

SW 소모임\_Orange3

# 청소년 비행 행동 예측 모델

팀원 : 총 4명 (이름 미공개)

# CONTENTS

01

INTRODUCTION



03

PREDICTION



02

EDA



04

Q&A



# INTRODCUTION

01 주제 선정 이유 (Reason for selecting topic)



02 연구 기획 (Research Project)



03 연구 문제 (Research Question)



# 주제 선정 이유

- 국가는 청소년을 유해한 환경으로부터 보호함으로써 청소년이 건전한 인격체로 성장할 수 있는 환경을 마련해야 할 의무가 있음
- 하지만, 지역 사회 내 유해업소 난립, SNS 유해 게시물 범람, 부실한 성인인증 문제 등이 존재함
- 이로 인해 현대 청소년들은 다양한 유해 환경에 무분별하게 노출되어 있음

## 학교 앞 성인용품점, 걸어서 2분

✎ 화성저널 | ⓒ 승인 2024.07.19 09:33 | 💬 댓글 0

학교 앞 유해업소 72곳, 작년보다 17곳 늘어  
등하굣길 아이들에게 주기적으로 노출 매우 심각, 관리·감독 시급

## 청소년 79% "폭력·선정적 콘텐츠 봤다"...SNS가 주경로

송고시간 | 2016-02-14 08:01

## 허술한 성인인증에...청소년 무방비 노출 '전자담배 무인 판매점'

아시아투데이 원문 | 기사전송 2024-12-10 16:54 최종수정 2024-12-10 18:39

AI챗으로 요약

# 주제 선정 이유

- 또한, 청소년 비행 유형이 다양해지고 심화되고 있음

## 2022년 청소년 매체이용 유해환경 실태조사 결과

- 초등학생의 70.6% 확장가상세계(메타버스) 이용, 청소년 46% 온라인 도박성 게임 중 카드화투 경험
- 청소년 폭력 피해 1순위는 '언어폭력', 가해자로 '온라인에서 알게 된 사람' 증가
- 청소년 '음주 경험'은 13.7%로 다소 증가, '흡연 경험'은 4.2%로 약간 감소
  - 근로 청소년 중 최저임금을 지급받지 못한 비율은 12.6%로 크게 감소

여성가족부 전국 초(4~6학년).중.고등학교에 재학 중인 청소년 17,140명을 대상으로 실시한 '2022년 청소년 매체이용 유해환경 실태조사'

# 주제 선정 이유

청소년의 중첩적인 비행 유형을 파악하기 위해 청소년 비행에 영향을 미치는 **다차원적 예측 요인 고찰**해 청소년을 보호할 수 있는 **효과적인 방안을 마련**하기 위해 이와 같은 주제를 선정하게 됨

# 연구 기획



## 청소년 비행 잠재 집단 도출

- 비행 행동 패턴에 따라  
주요 집단 도출
- 각 집단의 특성과 관련 요인  
도출 후 해석



## 비행 행동 예측 모델 설계

- 2020-2021년 데이터를 학습 데이터로,  
2022년 데이터를 테스트 데이터로 사용
- 모델 간 성능 비교를 통해  
최적 모델 선택



## 비행행동 개선 프로그램 설계

- 잠재 집단과 예측 모델 결과를 바탕으로  
집단별로 맞춤형 개선 프로그램 구상
- 심리적/환경적 요인을 고려한  
맞춤형 개입 방안 도출

# 연구 문제

"청소년의 비행 행동을 효과적으로 예측할 수 있는 최적의 모델은 무엇인가?"

"비행 행동 예측 결과를 바탕으로 개선 방안을 어떻게 도출할 수 있는가?"



# EDA

01

활용 데이터셋 (Data Set)



02

데이터 전처리 (Data Preprocessing)



03

데이터 탐색 (Data Exploration)



# 활용 데이터셋

- 데이터 출처

한국 아동/청소년 데이터 아카이브에서 제공한 한국청소년패널조사(KYPS)의 중등 데이터

- 데이터 정보

기간: 2020년 - 2022년

구조: 2590개의 행(청소년 개별 응답)과 377개의 열(비행 행동, 심리적 상태, 환경 요인 등)

HID	PID	SCLIDw2	WEIGHTA1w2	WEIGHTA2w2	WEIGHTB1w2	WEIGHTB2w2	SURVEY1w2	SURVEY2w2	COHORTw2	ARA1Aw2	ARA2Aw2
780	2	20409	214.69370861	1.2606867272	215.72134739	1.2667210474	1	1	m1	4	1
1192	2	20912	87.627325106	0.5145498041	87.931552304	0.5163362337	1	1	m1	9	3
1193	2	40920	105.8349605	0.6214654861	105.58365004	0.6199897849	1	1	m1	9	3
1285	2	20920	278.94311942	1.6379608447	278.01667008	1.6325207114	1	1	m1	9	2
1590	2	20936	83.644264874	0.4911611766	83.125628718	0.4881157324	1	1	m1	9	3
2079	2	21301	61.114280392	0.3588645547	59.774876332	0.3509995412	1	1	m1	13	3
2145	2	41304	148.07327915	0.86948993	144.82804825	0.8504338545	1	1	m1	13	2






# 데이터 전처리

- 비행 행동 예측과 거리가 있는 문항 및 개인의 노력으로 변화시킬 수 없는 정보 제외  
→ 예를 들어 개인정보, 성별, 형제자매 수 등

Name	Type	Role
HID	<b>N</b> numeric	<b>skip</b>
PID	<b>C</b> categori...	<b>skip</b>
SCLIDw5	<b>N</b> numeric	<b>skip</b>

# 데이터 전처리

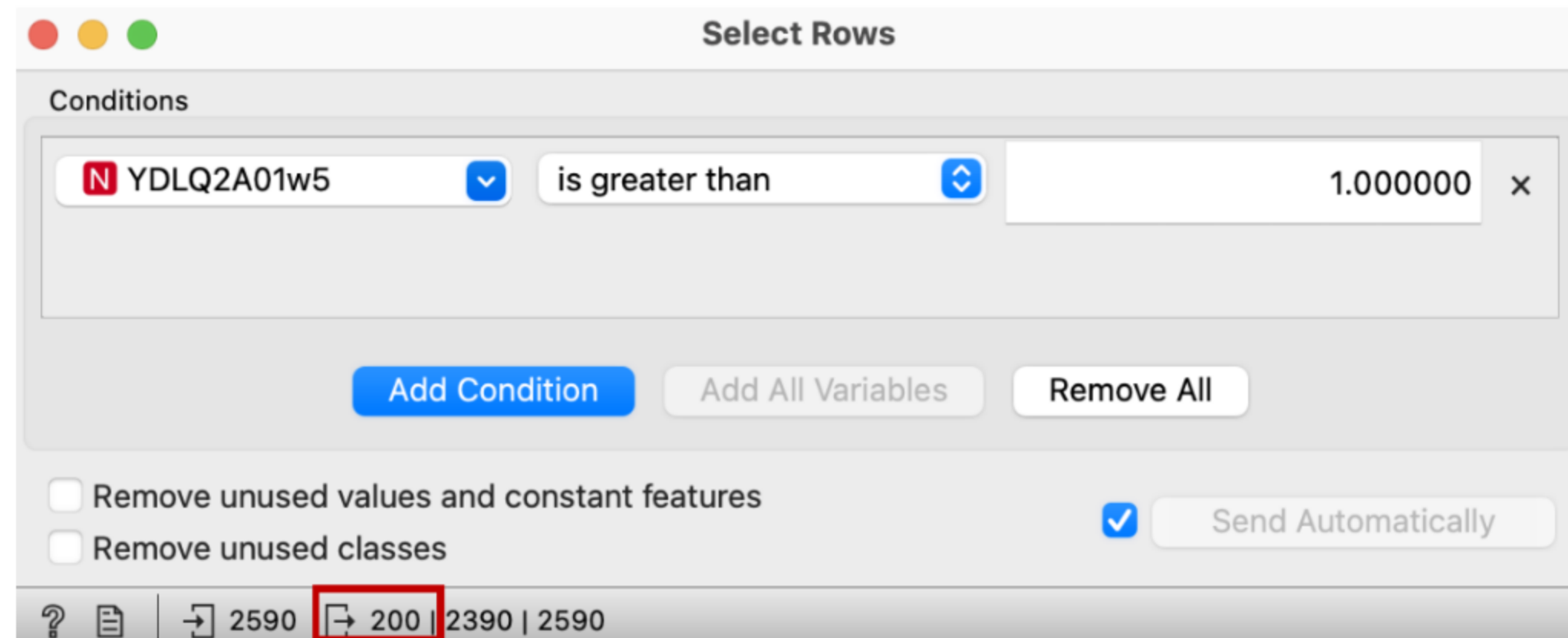
- 결측 값이 높은 문항 제외

^	Name	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.	Missing
N	YACT1B02w3		2.96		3	0.24	1	4	1930 (81 %)
N	YACT1B03w3		3.19		3	0.21	1	4	1871 (78 %)
N	YACT1B04w3		3.14		3	0.19	1	4	1479 (62 %)
N	YACT1B05w3		3.04		3	0.21	1	4	1462 (61 %)
N	YACT1B06w3		3.11		3	0.22	1	4	2247 (94 %)

# 데이터 전처리

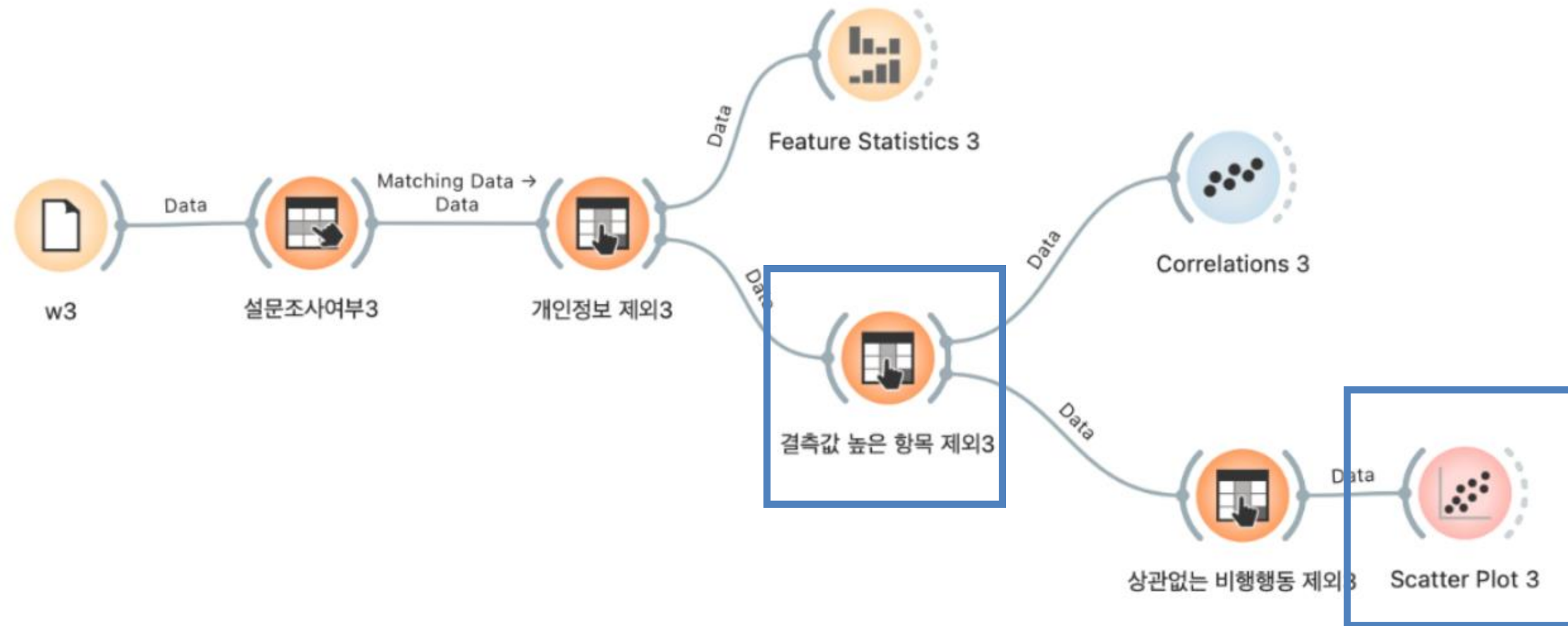
- 분석 대상

2020년 - 2022년 비행 행동 데이터 중, 세 연도 모두 응답자 수가 100명 이상인 데이터



→ YDLQ1A02(술 마시기)와 YDLQ2A01(누군가에게 욕이나 험한 말을 보냄)

# 데이터 전처리



# 데이터 탐색

- 비행행동 간 상관관계 분석 결과 (2020년)

1	YDLQ1A02w3, YTIM1E01w3	YDLQ1A02(술 마시기)
2	YDLQ1A02w3, YFUR2A07w3	
3	YDLQ1A02w3, YINT2B15w3	1) YTIM1E01 학기 중 학원 다님
4	YDLQ1A02w3, YPSY4A01w3	2) YFUR2A07 진로 관련하여 학원
5	YDLQ1A02w3, YPSY4A03w3	선생님과 상담
6	YDLQ1A02w3, YDIS1A12w3	3) YINT2B15 수업에 적극적으로
7	YDLQ1A02w3, YTIM1B03w3	참여하지 않음
8	YDLQ1A02w3, YPSY1A03w3	4) YPSY4A01 칭찬을 받거나 벌을
9	YDLQ1A02w3, YPSY4A04w3	받아도 금방 다시 주의가
10	YDLQ1A02w3, YINT2B14w3	산만해짐
11	YDLQ1A02w3, YPSY4C07w3	5) YPSY4A03 오랫동안 집중해야
12	YDLQ1A02w3, YDIS1A10w3	하는 과제는 하고 싶지 않음
13	YDLQ1A02w3, YEDU2A09w3	
14	YDLQ1A02w3, YPSY4C02w3	
15	YDLQ1A02w3, YPSY4B03w3	
16	YDLQ1A02w3, YTIM1N02w3	
17	YDLQ1A02w3, YEDU1A00w3	
18	YDLQ1A02w3, YPSY5A12w3	
19	YDLQ1A02w3, YDIS1A09w3	
20	YDLQ1A02w3, YTIM1A03w3	

1	YDLQ2A01w3, YINT2A09w3	YDLQ2A01(누군가에게 욕이나 험한 말을 보냄)
2	YDLQ2A01w3, YDIS1A12w3	
3	YDLQ2A01w3, YINT2B16w3	1) YINT2A09 나는 공부를 잘함
4	YDLQ2A01w3, YDIS1A11w3	2) YINT2B16 스스로 학습계획을 짜본
5	YDLQ2A01w3, YEDU2A03w3	적이 없음
6	YDLQ2A01w3, YTIM1C01w3	3) YEDU2A03 친구들에게 내 이야기를
7	YDLQ2A01w3, YPSY2A01w3	잘함
8	YDLQ2A01w3, YINT2A10w3	4) YTIM1C01 수면의 질 여부
9	YDLQ2A01w3, YINT2A13w3	5) YPSY2A01 자신이 불행한지 느낌 여부
10	YDLQ2A01w3, YTIM1A04w3	
11	YFUR2A06w3, YDLQ2A01w3	
12	YDLQ2A01w3, YPSY4B02w3	
13	YDLQ2A01w3, YACT2A02w3	
14	YDLQ2A01w3, YPSY4E09w3	
15	YPHY1C00w3, YDLQ2A01w3	
16	YFUR2A05w3, YDLQ2A01w3	
17	YPSY1A05w3, YDLQ2A01w3	
18	YDLQ2A01w3, YEDU1A00w3	
19	YDLQ2A01w3, YDIS3A08w3	
20	YDLQ2A01w3, YMDA1C12w3	



# 데이터 탐색

- 비행행동 간 상관관계 분석 결과 (2021년)

1	YDLQ1A02w4, YPSY4B03w4
2	YDLQ1A02w4, YPSY4B04w4
3	YDLQ1A02w4, YFUR3C04w4
4	YDLQ1A02w4, YPSY3A01w4
5	YDLQ1A02w4, YACT1A07w4
6	YDLQ1A02w4, YPSY4B05w4
7	YDLQ1A02w4, YFUR3B02w4
8	YDLQ1A02w4, YFUR3C01w4
9	YDLQ1A02w4, YPSY4A06w4
10	YDLQ1A02w4, YPSY6A27w4
11	YPSY4A01w4, YDLQ1A02w4
12	YDLQ1A02w4, YPSY4E04w4
13	YDLQ1A02w4, YPSY7A01w4
14	YDLQ1A02w4, YFAM2C01w4
15	YDLQ1A02w4, YPSY5A08w4
16	YDLQ1A02w4, YINT2A07w4
17	YDLQ1A02w4, YPSY6A26w4
18	YDLQ1A02w4, YPSY6A15w4
19	YDLQ1A02w4, YFAM2D03w4
20	YDLQ1A02w4, YEDU3A07w4

## YDLQ1A02(술 마시기)

- 1) YPSY4B03 내가 원하는 것을 못하게 하면 따지거나 덤빔
- 2) YPSY4B04 별 것 아닌 일로 싸움
- 3) YFUR3C04 직업을 선택할 때 고려사항이 많이 결정이 어려움
- 4) YPSY3A01 나는 나에게 만족함 여부
- 5) YACT1A07 건강 관련 참여 활동 여부

1	YDLQ2A01w4, YRME1A26w4
2	YDLQ2A01w4, YEDU1C02w4
3	YDLQ2A01w4, YMDA1C07w4
4	YDLQ2A01w4, YPSY4A02w4
5	YDLQ2A01w4, YPSY7A01w4
6	YDLQ2A01w4, YPSY4A07w4
7	YDLQ2A01w4, YFAM2F03w4
8	YDLQ2A01w4, YMDA1B12w4
9	YDLQ2A01w4, YRME1A18w4
10	YDLQ2A01w4, YPSY4A01w4
11	YDLQ2A01w4, YEDU1B02w4
12	YDLQ2A01w4, YFUR2A03w4
13	YDLQ2A01w4, YPSY4A04w4
14	YDLQ2A01w4, YTIM1E02w4
15	YDLQ2A01w4, YPSY7A05w4
16	YDLQ2A01w4, YFAM2A01w4
17	YDLQ2A01w4, YMDA1C13w4
18	YDLQ2A01w4, YPSY5A03w4
19	YDLQ2A01w4, YTIM1K01w4
20	YDLQ2A01w4, YPSY6A24w4

## YDLQ2A01(누군가에게 욕이나 험한 말을 보냄)

- 1) YMDA1C07 가족들이나 친구들과 함께 있는 것보다 스마트폰을 사용하는 것이 더 좋음
- 2) YPSY4A02 문제를 풀 때 문제를 끝까지 읽지 않음
- 3) YPSY7A01 나는 무엇을 하다가 다른 생각이 나면 집중하기 어려움
- 4) YPSY4A07 글자를 잘 빠뜨리고 쓰는 편임
- 5) YFAM2F03 부모님은 나에게 대한 규칙을 자주 바꾸심

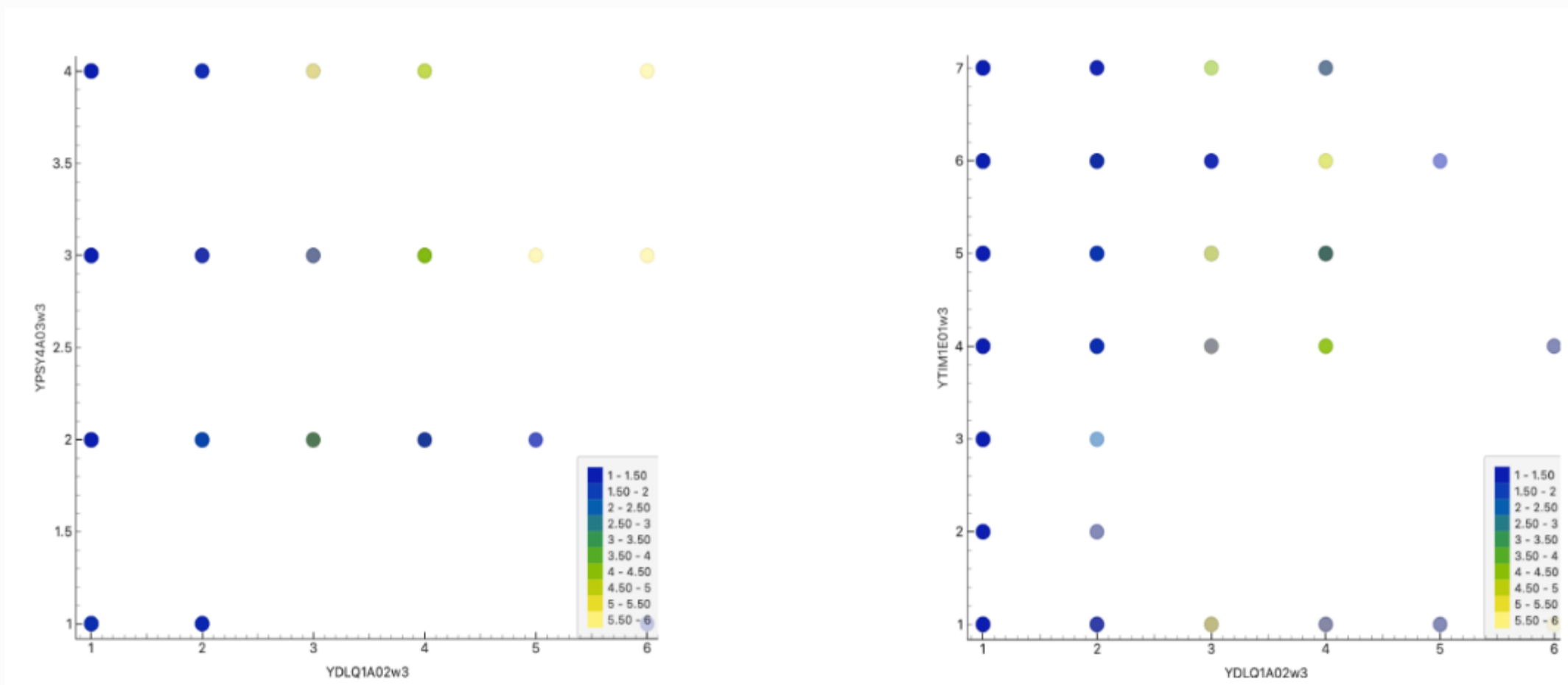


# 데이터 탐색

- 비행행동 데이터의 산점도 분석

점들이 특정한 직선적 방향으로 나열되지 않음

X축과 Y축 모두 이산형 값으로 나뉘며, 특정 값에 점들이 모여 있음



→ 비선형 데이터임을 확인

# PREDICTION

01

모델 세팅 (Model Setting)



02

결과 분석 (Analysis of Results)



# 모델 세팅

Train Data



2020-2021년  
한국청소년  
패널조사  
데이터

Test Data



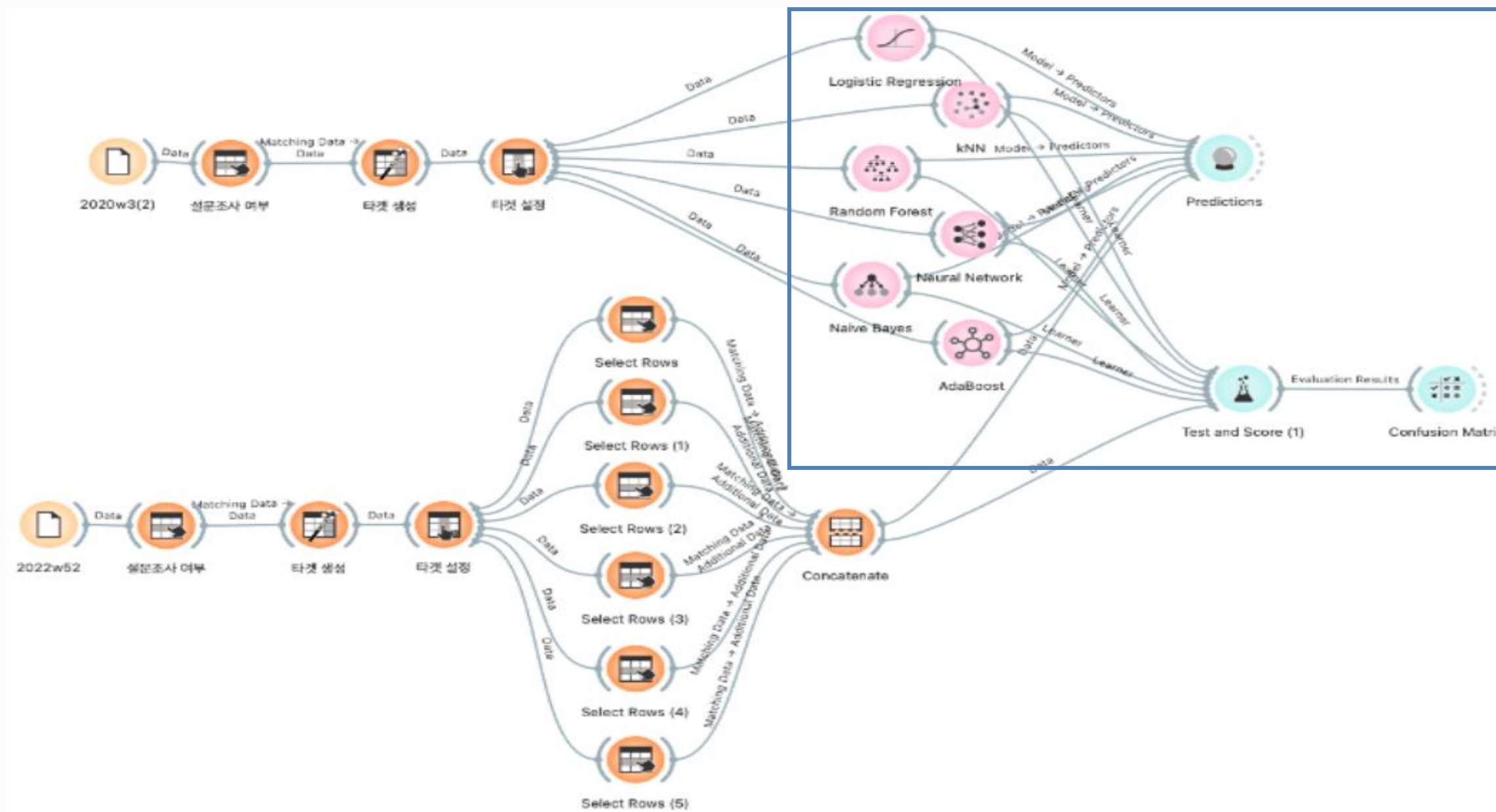
2022년  
한국청소년  
패널조사  
데이터

Target



비행행동  
데이터  
YDLQ1A02  
YDLQ2A01

# 모델 세팅



# 결과 분석

- 2020년

Model	AUC	CA	F1
Logistic Regression	0.698	0.774	0.722
kNN	0.824	0.788	0.795
Random Forest	0.904	0.847	0.839
Neural Network	0.764	0.810	0.777
Naive Bayes	0.731	0.791	0.751
AdaBoost	0.930	0.859	0.854

- 2021년

Model	AUC	CA	F1
Logistic Regression (1)	0.762	0.755	0.742
kNN (1)	0.871	0.802	0.803
Random Forest (1)	0.918	0.876	0.872
Neural Network (1)	0.855	0.864	0.858
Naive Bayes (1)	0.832	0.826	0.819
AdaBoost (1)	0.934	0.874	0.872

YDLQ1A02(술 마시기)

- 모든 모델에서 2021년 데이터가 더 높은 AUC와 F1-score 기록
- 두 연도 모두에서 AdaBoost와 Random Forest가 가장 우수한 성능
- Logistic Regression이 가장 낮은 성능

# 결과 분석

- 2020년

Model	AUC	CA	F1
Logistic Regression	0.740	0.800	0.752
kNN	0.802	0.767	0.779
Random Forest	0.882	0.872	0.861
Neural Network	0.823	0.861	0.842
Naive Bayes	0.768	0.794	0.753
AdaBoost	0.893	0.867	0.858

- 2021년

Model	AUC	CA	F1
Logistic Regression	0.755	0.706	0.702
kNN	0.908	0.804	0.805
Random Forest	0.960	0.867	0.866
Neural Network	0.888	0.830	0.827
Naive Bayes	0.857	0.785	0.784
AdaBoost	0.973	0.895	0.895

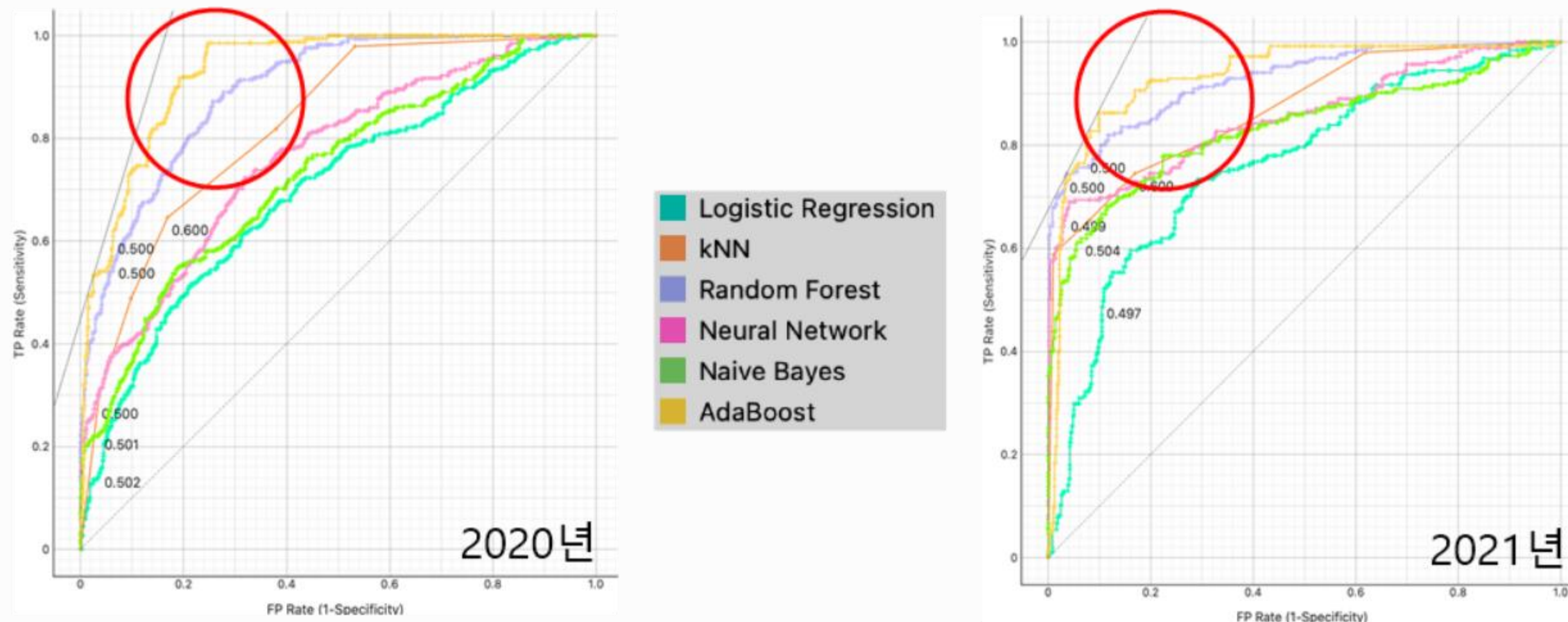
YDLQ2A01(누군가에게 욕이나 험한 말을 보냄)

→ 술 마시기 행동 예측과 같은 결과가 나타남



# 결과 분석

- ROC 곡선 : 모델의 민감도와 특이도 간의 관계로 좌측 상단에 가까울수록 우수한 성능



→ 전체적으로 2021년 곡선이 더 높은 곡선의 형태

→ AdaBoost와 Random Forest의 곡선이 좌측 상단에 가까움

# 결과 분석

- 결론

AdaBoost와 Random Forest의  
우수성

두 모델이 모든 비행 행동과 연도에서 최고의 성능을 기록하여,  
비선형적 특성을 가진 데이터 분석에 매우 적합한 모델로 확인

최신 데이터의 중요성

2021년 데이터가 전체적으로 성능을 향상시킨 것은 최신  
데이터 품질이 분석과 예측에 중요하다는 것을 시사함



# 결과 분석

- 한계와 개선 방안

## DATA 관점

- 외부 변수 반영 부족
- 민감한 주제로 인한 응답률 저조

### HOW?

- 해당 시점에 해당하는 뉴스 트렌드나 사회적 사건 등을 연계
- 익명성을 강화하고 우회적 질문 도입

## MODELING 관점

- Orange3 내 딥러닝 위젯 부재
- 모델 간 비교의 한계

### HOW?

- 데이터 전처리는 Orange3에서 수행
- 파이썬으로 LSTM이나 Transformer와 같은 딥러닝 모델 구현



Q&A