

마키아벨리즘 성향 테스트 결과 데이터 기반 국가선거 투표참여 여부 예측



1. 머신러닝 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



#1. GOAL



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



목표 = “어떤 사람이 투표했나 알아맞히기”

마키아벨리즘 성향 테스트 답변 데이터



머신러닝 분류모델 활용



투표참여 여부 예측

#2. DATA INFO



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



EDA

Theme of the DATA

#2. DATA INFO



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



마키아벨리즘이란 ?

국가의 발전과 인민의 복리증진을 위해서는
어떠한 수단이나 방법도 허용된다는
국가 지상주의적인 정치 이념

출처 | 표준국어대사전

그래서, 마키아벨리즘 성향이면 어떻다는 거?



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



+ ▲
마키아벨리즘
성향이
높은 사람

낮은 친화력, 높은 성실성
사람들과 소통할 때
계산적이고 **신중**하게 접근하는 경향

— 평균 약70점 —

마키아벨리즘
성향이
낮은 사람
- ▼

비교적 수동적, 순응적
사람들과 소통할 때
개인적이며 **감정을 이입**하여 접근하는 경향

DATASET = 테스트에 대한 각 참여자의 답변들



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



< DATASET 기본 정보 >

1

데이터 출처

심리학 공공데이터 사이트 (미국)<https://openpsychometrics.org/>

2

데이터 내용
(73,489 , 105)

**2017년 7월부터 2019년 3월까지의 기간 중
온라인에서 진행된 마키아벨리즘 테스트의
각국 참여자 답변 데이터 (영어 사용자)**

3

데이터 구성

- ① **마키아벨리즘 성향 판단 테스트 답변**
- ② **성격, 연령 등 인적사항 설문조사 답변**



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



마키아벨리즘 테스트 질문지 구성

< 마키아벨리즘 성향 판단 테스트 예시 >

Q8

마키아벨리즘 +▲

대부분의 사람들은
강제하지 않으면
열일하지 않는다

- | | |
|----------|-------|
| 1점 ----- | 매우 반대 |
| 2점 ----- | 조금 반대 |
| 3점 ----- | 보통 |
| 4점 ----- | 조금 동의 |
| 5점 ----- | 매우 동의 |

Q13

마키아벨리즘 +▲

일반인과 범죄자의
가장 큰 차이는
발각될 만큼 멍청한지
여부에 있다

- | | |
|----------|-------|
| 1점 ----- | 매우 반대 |
| 2점 ----- | 조금 반대 |
| 3점 ----- | 보통 |
| 4점 ----- | 조금 동의 |
| 5점 ----- | 매우 동의 |

Q11

마키아벨리즘 -▼

세계적인 리더들은
청렴하고,
도덕적이다

- | | |
|----------|-------|
| 1점 ----- | 매우 반대 |
| 2점 ----- | 조금 반대 |
| 3점 ----- | 보통 |
| 4점 ----- | 조금 동의 |
| 5점 ----- | 매우 동의 |



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



마키아벨리즘 테스트 질문지 구성

< 마키아벨리즘 성향 측정 척도에 따른 질문지 구분 >

+▲ 마키아벨리즘 지지 성향 판단 (10)		-▼ 마키아벨리즘 반대 성향 판단 (10)	
T	<p>Q12. 아무나 믿는 사람은 문제에 잘 휘말린다 Q15. 중요한 사람에게는 아첨을 해야 한다 Q2. 사람을 잘 다루려면 듣기 좋은 소리를 해야 한다 Q1. 일에 대한 진짜 목적은 도움이 되지 않는 한, 남들에게 말하지 않아야 한다</p>	T	<p>Q3. 도덕적으로 옳은 일만 행해야 한다 Q6. 정직함이 최선의 방책이다 Q10. 무엇인가를 부탁할 때는 그 일이 중요한 듯이 말하기보다 솔직한 이유를 말해야 한다 Q16. 모든 면에서 선하게 행하는 것이 가능하다 Q7. 거짓말을 했다면, 어떤 상황이라도 변명의 여지는 없다</p>
V	<p>Q5. 사람은 기본적으로 욕망이 있으며, 기회가 있으면 그 본성이 드러난다 Q8. 대부분의 사람들은 강제하지 않으면 열심히 일하지 않는다 Q18. 남들보다 앞서기 위해서는 꼼수를 써야 한다 Q13. 일반인과 범죄자의 가장 큰 차이는 범죄사실이 발각될 만큼 명청한 지 여부에 있다 Q20. 대부분의 사람들은 자산을 잃은 사실보다 부모를 잃은 사실을 더 쉽게 잊는다</p>	V	<p>Q11. 세계를 리드하는 사람은 청렴하고 도덕적이다 Q14. 대부분의 사람들은 용감하다 Q17. “매 초마다 나쁜 놈이 태어난다”는 말은 틀렸다 Q4. 대부분의 사람들은 기본적으로 착하고 친절하다</p>
M	Q19. 안락사는 인정되어야 한다	M	Q9. 부도덕하고 유명한 것보다 겸손하고 정직한 것이 모든 면에서 더 낫다

T : tactics 전략적 사고 V : views 관점 M : morality 도덕성

#2. DATA INFO



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



인적사항 설문지 구성

< 인적사항 설문지 예시 >

10개

성격판단 설문지

나는 외향적이고,
열정적인 편이다

1점	----- 매우 반대
2점	----- 조금 반대
3점	----- 쫌끔 반대
4점	----- 뭇도 아님
5점	----- 쫌끔 동의
6점	----- 조금 동의
7점	----- 매우 동의

16개

신뢰도판단 설문지

나는 다음 단어의
뜻을 알고있다

< pastiche >

1	----- 알고 있다
0	----- 모른다

13개

인적사항 설문지

대학에 다닌다면,
(다녔다면)
전공은 무엇인가?

주관식 답변 방식
ex) '심리학', '토목공학' 등

#2. DATA INFO



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



EDA

DATA in each column

#2. DATA INFO



EDA – 컬럼 구성



- 1. 프로젝트 목표
- 2. 데이터 소개
- 3. 모델링 과정
- 4. 모델 시연
- 5. 분류 결과 해석
- 6. 한계 및 과제



1 Q1A~Q20A

마키아벨리즘 테스트 질문지에 대한 답변
(int) 1~5 : 매우 반대 ... 매우 동의

2 Q1I ~ Q20I

마키아벨리즘 테스트 질문지 순서
(int) 1~20 : 테스트 과정에서의 각 질문지 순서

3 Q1E ~ Q20E

마키아벨리즘 테스트 질문지에 대한 답변 시간
(float) 0~??? : 밀리초

이상치 존재

4 TIPI1 ~ TIPI10

참여자의 성격 판단 질문지에 대한 답변
(int) 1~7 : 매우 반대 ... 매우 동의

#2. DATA INFO



EDA – 컬럼 구성



- 1. 프로젝트 목표
- 2. 데이터 소개
- 3. 모델링 과정
- 4. 모델 시연
- 5. 분류 결과 해석
- 6. 한계 및 과제



5 VCL1 ~ VCL16

참여자 답변 신뢰도 측정 질문지 답변
(int) 0~1 : 단어 뜻을 안다, 모른다

6 education

참여자의 학력
(int) 1~4 : 중졸, 고졸, 학사, 석박사

7 urban

참여자의 고향 설문지 답변
(int) 1~3 : 농어촌, 중소도시, 대도시

8 gender

참여자의 성별
(int) 1~3 : 남, 여, 기타

#2. DATA INFO



EDA – 컬럼 구성



- 1. 프로젝트 목표
- 2. 데이터 소개
- 3. 모델링 과정
- 4. 모델 시연
- 5. 분류 결과 해석
- 6. 한계 및 과제



9	engnat	참여자의 모국어 (int) 1~2 : 영어, 비영어
10	age	참여자의 나이 (int) 13~??? : 주관식 답변
11	hand	참여자의 주사용 손 (int) 1~3 : 오른손, 왼손, 양손
12	religion	참여자의 종교 (int) 1~12 : 불신론자, 크리스챤, 불교도..

이상치 존재

#2. DATA INFO



EDA – 컬럼 구성



- 1. 프로젝트 목표
- 2. 데이터 소개
- 3. 모델링 과정
- 4. 모델 시연
- 5. 분류 결과 해석
- 6. 한계 및 과제



13	orientation	참여자의 성정체성 (int) 1~5 : 이성애자, 동성애자, 양성애자..
14	race	참여자의 인종 (int) 10~70 : 10단위 수 : 백인, 흑인, 아시아인..
15	voted	참여자의 과거 국가 선거 투표 참여 여부 (int) 1~2 : 참여, 불참
16	married	참여자의 결혼 상태 (int) 1~3 : 미혼, 최근 결혼, 과거 결혼

label column

#2. DATA INFO



EDA – 컬럼 구성



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



17	familysize	참여자의 형제자매 수 (int) 0~??? : 주관식 답변	이상치 존재
18	major	참여자의 학부 전공 (str) ??? : 주관식 답변	군집화 불가
19	country	참여자의 인터넷 서버 지역 (str) US~KR : 알파벳 2자리 국가 코드	자동수집 정보
20	screenw~h	참여자의 스크린 사이즈 (float) 0~??? : 스크린 가로/세로 사이즈	자동수집 정보
21	intro~...elapse	참여자의 테스트 구간별 답변 시간 (float) 0~??? : 밀리초	자동수집 정보

#2. DATA INFO



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



EDA

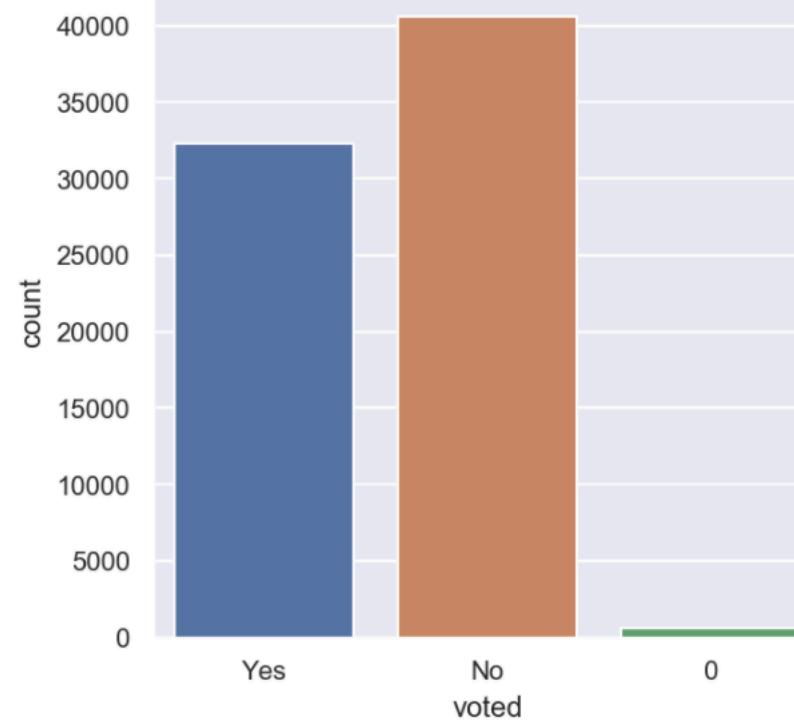
voted & other columns



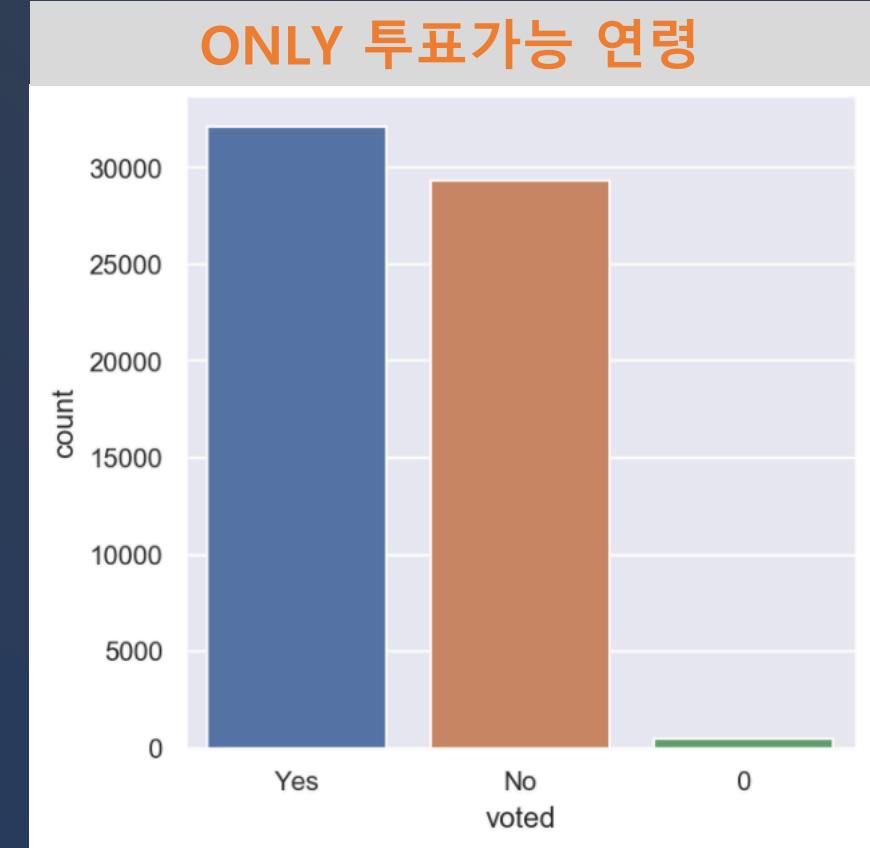
1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



투표 불가능 연령 포함



ONLY 투표 가능 연령



Q1~Q20 마키아벨리즘 답변

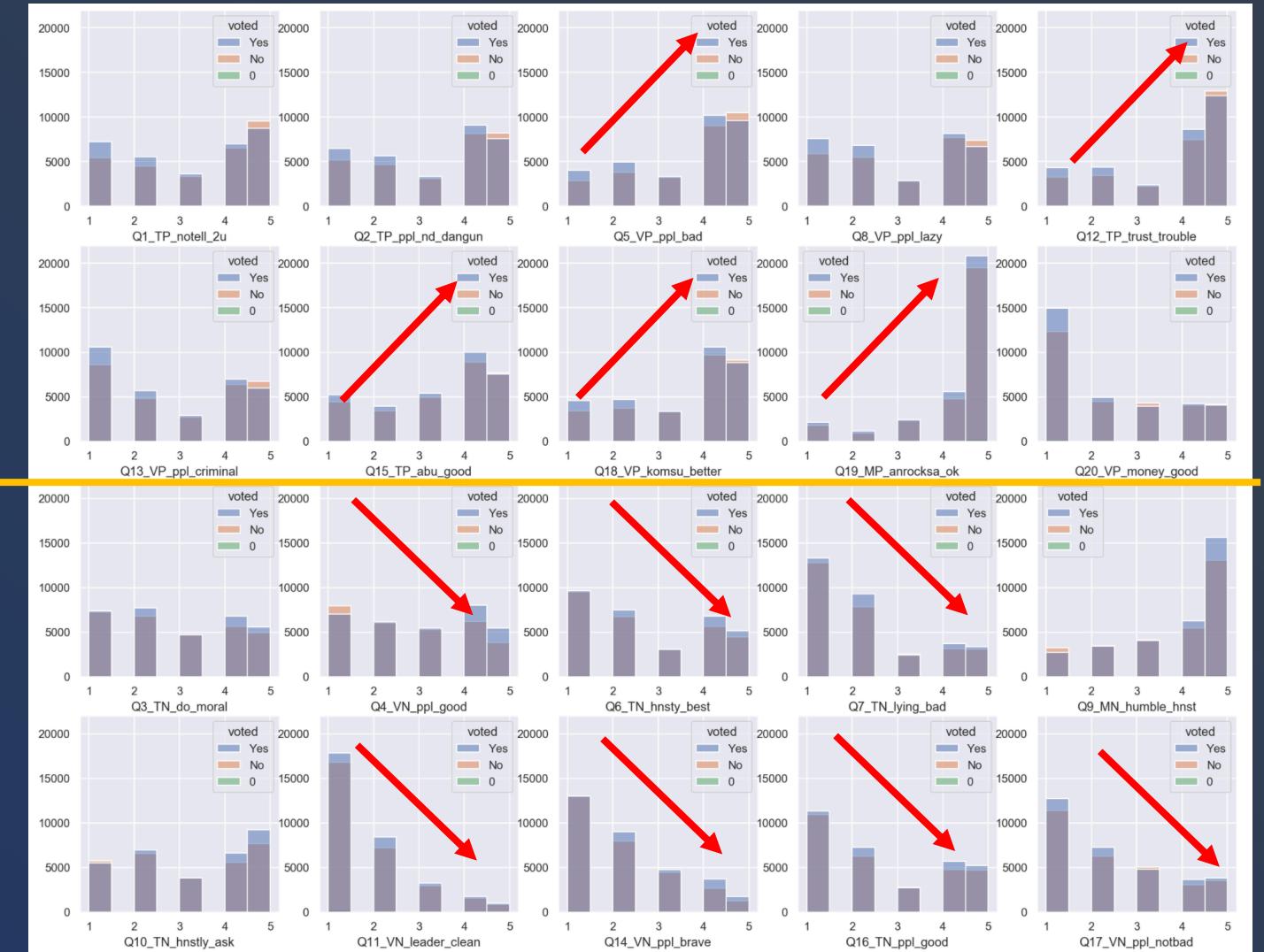


1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



+ ▲
마키아벨리즘
지지성향 질문

- ▼
마키아벨리즘
반대성향 질문



Q1~Q20 마키아벨리즘 답변



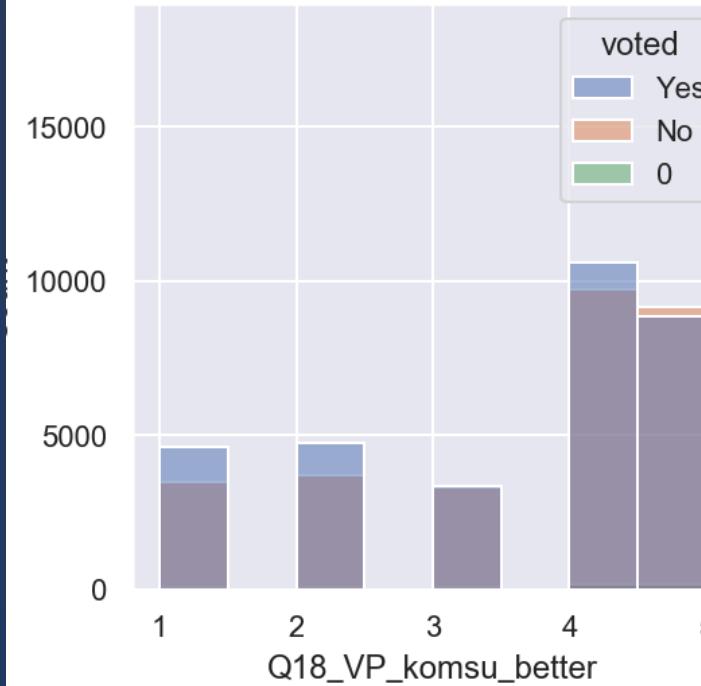
1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



Q18

마키아벨리즘 +▲

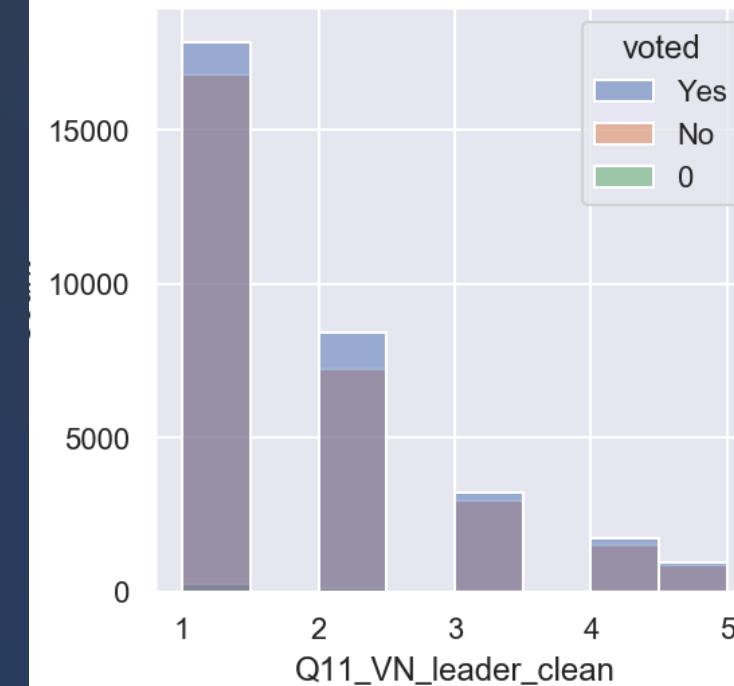
꼼수를 부리지 않고서는
앞서나갈 수 없다



Q11

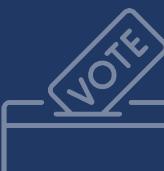
마키아벨리즘 -▼

세계적인 리더들은 청렴하고,
도덕적이다





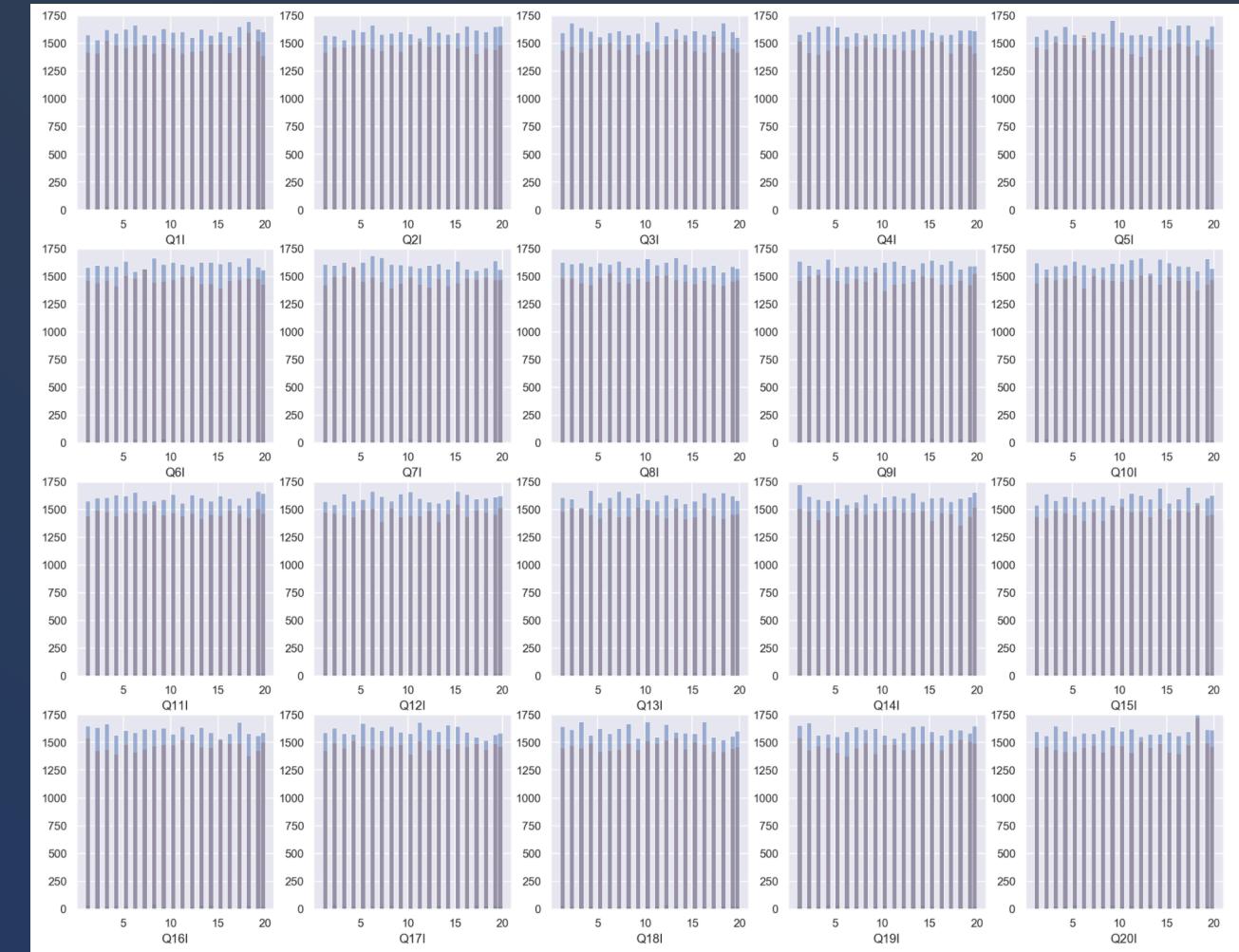
1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



QI
: the position of that item in the survey

II

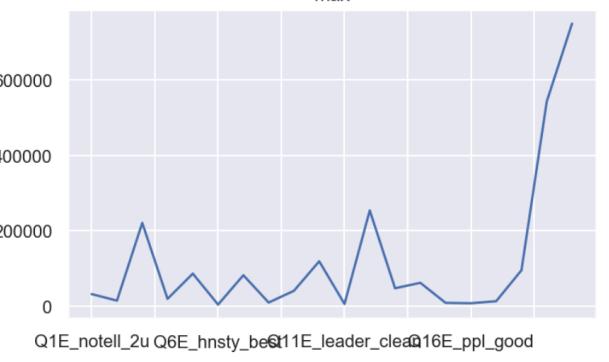
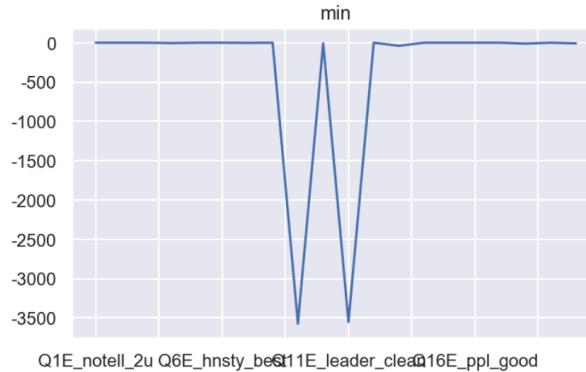
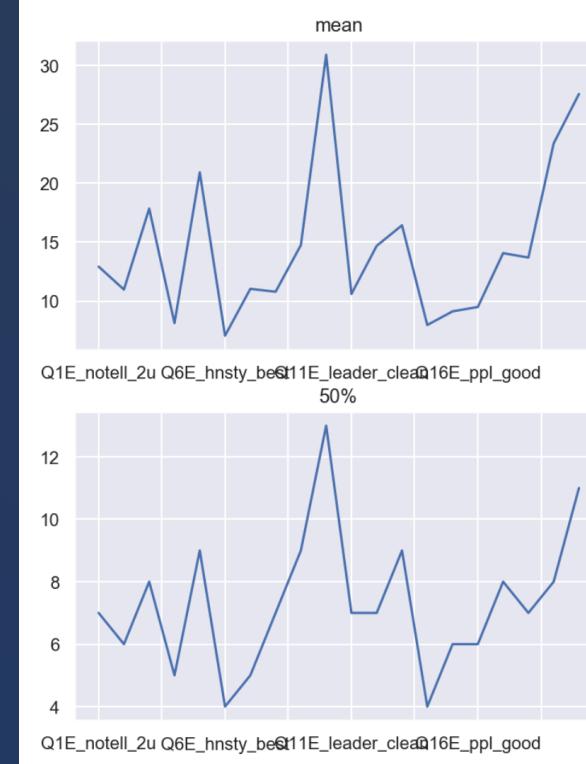
성향 테스트 결과 데이터의 일관성 유지를 위한 심리검사 방법론적 장치이므로 **연관 없는 정보**



Q1E~Q20E 마키아벨리즘 답변 시간



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제

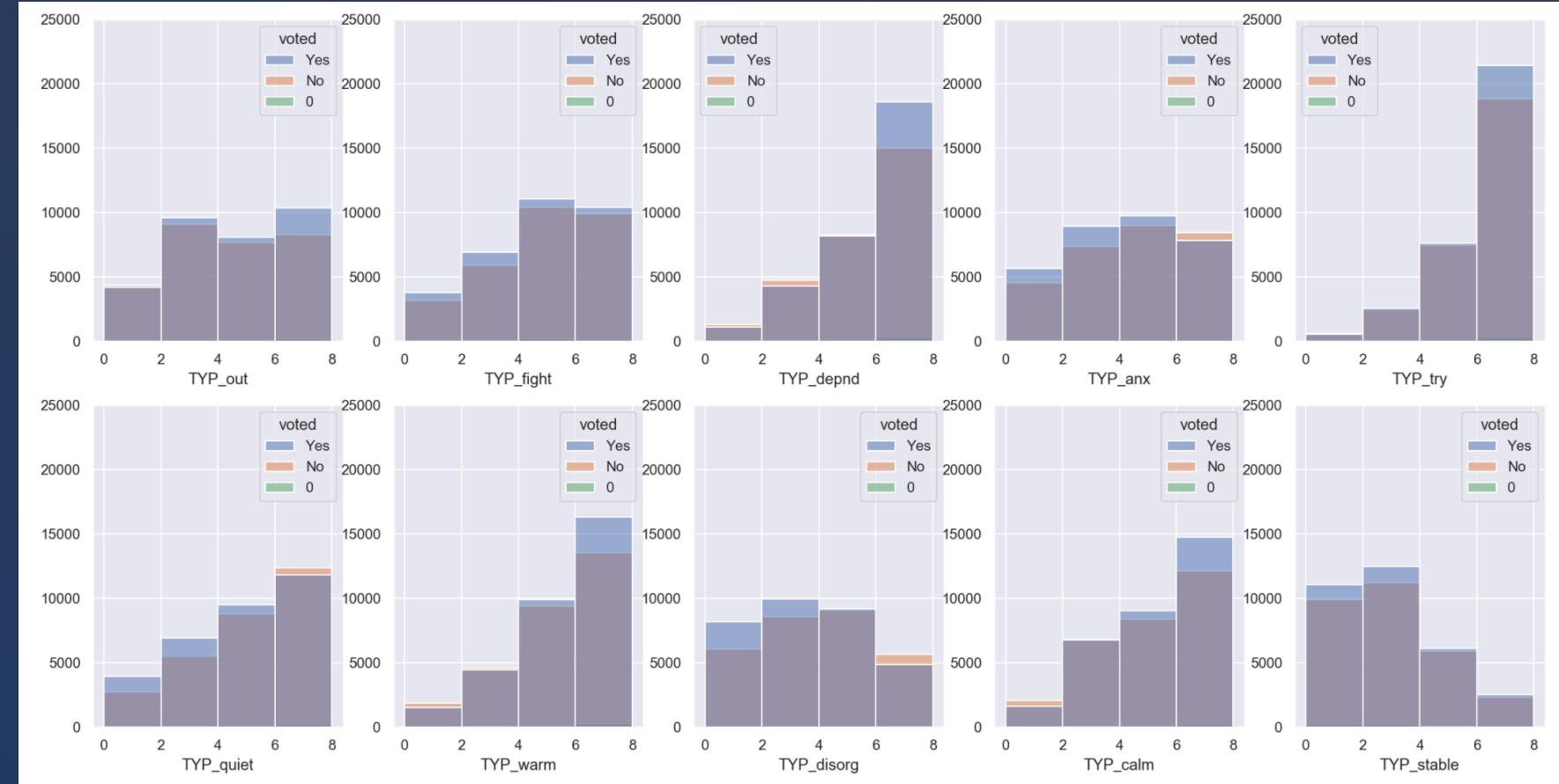


마키아벨리즘 테스트 질문지별 답변시간 : 유난히 길게 걸린 답변, 짧게 걸린 답변 있음
+ minimum 등 이상치 있음

TIPI1~TIPI20 참여자의 성격



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



성격이 어떤 경우, 투표참여자가 더 많은 경향??

VCL1~VCL16 단어 인지 여부 확인 질문



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



<VCL 문제>

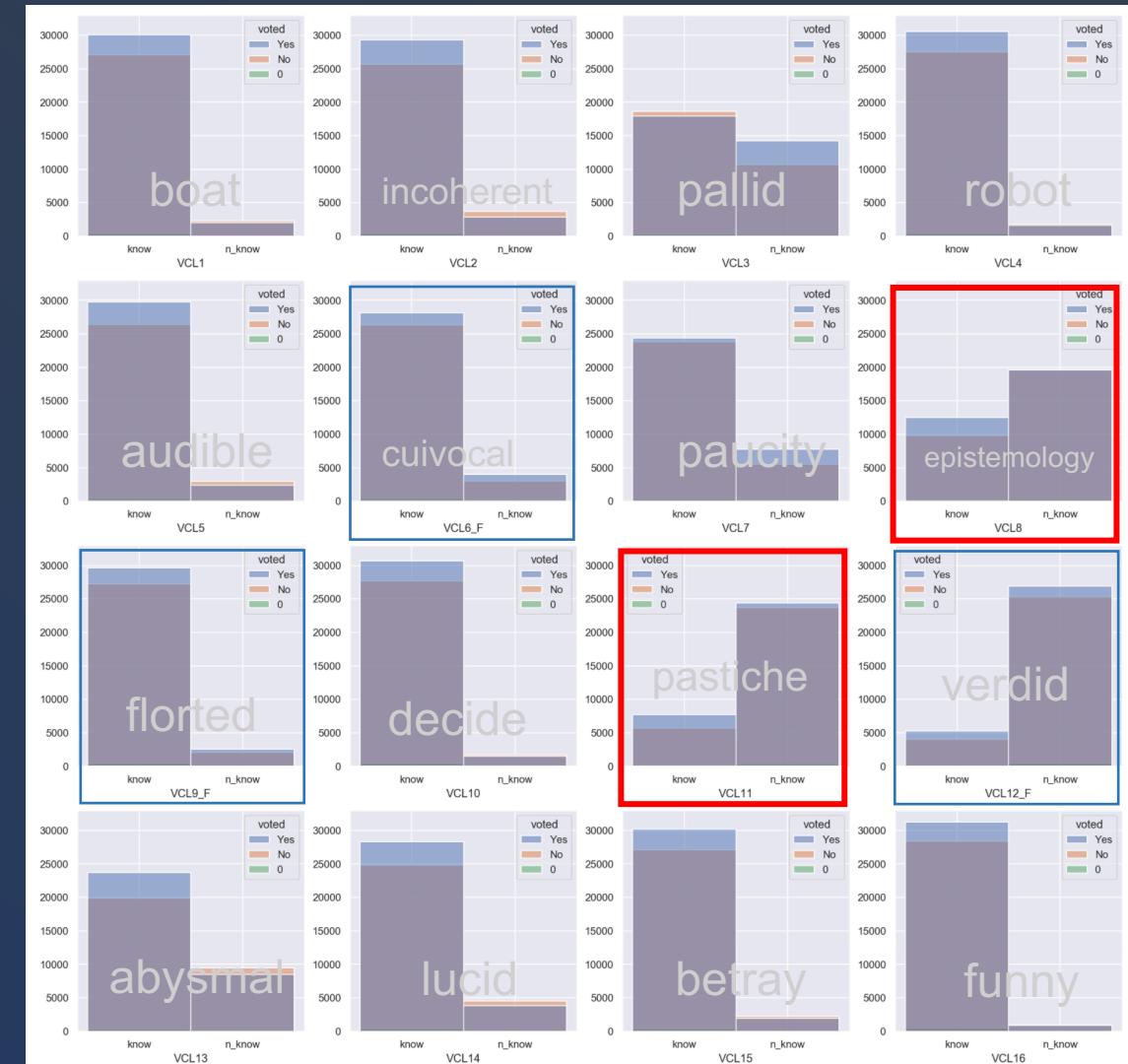
" 아래 단어의 뜻을 아십니까?"

16개 질문지 중 3개는 허구로
만들어낸 가상의 단어가 포함됨
-> cuivocal, florted, verdid



지어낸 단어마저 "알고있다"라고
거짓 답변하는 참여자 확인 목적

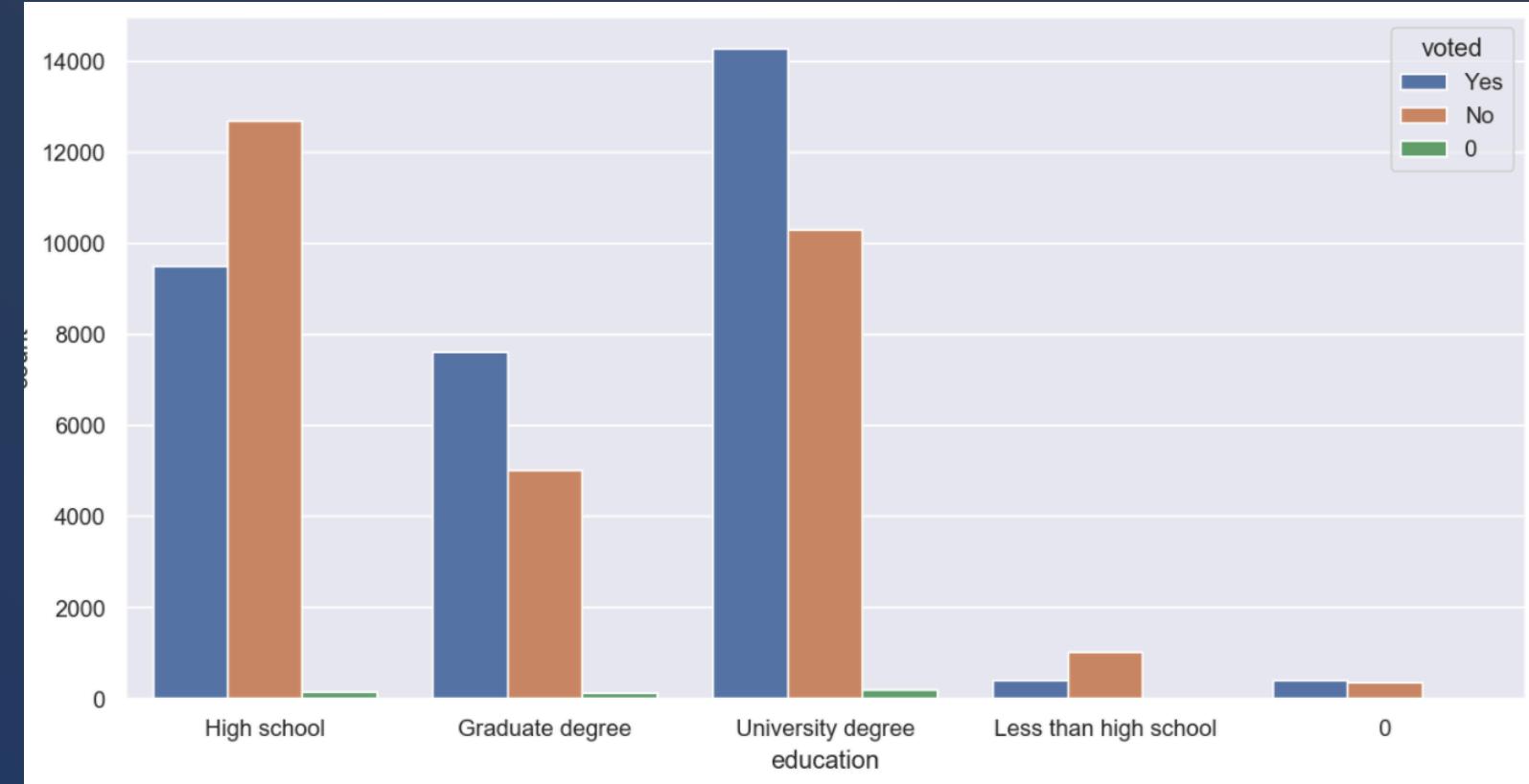
거짓답변을 한 참여자의 투표여부에는
차이가 미미. BUT 대다수가 모르는
단어를 안다고 답변한 참여자의
투표 참여율이 더 높음



education 참여자의 학력



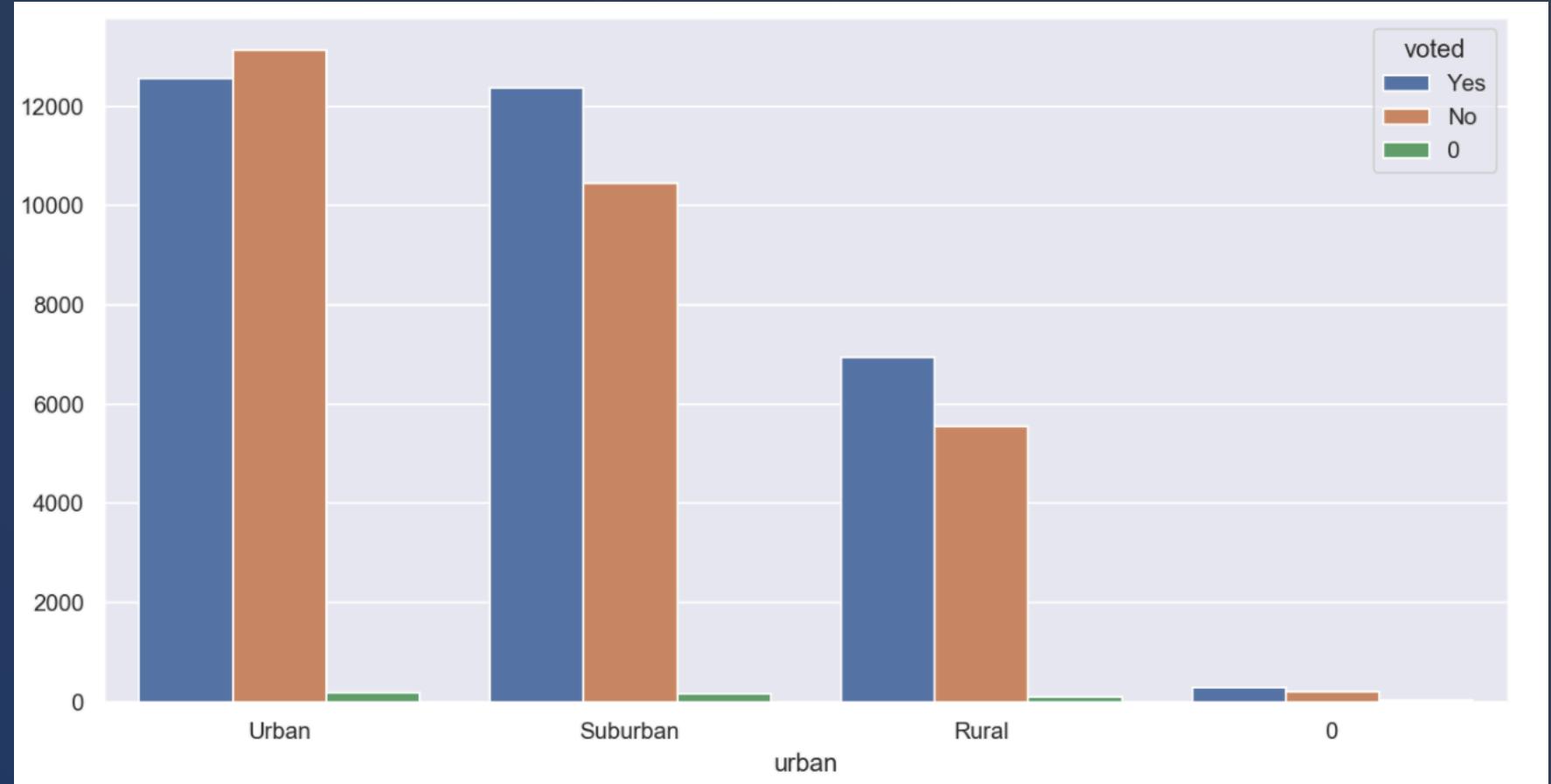
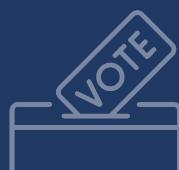
1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



오우



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제

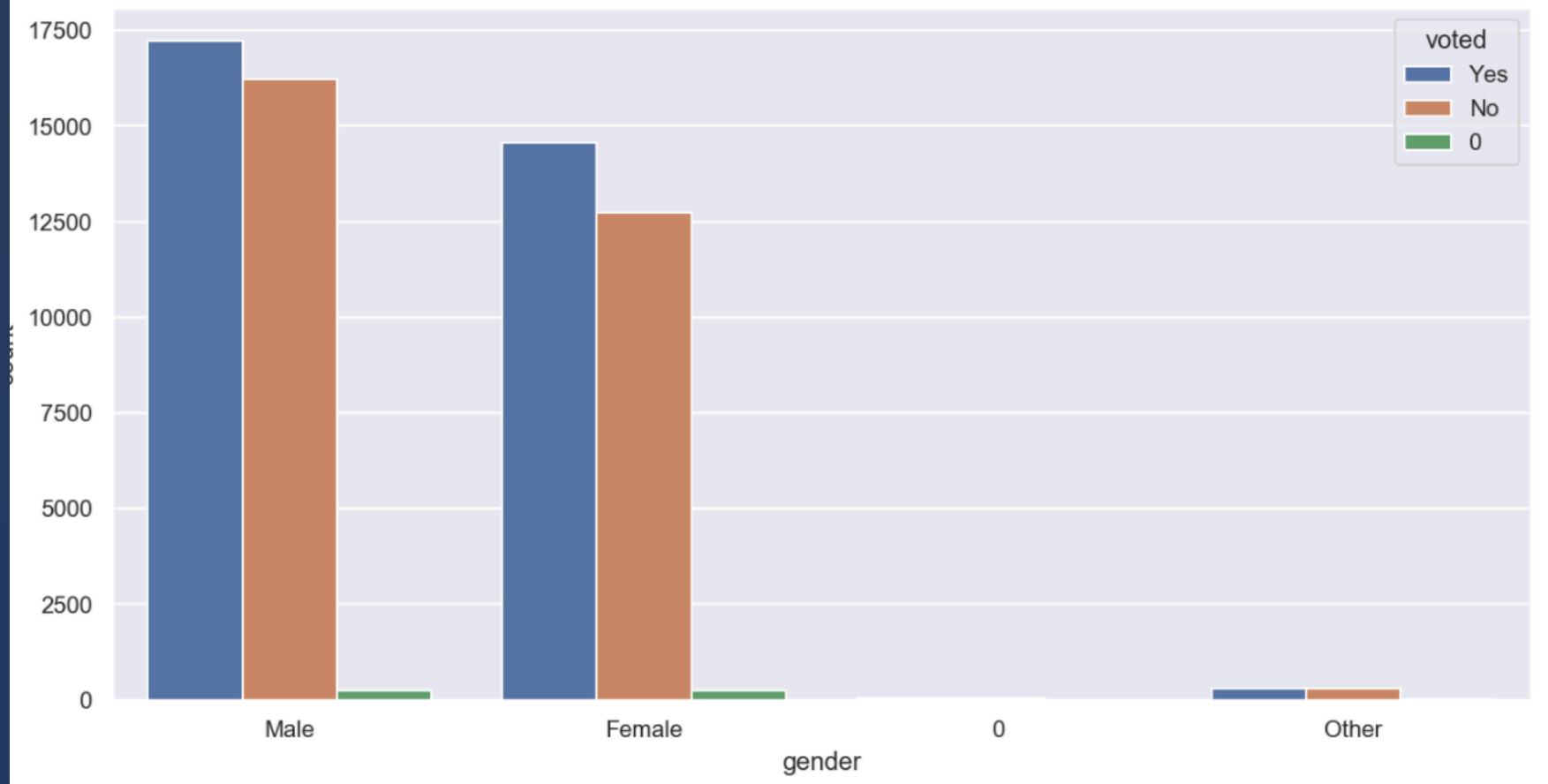


○미ㅏ○ㄴ

gender 참여자의 성별



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제

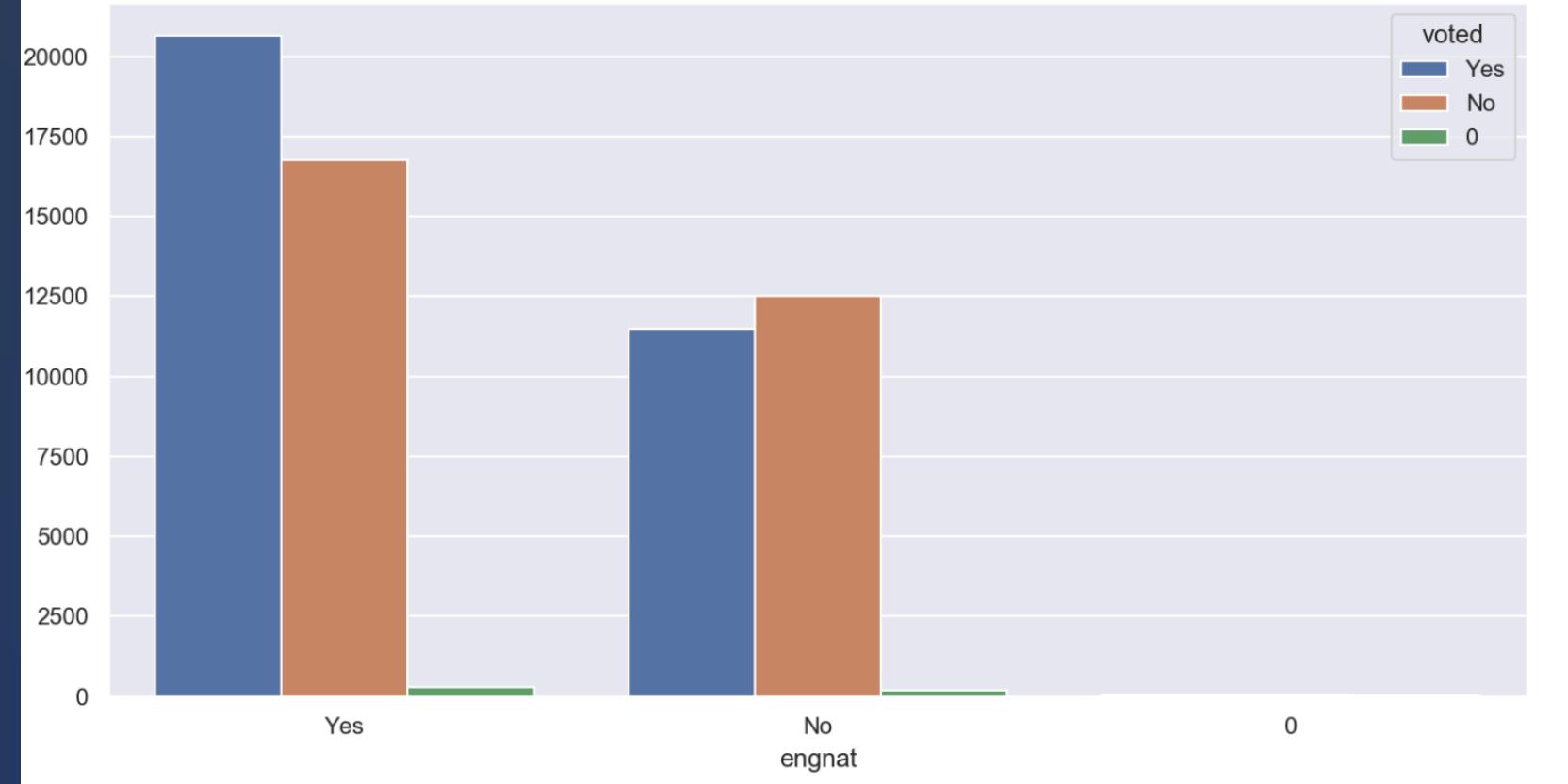


ㅋㅋ큐큐

engnat 참여자의 모국어



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



ㅋㅋ

#2. DATA INFO

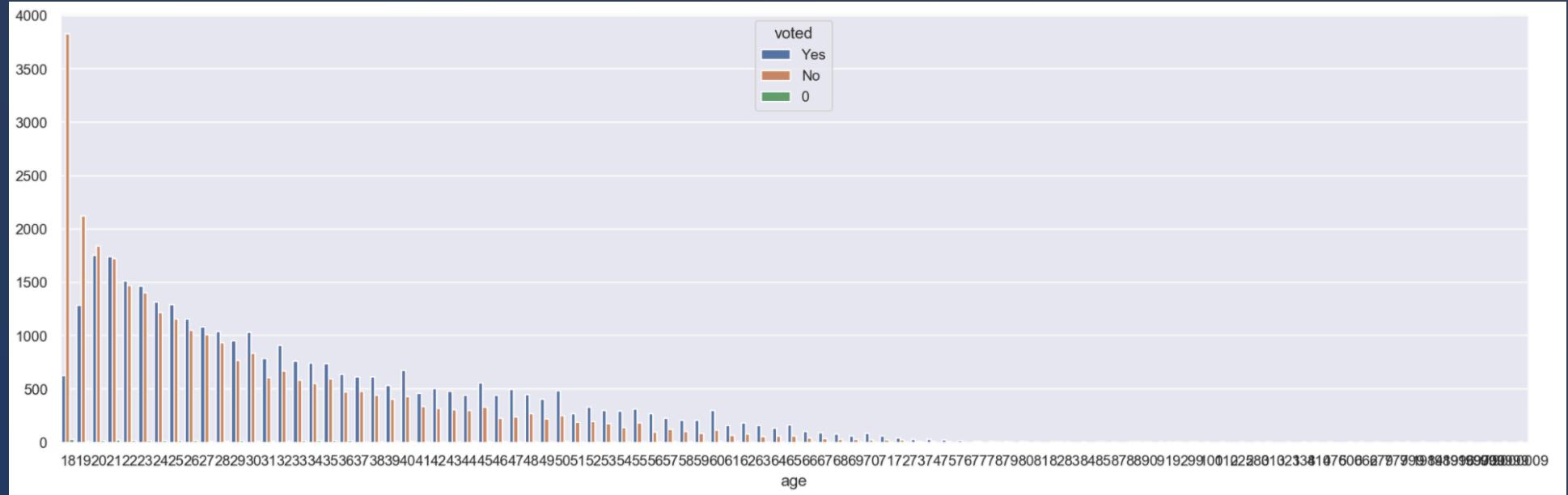


1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



EDA
voted &

age 참여자의 나이

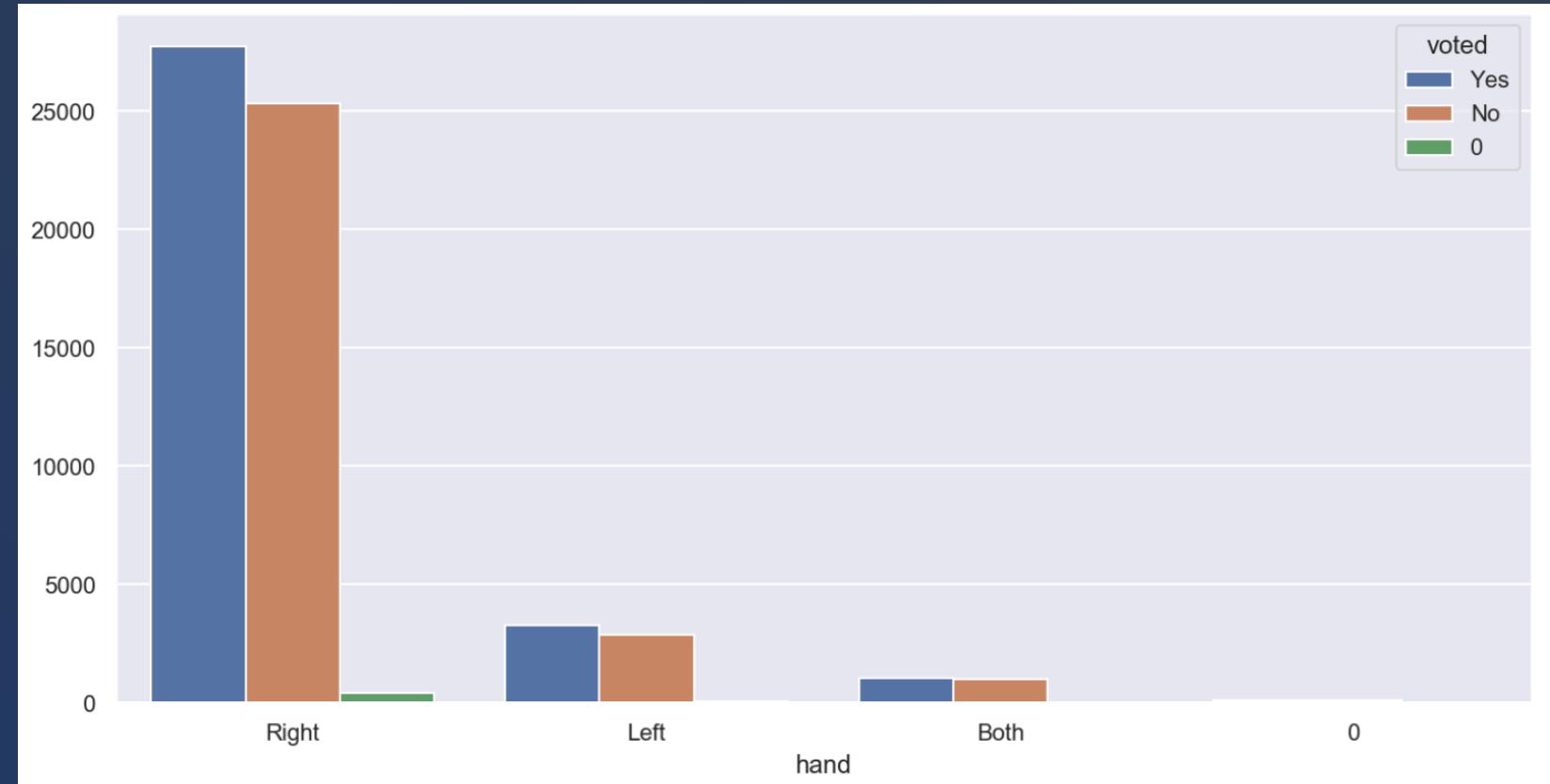


퍽

hand 참여자의 주사용 손



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제

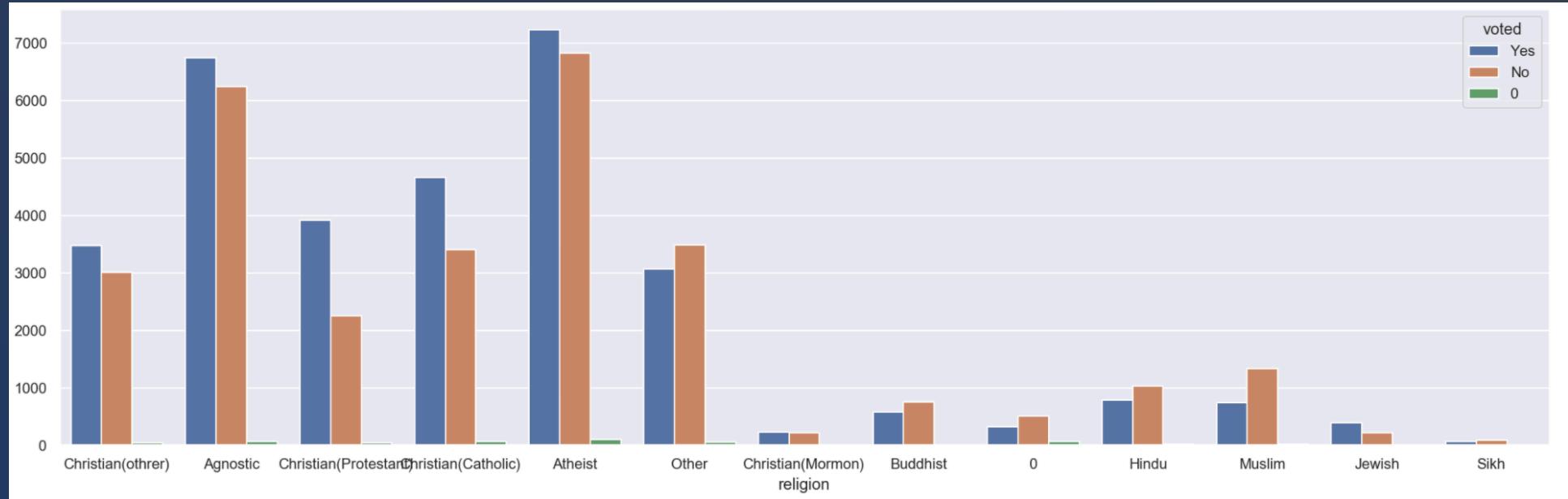


끼싱

religion 참여자의 종교



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



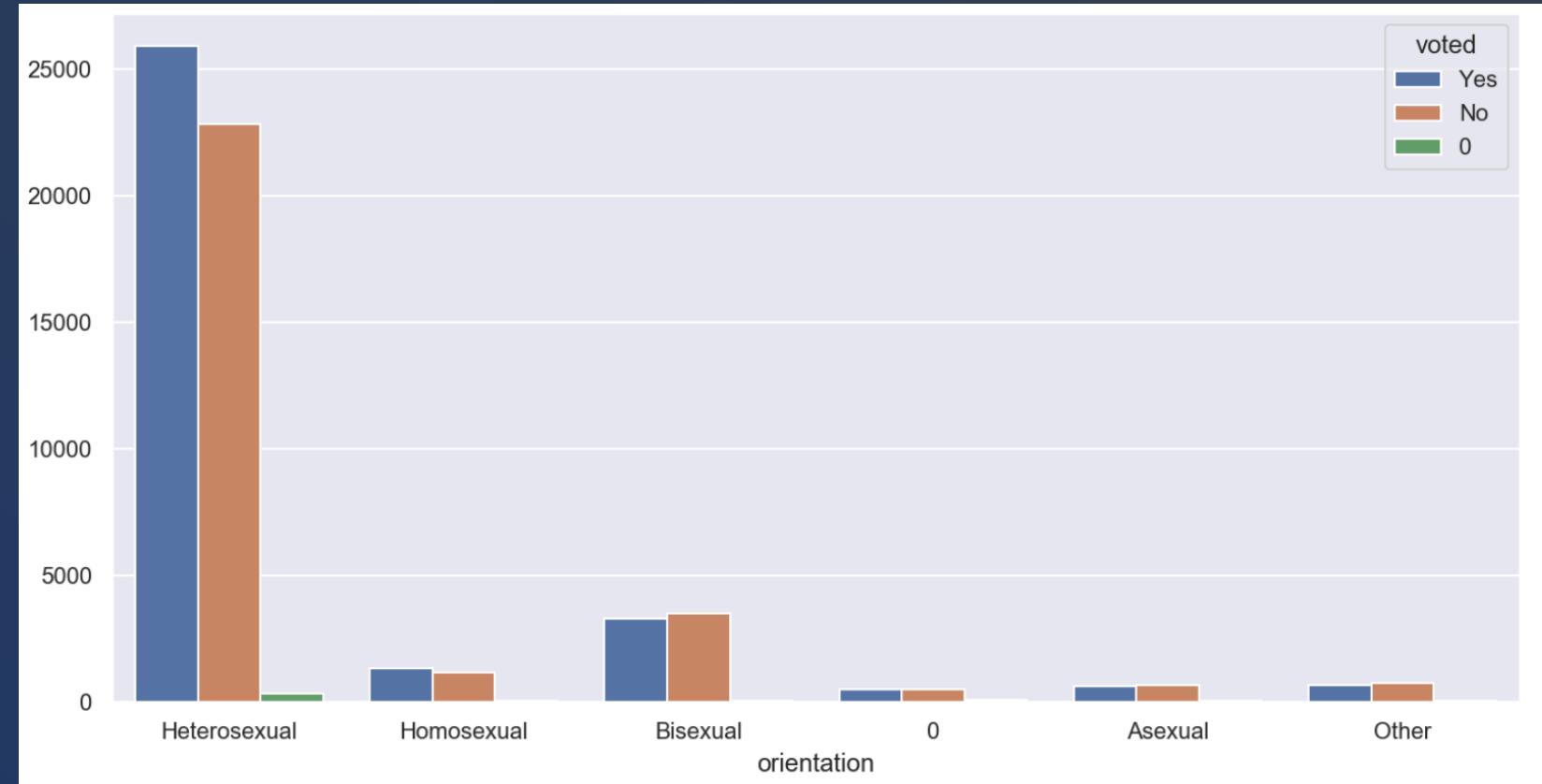
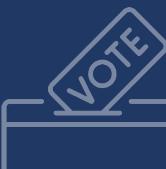
꺄항



orientation 참여자의 성정체성



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제

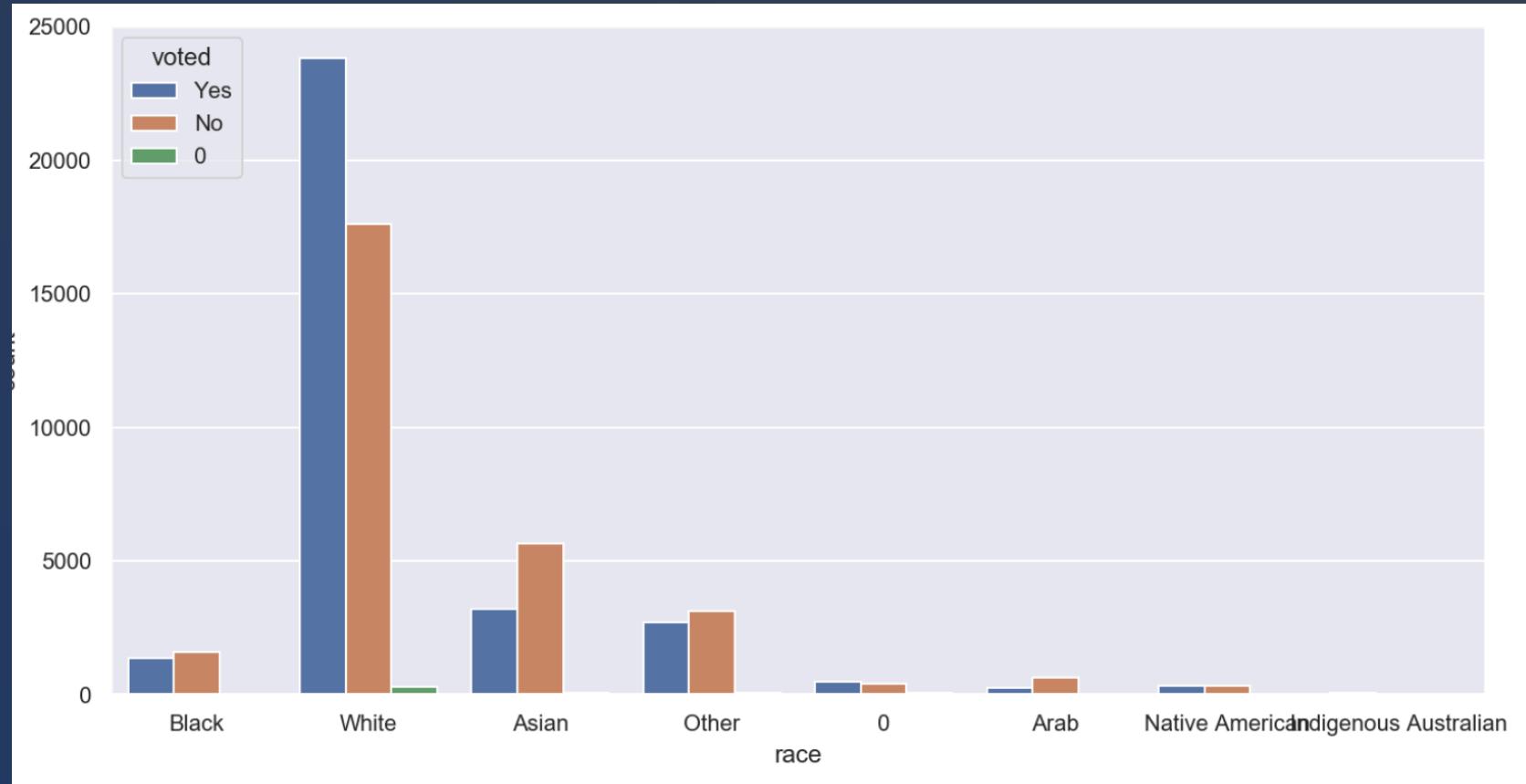


호영

race 참여자의 인종



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제

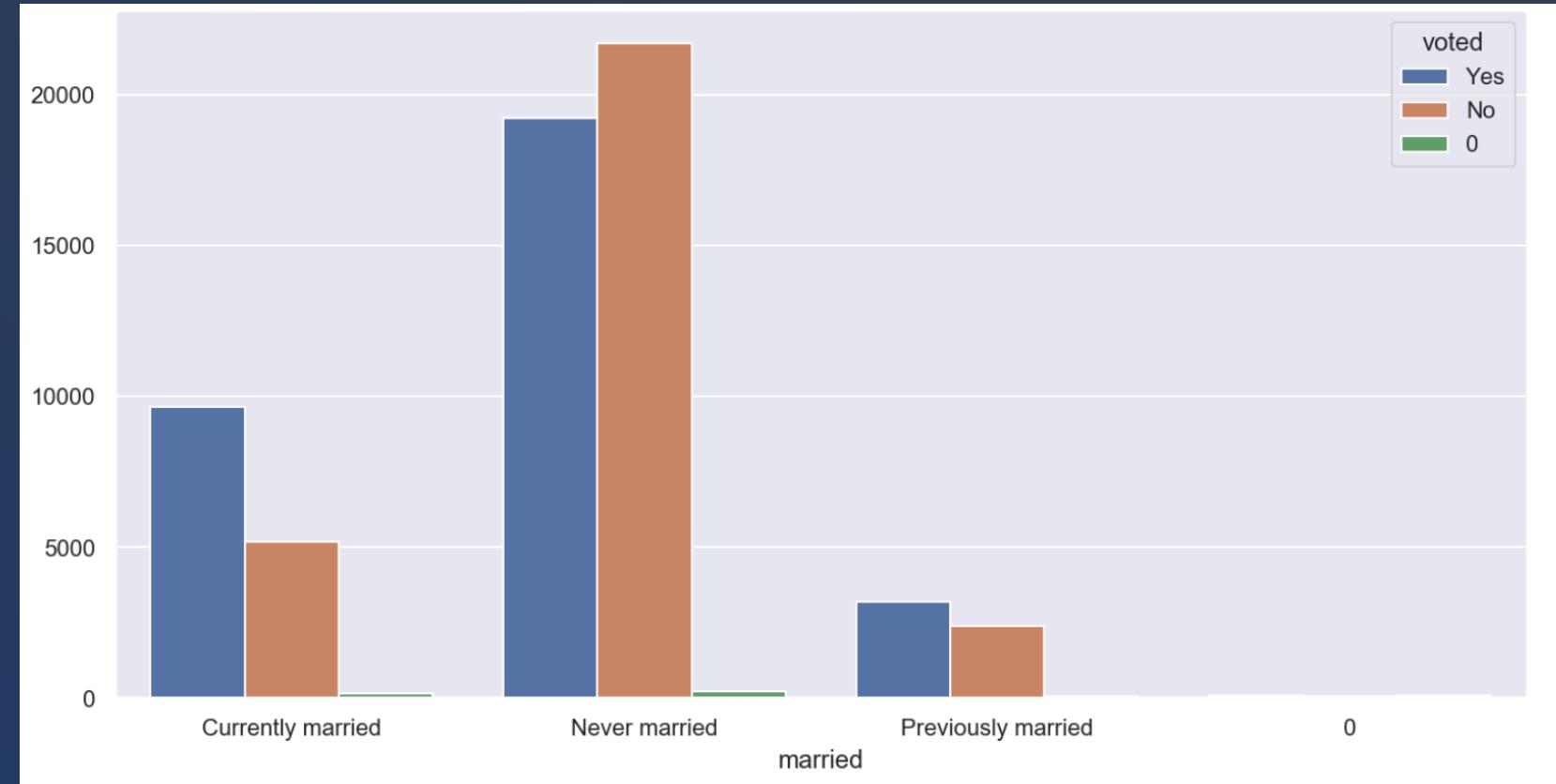
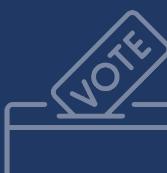


후잇

married 참여자의 결혼상태



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



호잇

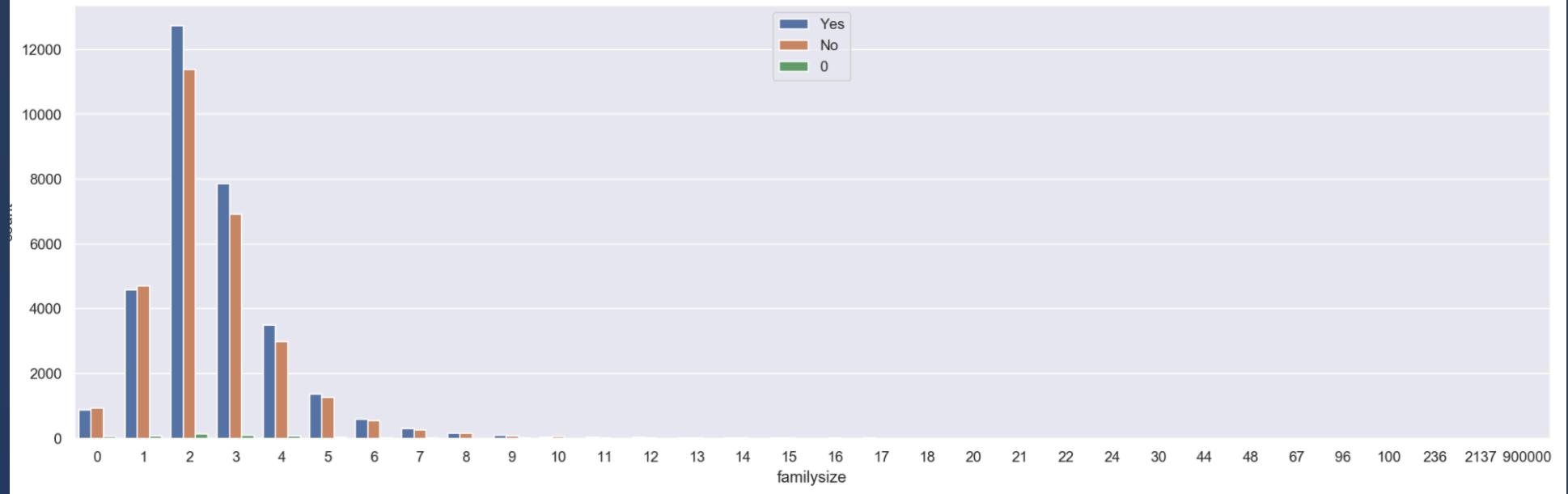
#2. DATA INFO



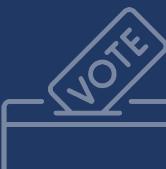
familysize 참여자의 형제자매수



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



호호



major 참여자의 학부 전공



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



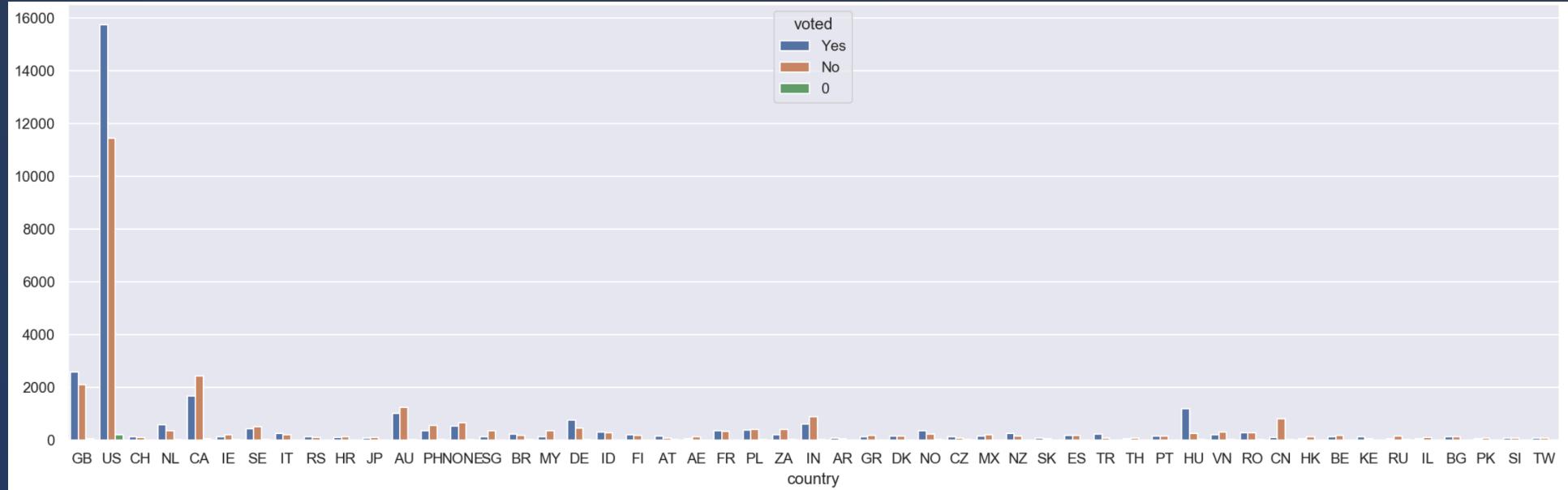
psychology	3367
business	1988
engineering	1734
english	1501
computer science	1435
	...
politic scienze	1
chinese language & literature program;	1
radio television film	1
training/education	1
journalism and mass media studies	1
Name: major, Length: 6380, dtype: int64	

헤헤

country 참여자의 국적



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



히히히

#2. DATA INFO



screenw~h 참여자의 스크린 사이즈



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



#2. DATA INFO



intro~survey elapse 참여자의 답변 시간



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제





Insights from EDA



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



분류모델 개발을 위한 핵심과제

1

수치 데이터 이상치 처리

age, familysize (참여자 직접입력 데이터)

QEn, intro~surveylapse (자동수집 시간 데이터)

2

유의미한 범주데이터 선별

country, race, TIPI1~10 등

(특정 범주에서의 voted 값이 특이한 데이터)

3

마키아벨리즘 관련 데이터에 가중치 부여

Q1~20, QE1~20, + score, T, V, M

(마키아벨리즘 테스트 답변 및 답변시간 데이터)

#2. DATA INFO



Insights from EDA



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



분류모델 개발을 위한 핵심과제

1

수치 데이터
이상치 처리

age, familysize (참여자 직접입력 데이터)

QEn, intro~surveylapse (자동수집 시간 데이터)

2

유의미한
범주데이터 선별

전처리

3

마키아벨리즘
관련 데이터에
가중치 부여

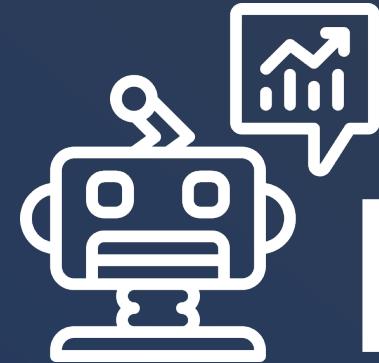
Q1~20, QE1~20

(마키아벨리즘 테스트 답변 및 답변시간 데이터)

#3. MODELING



1. 프로젝트 목표
2. 데이터 소개
- 3. 모델링 과정**
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제

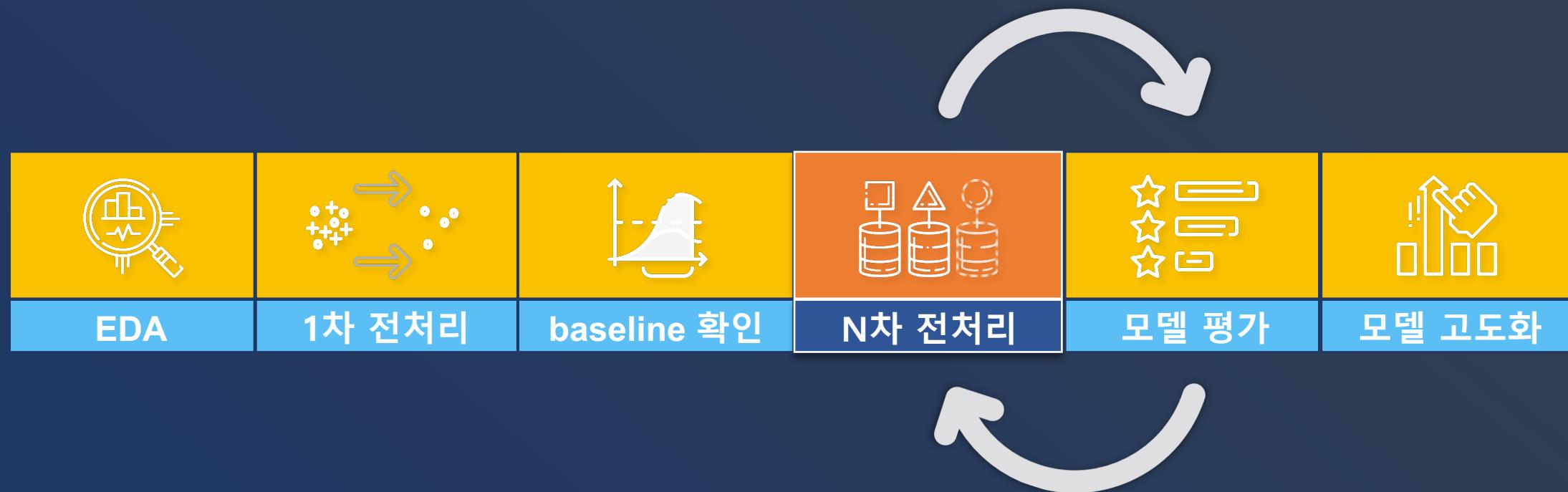


MODELING

try again & again...

#3. MODELING

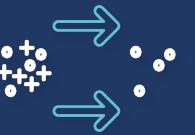
분류모델 개발 과정



#3. MODELING



EDA



1차 전처리



baseline 확인



N차 전처리



모델 평가



모델 고도화



1. 프로젝트 목표
2. 데이터 소개
- 3. 모델링 과정**
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



1

0, null 제거

 $\text{len}(0, \text{null data}) == 8,920$

2

선거 가능 연령
기준 이하 제거 $\text{len}(\text{age} < 18) == 11,521$

- 전세계 기준 선거 가능 연령 : 18세

3

마키아벨리즘 점수
합산 데이터 추가

지지 경향 답변 점수의 합
 $\text{score} =$
 +
 반대 경향 답변 점수의 역변환 값 ($6 - \text{점수}$)의 합

마키아체도별 점수
합산 데이터 추가 $T, V, M =$ 각 척도별 점수를 score 산출 방식으로 합산

4

밀리초 단위 환산

 $\text{round}(\text{밀리초} * 0.001)$



1차 전처리



baseline 확인



N차 전처리



모델 평가



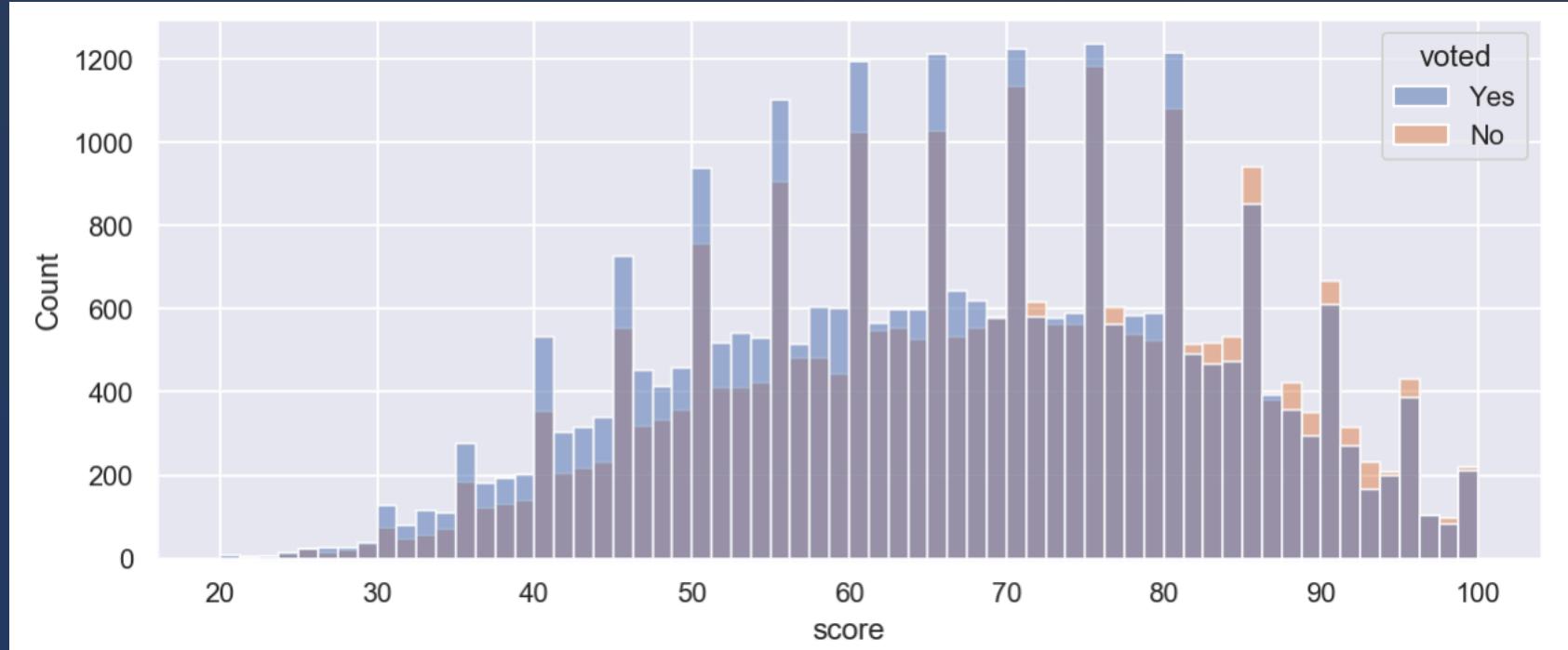
모델 고도화



- 1. 프로젝트 목표
- 2. 데이터 소개
- 3. 모델링 과정**
- 4. 모델 시연
- 5. 분류 결과 해석
- 6. 한계 및 과제



SCORE & voted 관계



마키아벨리즘 성향이 강할수록 투표에 참여하지 않음

#3. MODELING



EDA



1차 전처리



baseline 확인



N차 전처리



모델 평가



모델 고도화



1. 프로젝트 목표
2. 데이터 소개
- 3. 모델링 과정**
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



성능 향상!

(최대 성능 모델의 AUC값 기준)

Before (RAW)

	accuracy	AUC	precision	recall	f1
Ada	0.648153	0.638882	0.631036	0.803699	0.706978
GBC	0.655801	0.645951	0.634579	0.821078	0.715881
XGB	0.639680	0.633052	0.634174	0.750891	0.687615
LGBM	0.660861	0.652268	0.642883	0.805036	0.714879

After (1차 전처리)

	accuracy	AUC	precision	recall	f1
Ada	0.650402	0.644314	0.641925	0.759590	0.695818
GBC	0.656159	0.649482	0.643907	0.775907	0.703771
XGB	0.648209	0.644031	0.648809	0.723138	0.683960
LGBM	0.657255	0.651714	0.649821	0.756639	0.699174

0.6459



+▲ 0.0055



0.6494

#3. MODELING



EDA



1차 전처리



baseline 확인



N차 전처리



모델 평가



모델 고도화

- ① Feature importance 기준 컬럼 선별
- ② 수치형데이터 scaling
- ③ score로 가중치 생성
- ④ feature 정리 + 가중치



목표

주요 모델 4개에서 공통으로 Feature importance가 0인 컬럼 제거

- 1. 프로젝트 목표
- 2. 데이터 소개
- 3. 모델링 과정**
- 4. 모델 시연
- 5. 분류 결과 해석
- 6. 한계 및 과제

feature importance가 0인 컬럼 LIST





EDA



1차 전처리



baseline 확인



N차 전처리



모델 평가



모델 고도화



1. 프로젝트 목표
2. 데이터 소개
- 3. 모델링 과정**
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



- ① Feature importance 기준 컬럼 선별 ② 수치형데이터 scaling ③ score로 가중치 생성 ④ feature 정리 + 가중치

성능 아주매우조금 향상 !

(feature importance 값이 0인 컬럼 제거)

Before

	accuracy	AUC	precision	recall	f1
Ada	0.650402	0.644314	0.641925	0.759590	0.695818
GBC	0.656159	0.649482	0.643907	0.775907	0.703771
XGB	0.648209	0.644031	0.648809	0.723138	0.683960
LGBM	0.657255	0.651714	0.649821	0.756639	0.699174

After

	accuracy	AUC	precision	recall	f1
Ada	0.650402	0.644314	0.641925	0.759590	0.695818
GBC	0.656250	0.649578	0.643999	0.775907	0.703826
XGB	0.648209	0.644031	0.648809	0.723138	0.683960
LGBM	0.657255	0.651714	0.649821	0.756639	0.699174

0.6494

+▲ 0.0001

0.6495



EDA



1차 전처리



baseline 확인



N차 전처리



모델 평가



모델 고도화



1. 프로젝트 목표
2. 데이터 소개
- 3. 모델링 과정**
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제

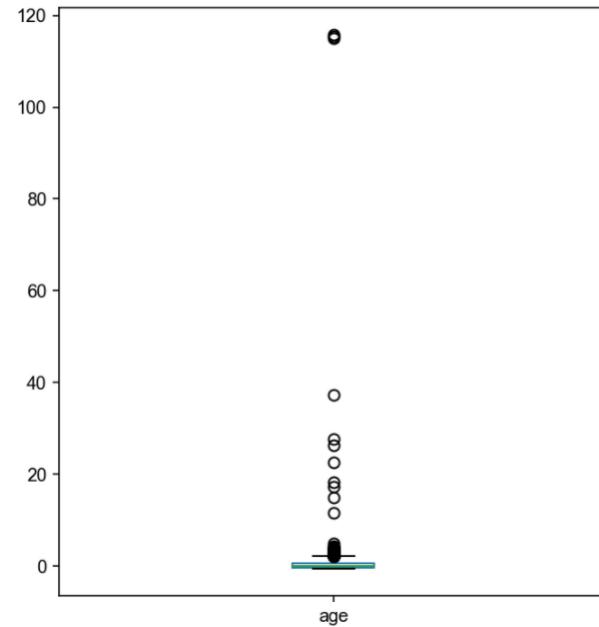
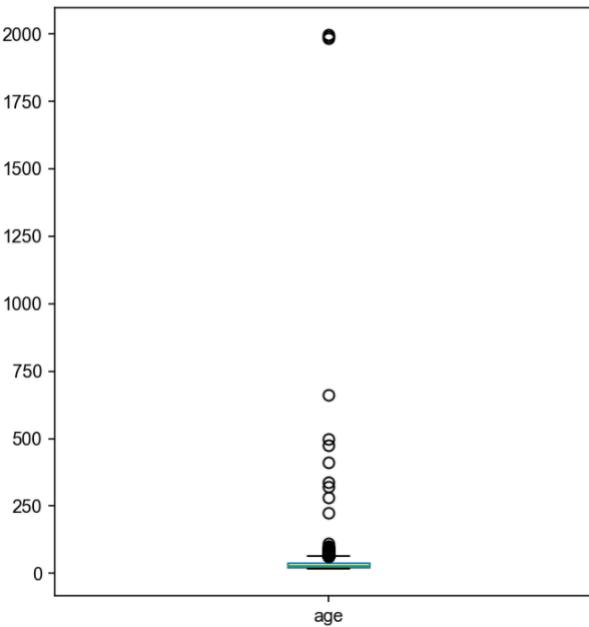


- ① Feature importance 기준 컬럼 선별
- ② 수치형데이터 scaling
- ③ score로 가중치 생성
- ④ feature 정리 + 가중치

이상치를 포함한 수치형데이터에 Robust scaling

```
df_18['age'].describe(), df_18_cp['age'].describe()
```

```
(count    54718.000000
 mean     31.739537
 std      19.666469
 min     18.000000
 25%    22.000000
 50%    28.000000
 75%    39.000000
 max    1997.000000
Name: age, dtype: float64,
count    54718.000000
mean     0.219973
std      1.156851
min    -0.588235
25%   -0.352941
50%    0.000000
75%    0.647059
max    115.823529
Name: age, dtype: float64)
```



울랄라



EDA



1차 전처리



baseline 확인



N차 전처리



모델 평가



모델 고도화



1. 프로젝트 목표
2. 데이터 소개
- 3. 모델링 과정**
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



- ① Feature importance 기준 컬럼 선별 ② 수치형데이터 scaling ③ score로 가중치 생성 ④ feature 정리 + 가중치

성능 하락 ↘—↘

(이상치가 있는 수치형 데이터 robust scaling)

Before

	accuracy	AUC	precision	recall	f1
Ada	0.650402	0.644314	0.641925	0.759590	0.695818
GBC	0.656159	0.649482	0.643907	0.775907	0.703771
XGB	0.648209	0.644031	0.648809	0.723138	0.683960
LGBM	0.657255	0.651714	0.649821	0.756639	0.699174

After

	accuracy	AUC	precision	recall	f1
Ada	0.648940	0.642983	0.641332	0.755772	0.693865
GBC	0.653417	0.647129	0.643440	0.766186	0.699469
XGB	0.645285	0.641486	0.648163	0.713418	0.679227
LGBM	0.652778	0.648680	0.653036	0.726263	0.687705

0.6494



+▼ 0.0023



0.6471

#3. MODELING



EDA



1차 전처리



baseline 확인



N차 전처리



모델 평가



모델 고도화



1. 프로젝트 목표
2. 데이터 소개
- 3. 모델링 과정**
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



- ① Feature importance 기준 컬럼 선별 ② 수치형데이터 scaling ③ score로 가중치 생성 ④ feature 정리 + 가중치

#3. MODELING



EDA



1차 전처리



baseline 확인



N차 전처리



모델 평가



모델 고도화



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



- ① Feature importance 기준 컬럼 선별 ② 수치형데이터 scaling ③ score로 가중치 생성 ④ feature 정리 + 가중치

성능 향상 ?

(score 컬럼)

Before

After

0.6494

+▲ 0.0001

0.6495

#3. MODELING



EDA



1차 전처리



baseline 확인



N차 전처리



모델 평가



모델 고도화

- ① Feature importance 기준 컬럼 선별
- ② 수치형데이터 scaling
- ③ score로 가중치 생성
- ④ feature 정리 + 가중치



1. 프로젝트 목표
2. 데이터 소개
- 3. 모델링 과정**
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



feature importance 0인 컬럼 제거 & SCORE에 가중치 컬럼 추가



EDA



1차 전처리



baseline 확인



N차 전처리



모델 평가



모델 고도화



- ① Feature importance 기준 컬럼 선별
- ② 수치형데이터 scaling
- ③ score로 가중치 생성
- ④ feature 정리 + 가중치

feature importance 0인 컬럼 제거 + Score 가중치컬럼 추가 전후 성능 비교

1. 프로젝트 목표
2. 데이터 소개
- 3. 모델링 과정**
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제

성능 하락 π—π

(score)

Before		
feature	GBC	0.539382 0.527113
score	GBC	0.539382 0.527113

+ ▾ 0.0055



After		
combi	GBC	0.539382 0.527113





EDA



1차 전처리



baseline 확인



N차 전처리



모델 평가



모델 고도화



1. 프로젝트 목표
2. 데이터 소개
- 3. 모델링 과정**
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



N차 전처리 4가지 DATASET 성능 비교

① feature importance 기준 컬럼 선별

	accuracy	AUC	precision	recall	f1
Ada	0.648153	0.638882	0.631036	0.803699	0.706978
GBC	0.655801	0.645951	0.634579	0.821078	0.715881
XGB	0.639680	0.633052	0.634174	0.750891	0.687615
LGBM	0.660861	0.652268	0.642883	0.805036	0.714879

② 수치형데이터 scaling

	accuracy	AUC	precision	recall	f1
Ada	0.650402	0.644314	0.641925	0.759590	0.695818
GBC	0.656159	0.649482	0.643907	0.775907	0.703771
XGB	0.648209	0.644031	0.648809	0.723138	0.683960
LGBM	0.657255	0.651714	0.649821	0.756639	0.699174

③ SCORE에 가중치 부여

	accuracy	AUC	precision	recall	f1
Ada	0.648153	0.638882	0.631036	0.803699	0.706978
GBC	0.655801	0.645951	0.634579	0.821078	0.715881
XGB	0.639680	0.633052	0.634174	0.750891	0.687615
LGBM	0.660861	0.652268	0.642883	0.805036	0.714879

④ feature 정리 + SCORE 가중치

	accuracy	AUC	precision	recall	f1
Ada	0.650402	0.644314	0.641925	0.759590	0.695818
GBC	0.656159	0.649482	0.643907	0.775907	0.703771
XGB	0.648209	0.644031	0.648809	0.723138	0.683960
LGBM	0.657255	0.651714	0.649821	0.756639	0.699174

0.6388



+ ▲ 0.0055

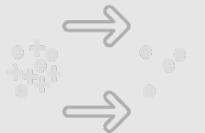


0.6443

#3. MODELING



EDA



1차 전처리



baseline 확인



N차 전처리



모델 평가



모델 고도화



- 1. 프로젝트 목표
- 2. 데이터 소개
- 3. 모델링 과정**
- 4. 모델 시연
- 5. 분류 결과 해석
- 6. 한계 및 과제

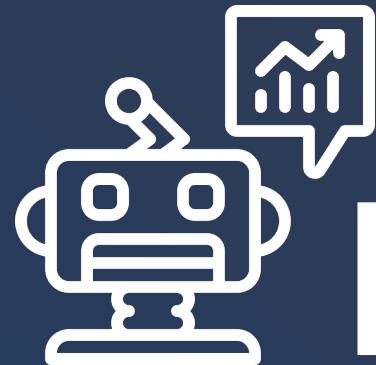


하이퍼파라미터 튜닝

#4. PREDICT



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
- 4. 모델 시연**
5. 분류 결과 해석
6. 한계 및 과제



PREDICT

who has voted?

#4. PREDICT

마키아벨리안 미국인은 투표했을까?



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
- 4. 모델 시연**
5. 분류 결과 해석
6. 한계 및 과제



가상의 인물 설정 후 분류모델 이용

#4. PREDICT

우리 분류모델이 판단한 투표참여자 성향



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
- 4. 모델 시연**
5. 분류 결과 해석
6. 한계 및 과제



???

#5. ANALYSIS



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
- 5. 분류 결과 해석**
6. 한계 및 과제



ANALYSIS

Do Not Touch...?

인간의 손때가 묻으면 더러워지는 머신모델..?



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
- 5. 분류 결과 해석**
6. 한계 및 과제



raw_data >> 1차 전처리 >> N차 전처리 >> 하이퍼파라미터
단계별 모델 성능 비교

인간의 손때가 묻으면 더러워지는 머신모델..?



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
- 5. 분류 결과 해석**
6. 한계 및 과제



`raw_data >> 1차 전처리 >> N차 전처리 >> 하이퍼파라미터
단계별 모델 confusion matrix 비교`

인간의 손때가 물으면 더러워지는 머신모델..?



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
- 5. 분류 결과 해석**
6. 한계 및 과제



TEST DATA vs PREDICTED DATA 비교??

#6. REVIEW



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



 **REVIEW**
where to go...

#6. REVIEW



1. 프로젝트 목표
2. 데이터 소개
3. 모델링 과정
4. 모델 시연
5. 분류 결과 해석
6. 한계 및 과제



머신러닝 분류모델, 잘 만드는 방법?

- 1) 이상치 처리 문제

: age, family size, 각종 소요시간 컬럼의 이상치 무조건 제거 x
-> feature importance를 확인하여 <scaling>으로 데이터 손실 최소화

- 2) 가중치 부여 방식 문제

: feature importance의 중요도 만큼 가중치 부여한 경우
: 임의의 feature에 의도적으로 가중치 부여한 경우

- 3) Auto_ML 사용 문제

: 활용도 한계로 인한 문제
-> 동일 모델에 동일 하이퍼파라미터를 적용하여 raw_data로 뽑은 성능에도 차이가 있음

- 4) 하이퍼파라미터

:

->

- 5) 향후 시도 과제 : 앙상블 시도