 고속도로에서의 데이터 약 6400개를 사용하여 분석했지만 다양한 요소의 변수가 존재하여 6400개로는 충분하다고 보기는 어려워보임.

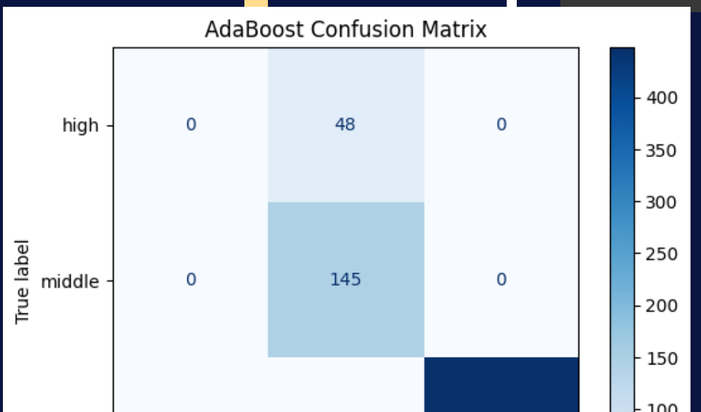
- 또한 특정 컬럼, 예를 들어 노면상태 같은 변수는 '건조 ' 값이 5647개로 전체의 88.2%를 차지하는 등 일부 컬럼에서 편향이 발생하여 결과의 모델의 성능을 떨어뜨릴 수도 있음



노면상태	count
건조	5647
젖음/습기	686
적설	32
서리/결빙	30
기타	6

기상상태	count
맑음	5600
비	539
흐림	179
눈	63
기타	12
안개	8

도로형태	count
기타	5741
교량	214
터널	192
지하차도	181
주차장	35
교차로	24
고가도로위	14



05.결과 분석 및 한계점

한계점

- TASS 교통사고 분석 시스템에서 제공하는 날씨 데이터는 노면상태, 기상상태인데 카테고리가 적어 유의미한 분석이 어려움 (온도, 습도 등의 컬럼이 추가된다면 지역별 날씨특성과 위험도의 관계까지 분석할 수 있을 것으로 예상됨)

- 사고심각도 지수(사망자수*3+중상자수*2+ 경상자수) 를 만들어 각각의 변수에 대한 관계를 회귀분석해보려 했지만 사고심각도가 정규성을 만족하지 않았고 회귀모델을 만든결과 R2값이 너무 낮아(데이터가 모델을 잘 설명하지 못함) 활용하지 못함

06. 향후 계획

- 정규성을 만족하는 종속변수를 새로 설정하여 회귀분석을 진행할 예정
- 4가지 모델의 정확도와 F1-score를 가장 높일 수 있는 최적의 하이퍼파라미터 조합을 찾을 예정
- 속도, 기온, 습도 등의 변수들을 추가하여 모델을 업데이트하고 네비게이션 기업에 배포