빅데이터 분석 기획서

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 분석 과제명 | 마키아벨리즘 심리테스트를 활용하여 테스트 참가자의 국가 선거 투표 여부 예측 | | | | |
| 분석 기간 | 2020.10.30 ~ 2020.11.13 | | | | |
| 입주자(팀) | 소속기관 | KIC캠퍼스 | | 소재지 | 서울특별시 강남구 역삼동 819-10 세경빌딩3층 |
| 구분 | 성명 | 이메일 | 연락처 | 역할 |
| 팀장 | 윤혜정 | yoonhe3721@gmail.com | 01040243721 | 데이터 전처리, 발표자료작성 |
| 팀원1 | 이재훈 | jaehoon.ds@gmail.com | 01058829090 | 데이터 전처리, 모델링 |
| 팀원 2 | 최영은 | chloe\_ch95@naver.com | 01056335096 | 모델링, 발표자료작성 |
| 팀원 3 | 이지현 | wlgus6313@nate.com | 01063135374 | 모델링, 발표 |
| 팀원 4 | 정성호 | harryjeong27@gmail.com | 01032994763 | 데이터 분석, 아이디어뱅크 |
| 연구 배경 및 목적 | 심리학 테스트의 범주가 넓어짐에 따라 해당 영역의 데이터 분석 방법에 대한 탐구도 심화되고 있다.  본 연구에서는 국가 선거 투표자/미투표자의 심리학적 성향을 알아본다. 개개인이 가진 마키아벨리즘의 정도를 알아보는 테스트인 마키아벨리즘 테스트를 활용하여 테스트 참가자의 국가 선거 투표 여부를 예측해본다. | | | | |
| 연구 및 분석 계획 | 1. 마키아벨리즘 테스트, 기타 설문내용(인구통계학적 정보, TP테스트 등)을 다각적으로 검토하여 1) 상호 간의 상관관계, 2) 투표율과의 상관관계 분석 실행 2. 위 상관관계 결과를 바탕으로 상관관계가 있다고 느껴질 경우 로지스틱 회귀분석을 통해 회귀계수 추정하여 변수 가중치 넣고, 상관관계 및 분류분석 실행 3. 분류분석 결과를 바탕으로 실제 투표여부 예측 4. AutoML 활용하여 가장 높은 예측력 보이는 모델 선택   (CatBoost Classfier, Gradient Boosting Classifer, LGBM) | | | | |
| 결론 및 한계점 | Auto ML 을 활용하여 최종 77.996의 예측률을 얻었으나3가지의 한계점을 도출하였음  첫째, 마키아벨리즘 심리테스트 척도의 신뢰도 문제점.  - 인구통계학적인 변인에 따라서 신뢰도의 차이가 심한 편으로 척도를 신뢰하기 어려움  둘째, 외부 데이터 활용 제한  -외부데이터 사용이 불가하여 마키아밸리즘 테스트 만을 이용하여 예측을 하는데 무리가 있음  셋째. 주최사 측의 데이터 임의 수정 및 삭제  -기존의 설문지에는 전공 선택 및 성적취향 문항도 있었으나 주최측에서 임의로 삭제함  이러한 한계점을 극복할 수 있었다면 더 좋은 예측력을 얻을 수 있을 것이라는 아쉬움이 있음. | | | | |

데이터

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 번호 | 데이터셋명 | 주요필드설명 | 데이터의 종류 | 원천 |
| 1 | train (훈련용데이터) | 마키아벨리즘 심리 테스트 항목 | CSV | DACON(주최사) 제공 |
| 2 | test\_x (정답용데이터) | 마키아벨리즘 심리 테스트 항목 (voted 제외) | DACON(주최사) 제공 |
| 3 |  |  |  |  |

[ 참고 : 대회 규칙상 외부 자료 사용 불가 ]

일정관리

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 10월 3주 | 10월 4주 | 11월 1주 | 11월 2주 |
| 프로젝트 기획 |  |  |  |  |
| 데이터 수집 및 정제 |  |  |  |  |
| 분석 알고리즘 설계 |  |  |  |  |
| 테스트 및 보완 |  |  |  |  |
| 발표자료 및 보고서 |  |  |  |  |