와인 품질 데이터 머신러닝

전주 ICT 이노베이션 스퀘어 온라인코딩교육

2조 양명훈 오한나 김아영 오창현

와인 품질 데이터 머신러닝

CONTENTS

1. 프로젝트 제안 배경 및 목적
2. 프로젝트 프로세스
3. 프로젝트 단계별 내용
4. 프로젝트 후기

와인 품질 데이터 머신러닝

1. 제안 배경 및 목적

- 배운 머신러닝 내용을 실제 데이터에 활용해보고 익숙해지는 것을 목표로 함
- 실생활과 관련된 데이터를 주제로 활용하면 흥미로울 것
- 와인을 구성하는 요소 데이터를 가지고 와인의 품질 예측

2. 프로젝트 프로세스

- 주제 선정 및 데이터 수집
- EDA(탐색적 데이터 분석)
- 머신러닝 모델 적용 및 학습
- 데이터 전처리 후 모델 학습
- 시각화

와인 품질 데이터 머신러닝

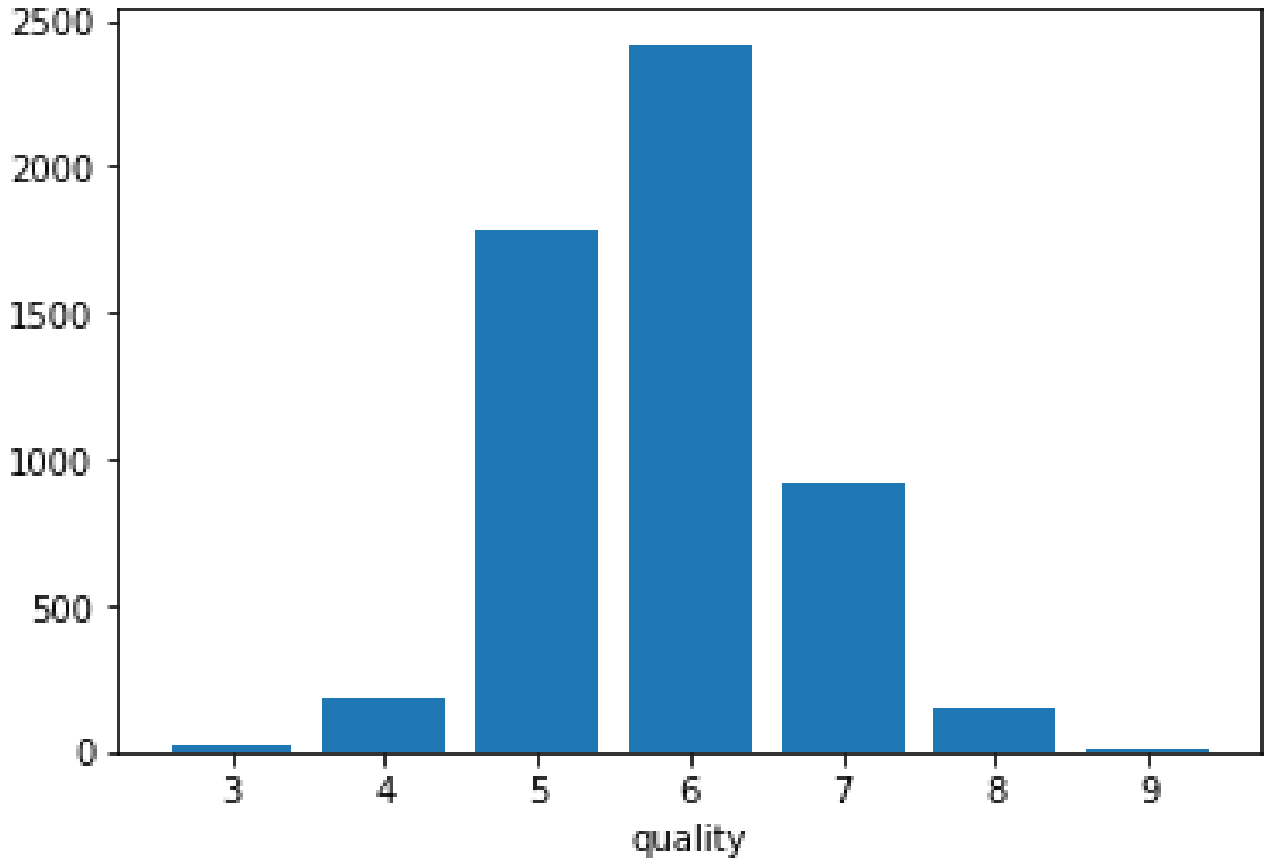
3. 프로젝트 단계별 내용

	quality	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	type
0	5	5.6	0.695	0.06	6.8	0.042	9.0	84.0	0.99432	3.44	0.44	10.2	white
1	5	8.8	0.610	0.14	2.4	0.067	10.0	42.0	0.99690	3.19	0.59	9.5	red
2	5	7.9	0.210	0.39	2.0	0.057	21.0	138.0	0.99176	3.05	0.52	10.9	white
3	6	7.0	0.210	0.31	6.0	0.046	29.0	108.0	0.99390	3.26	0.50	10.8	white
4	6	7.8	0.400	0.26	9.5	0.059	32.0	178.0	0.99550	3.04	0.43	10.9	white
5	6	6.0	0.190	0.37	9.7	0.032	17.0	50.0	0.99320	3.08	0.66	12.0	white
6	5	6.1	0.220	0.49	1.5	0.051	18.0	87.0	0.99280	3.30	0.46	9.6	white
7	6	7.1	0.380	0.42	11.8	0.041	32.0	193.0	0.99624	3.04	0.49	10.0	white
8	5	6.8	0.240	0.31	18.3	0.046	40.0	142.0	1.00000	3.30	0.41	8.7	white
9	5	6.8	0.390	0.35	11.6	0.044	57.0	220.0	0.99775	3.07	0.53	9.3	white

총 13 개의 column으로 구성

quality	품질(0~10, 높을 수록 좋음)
fixed acidity	고정 산도
volatile acidity	휘발성 산도
citric acid	시트르산(구연산)
residual sugar	잔당 : 발효 후 와인 속에 남아있는 당분
chlorides	염화물
free sulfur dioxide	독립 이산화황
total sulfur dioxide	총 이산화황
density	밀도
pH	수소이온농도
sulphates	황산염
alcohol	도수
type	종류

와인 품질 데이터 머신러닝

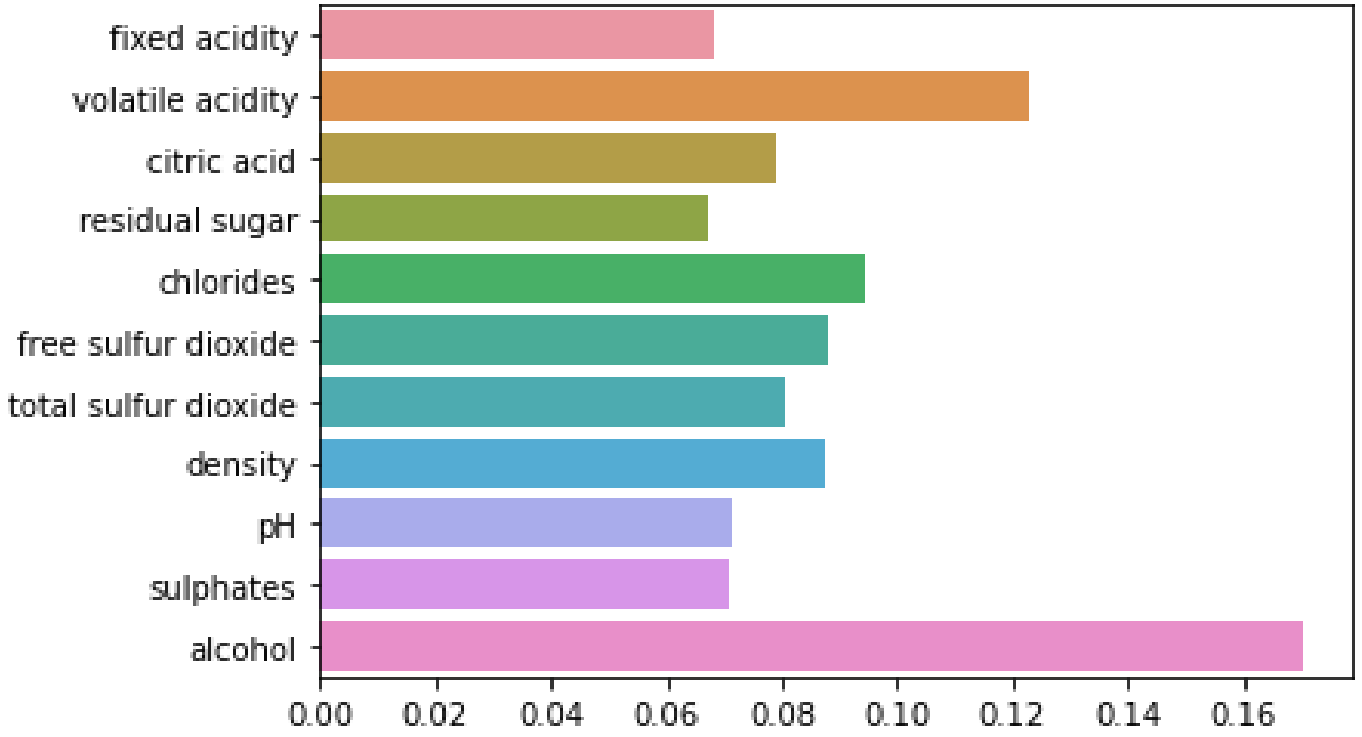


<Quality 데이터 분포도>

- quality는 3~9로 구성
- 데이터 간 분포 차이 큼

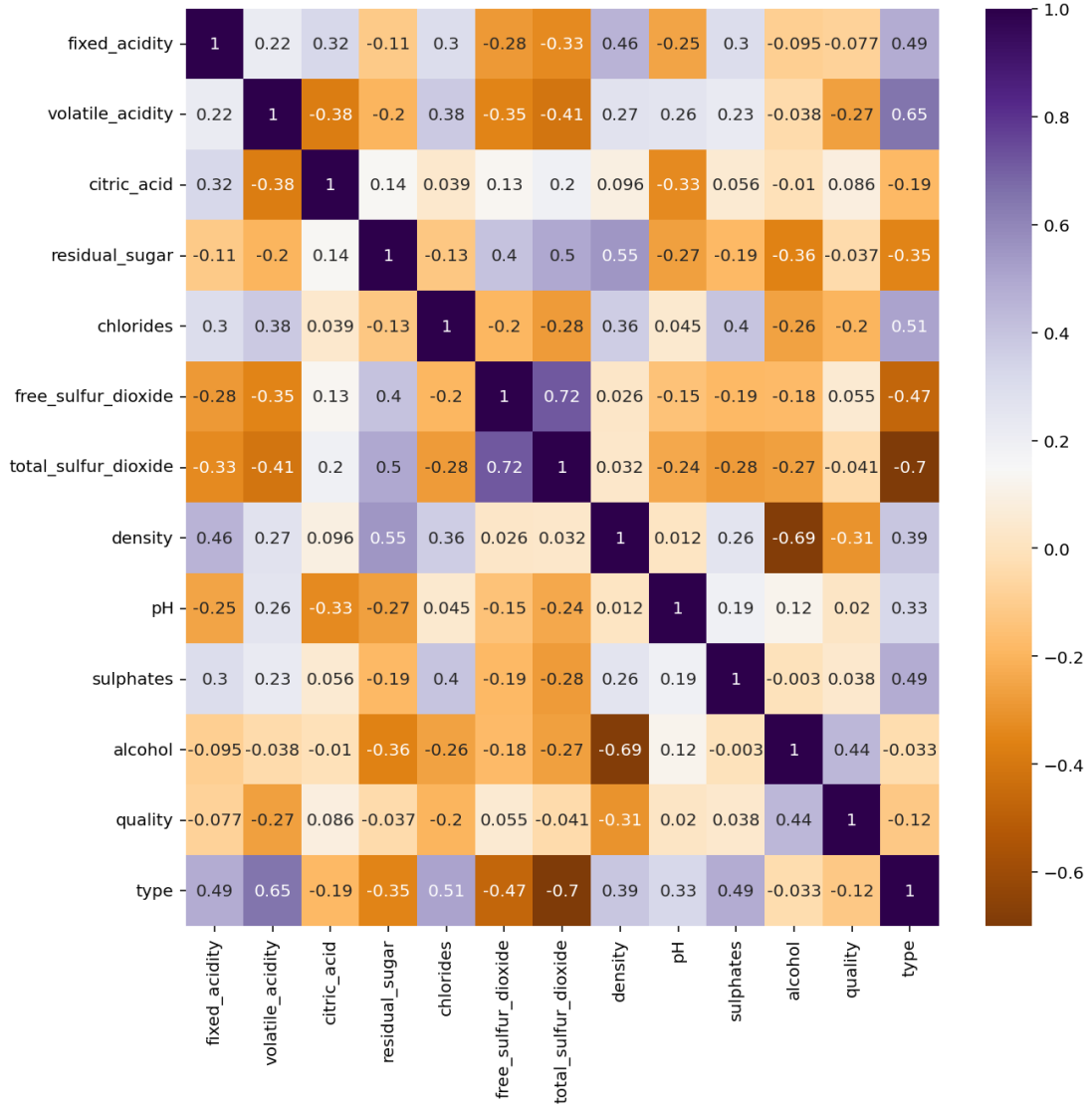
(26, 186, 1788, 2416, 924, 152, 5)

와인 품질 데이터 머신러닝



< 특성 중요도 >
- Quality에 영향을 미치는 정도

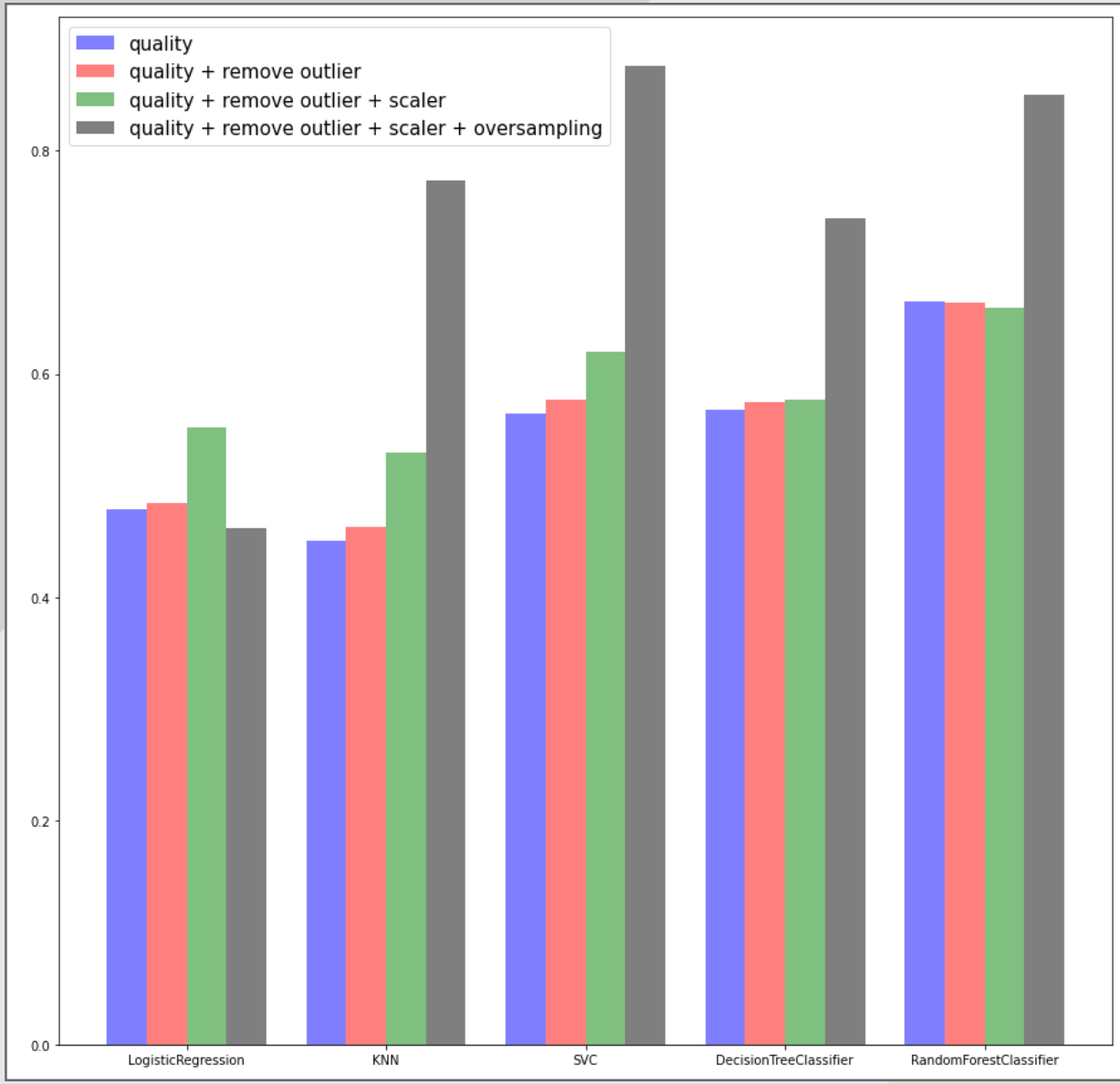
와인 품질 데이터 머신러닝



<데이터 상관관계 히트맵>

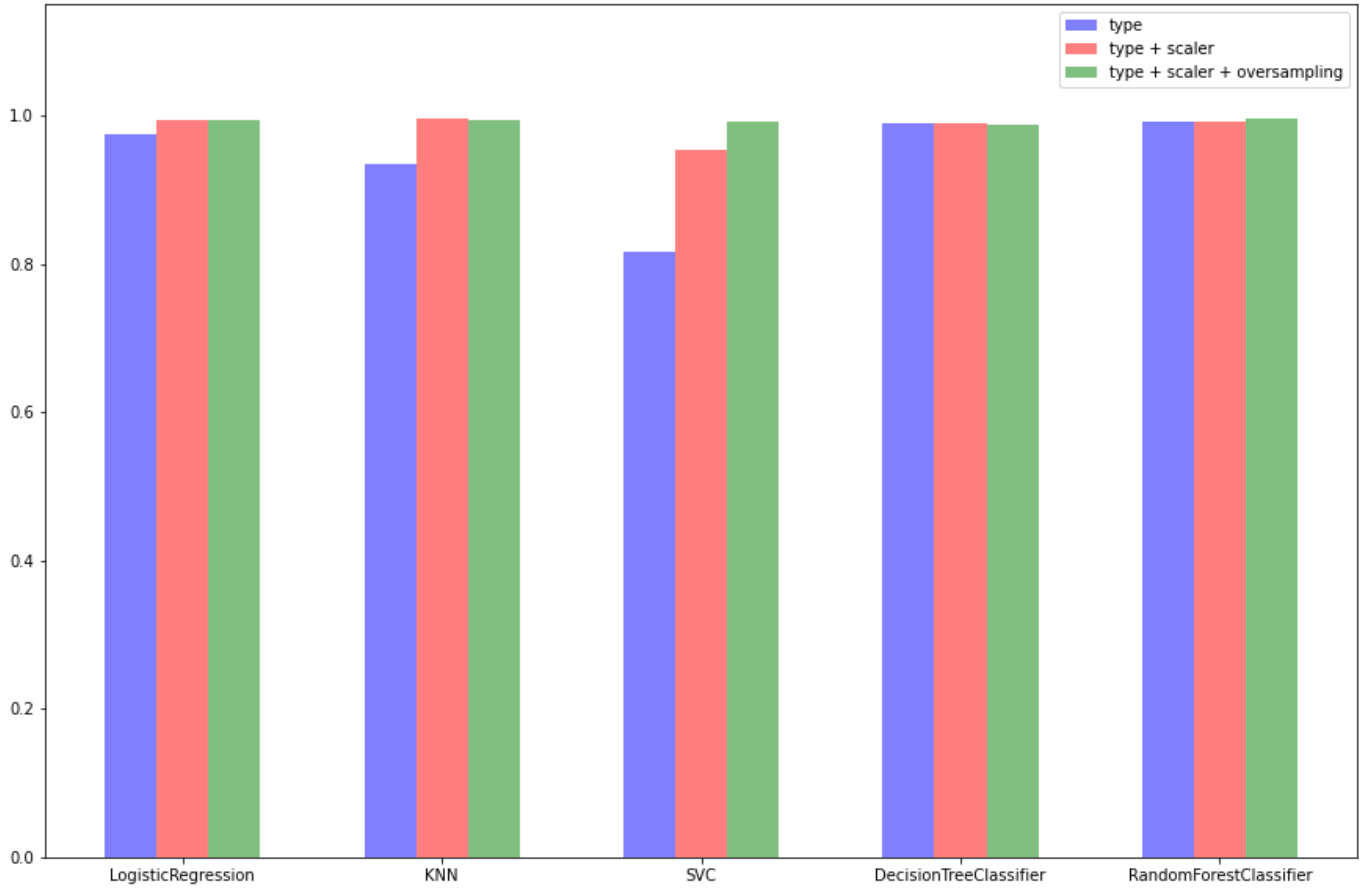
- type은 total_sulfur_dioxide(총 이산화황)에 높은 상관관계를 가짐
- quality는 alcohol, density(밀도)에 높은 상관관계를 가짐

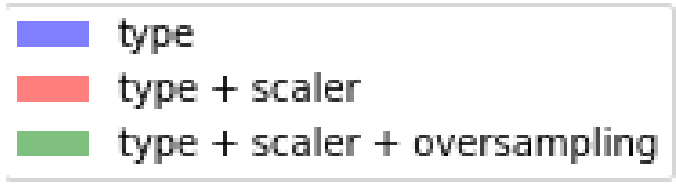
와인 품질 데이터 머신러닝



quality
quality + remove outlier
quality + remove outlier + scaler
quality + remove outlier + scaler + oversampling

와인 품질 데이터 머신러닝





와인 품질 데이터 머신러닝

4. 프로젝트 후기

- 데이터 수집에서 난항을 겪지 않아서 좋았음
- Scale 조정 -> 점수 하락 발생
- 데이터가 불균형하기 때문에 SMOTE 적용 -> 점수 하락 발생
- 다른 코드들 분석 결과 대부분 다항 분류가 아니라 이진 분류로
  데이터 타깃을 변형하여 예측
- Target을 Type으로 설정하여 red와 white 와인을 분류하는 문제 정의
  : Target을 Quality로 할 때와는 다르게 점수가 높게 측정됨

Q&A

들어주셔서 감사합니다