Wine Quality Analysis

2021008931 송주헌
2022000719 민지홍
2020037329 김충훈
2019003263 박민규
2021090546 정서현

CONTENTS

문제 설명

포르투갈 북부 지역의 레드 와인에 대한 화학적 특성과 품질에 대한
데이터를 사용하여 와인의 품질을 예측하는 모델을 구축하고자 한다.
데이터셋에는 다양한 화학적 특성(산도, 당도, 알코올 등)과 해당 와인의 품질 점수(0-10)가 포함되어 있다.
이 프로젝트의 목표는 이러한 화학적 특성이 와인의 품질에 어떤 영향을 미치는지 분석하고,
이를 바탕으로 와인의 품질을 예측하는 머신러닝 모델을 개발하는 것이다.

프로젝트 범위

◆데이터 탐색 및 전처리

데이터셋을 탐색하고 전처리 과정을 통해 분석과 모델링에 적합한 형태로 변환

◆특성 공학 및 선택

와인의 품질에 영향을 미치는 주요 화학적 특성을 선택

◆모델 구축 및 평가

여러 머신러닝 알고리즘을 사용하여 모델을 구축하고, 그 성능을 평가

◆결과 분석 및 시각화

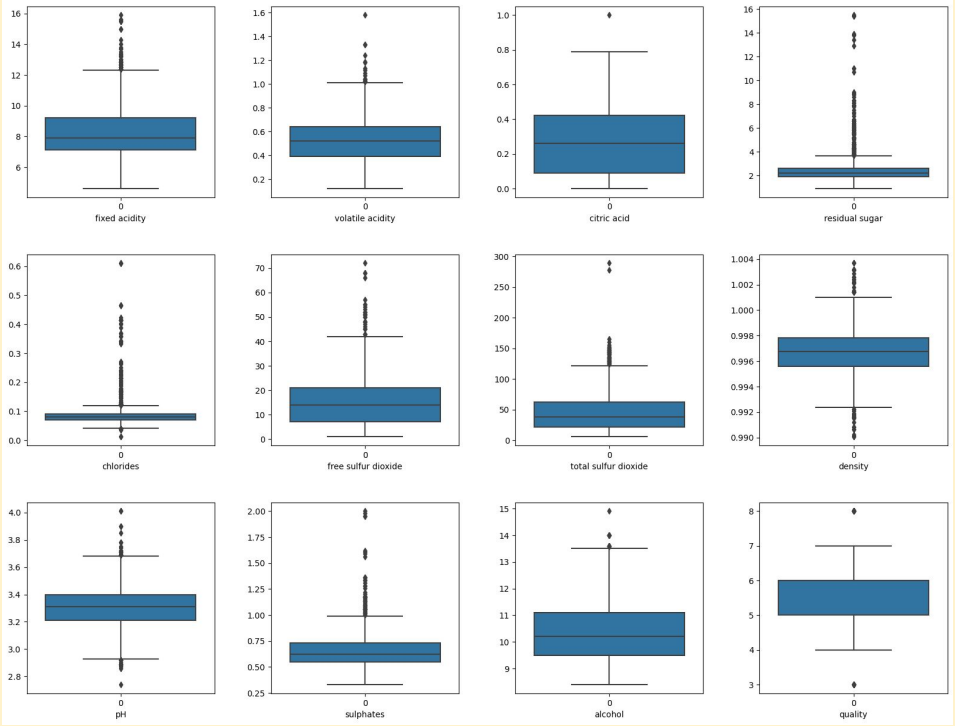
모델의 예측 결과를 분석하고, 주요 인사이트를 시각화

프로젝트 목표

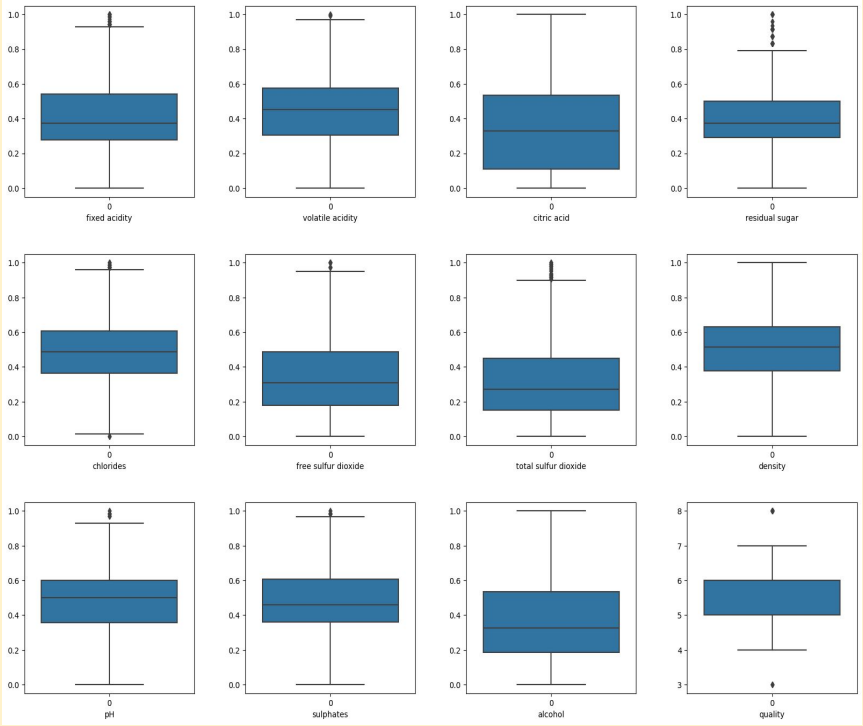
- 레드 와인의 화학적 특성을 분석하여 품질에 영향을 미치는 주요 요인을 식별한다
- 와인의 품질을 예측할 수 있는 머신러닝 모델을 개발한다
- 개발한 모델의 성능을 평가하고, 실제 응용 가능성을 검토한다

탐색적 데이터 분석 (EDA)

데이터 전처리(이상치 제거, 정규화) [데이터 시각화]



데이터 전처리 전



데이터 전처리 후

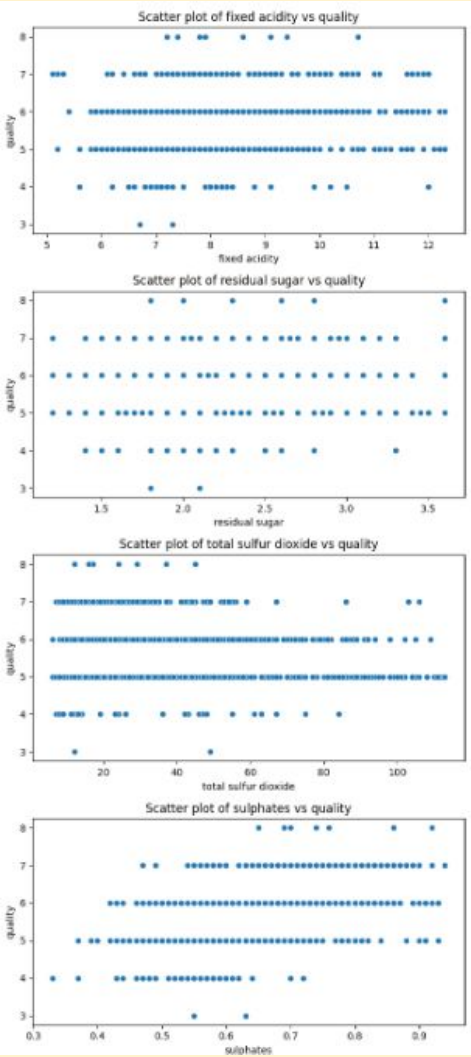
탐색적 데이터 분석 (EDA)

Fixed Acidity vs Quality
고정 산도와 품질 간의 관계에서, 고정 산도가 품질에 큰 영향을 미치지 않는 것으로 보인다.
품질 점수가 3에서 8까지 다양하게 분포되어 있으며, 고정 산도가 높아지거나 낮아짐에 따라 뚜렷한 패턴이 보이지 않는다.

Residual Sugar vs Quality
잔류 당과 품질 간의 관계는 뚜렷한 패턴을 보이지 않는다. 품질 점수가 잔류 당 함량과 상관없이 다양하게 분포되어 있다.

Total Sulfur Dioxide vs Quality
총 이산화황과 품질 간의 관계는 약한 음의 상관관계를 보인다.
총 이산화황 함량이 낮을수록 품질 점수가 높아지는 경향이 있다.

Sulphates vs Quality
황산염과 품질 간의 관계는 양의 상관관계를 보인다.
황산염 함량이 높을수록 품질 점수가 높아지는 경향이 있다.



[데이터 시각화]

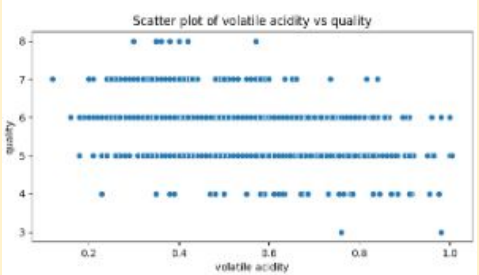
탐색적 데이터 분석 (EDA)

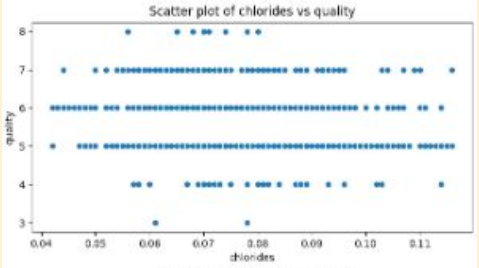
Volatile Acidity vs Quality
휘발성 산도와 품질 간의 관계는 음의 상관관계를 보인다.
휘발성 산도가 낮을수록 품질 점수가 높아지는 경향이 있다.
즉, 휘발성 산도가 높으면 와인의 품질이 낮아질 가능성이 크다

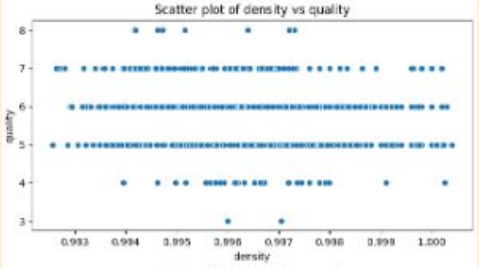
Chlorides vs Quality
염화물과 품질 간의 관계는 음의 상관관계를 보인다.
염화물 함량이 낮을수록 품질 점수가 높아지는 경향이 있다.

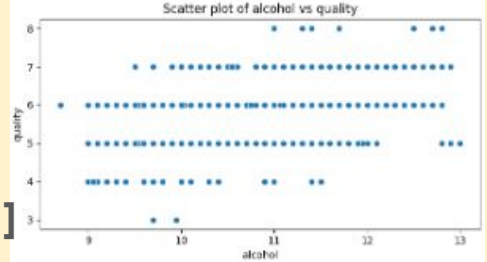
Density vs Quality
밀도와 품질 간의 관계는 음의 상관관계를 보인다.
밀도가 낮을수록 품질 점수가 높아지는 경향이 있다.

Alcohol vs Quality
알코올과 품질 간의 관계는 강한 양의 상관관계를 보인다.
알코올 함량이 높을수록 품질 점수가 높아지는 경향이 있다.









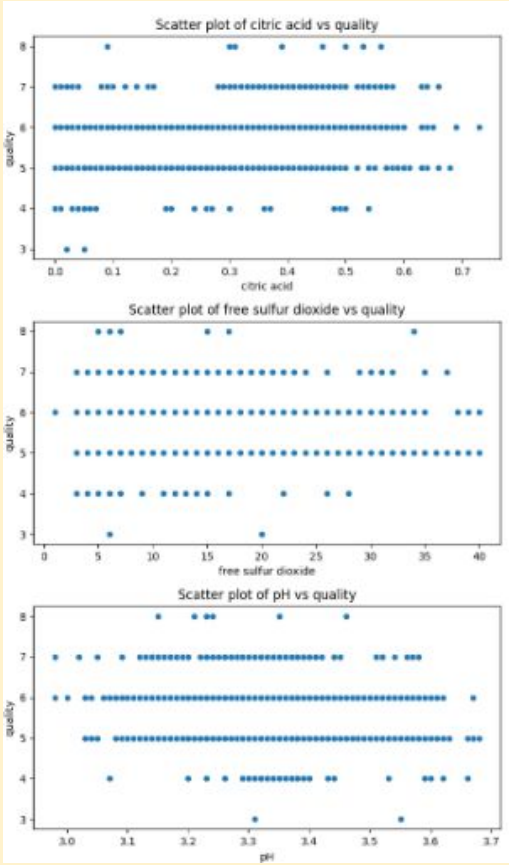
[데이터 시각화]

탐색적 데이터 분석 (EDA)

Citric Acid vs Quality
시트르산과 품질 간의 관계에서 약간의 양의 상관관계가 보인다.
시트르산 함량이 높을수록 품질 점수가 약간 증가하는 경향이 있다.

Free Sulfur Dioxide vs Quality
자유 이산화황과 품질 간의 관계는 뚜렷한 패턴이 보이지 않는다.
품질 점수가 자유 이산화황 함량과 상관없이 다양하게 분포되어 있다.

pH vs Quality
pH와 품질 간의 관계는 뚜렷한 패턴을 보이지 않는다.
품질 점수가 pH 값과 상관없이 다양하게 분포되어 있다.



탐색적 데이터 분석 (EDA)

[변수 간 상관관계 분석]

residual sugar와
free sulfur dioxide를 제외한
나머지 설명변수들의
Quality에 대한 P 값이
0.05보다 작으므로
통계적으로 유의미한
관계를 가진다.

	피어슨 상관계수	P-값
alcohol	0.48	1.241040e-69
sulphates	0.25	6.715461e-53
volatile acidity	-0.39	4.917301e-36
citric acid	0.23	6.128857e-14
density	-0.17	3.590696e-13
chlorides	-0.13	6.580961e-11
total sulfur dioxide	-0.18	7.748687e-09
fixed acidity	0.12	1.941172e-04
pH	-0.06	3.886856e-02
residual sugar	0.01	1.794222e-01
free sulfur dioxide	-0.05	9.126474e-01

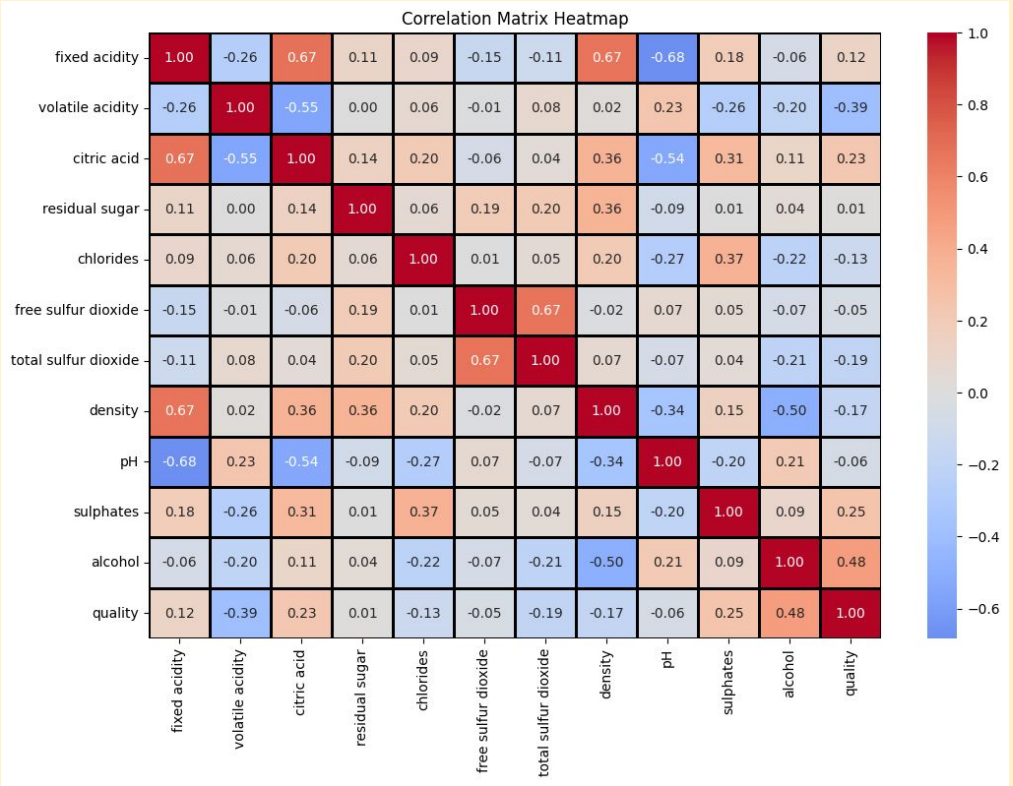
탐색적 데이터 분석 (EDA)

[변수간 상관관계 분석]

피어슨 상관계수를 사용하여

변수간의 상관관계를 나타낸 표

Quality가 **alcohol**과 **sulphates**와의 상관관계가 높은것으로 보임



모델 개발

[알고리즘 소개 (선형 회귀, 랜덤 포레스트)]

```
6 #선형 회귀 예측
7 model = LinearRegression()
8 model.fit(x_train, y_train)
9 y_pred = model.predict(x_test)
```

종속 변수와 하나 또는

여러 독립 변수 간의 선형 관계를

모델링하는 회귀 기법

```
1 #랜덤 포레스트
2 model_r = RandomForestRegressor(n_estimators=100, random_state=42)
3 model_r.fit(x_train, y_train)
```

다수의 결정 트리(decision trees)를 사용하여

예측을 수행하고, 각 트리의 결과를
평균화하여

최종 예측을 하는 앙상블 기법

모델 개발

[알고리즘 소개 (Bagging)]

```
30 # 배깅을 사용한 결정 트리 모델 훈련
31 bagging_regressor = BaggingRegressor(
32     base_estimator=DecisionTreeRegressor(random_state=42),
33     n_estimators=100,
34     random_state=42
35 )
36 bagging_regressor.fit(X_train_scaled, y_train)
37
38 # 예측
39 y_pred = bagging_regressor.predict(X_test_scaled)
```

동일한 학습 알고리즘을 여러 번 적용하여 각각의 결과를 평균 또는 투표로 결합해 예측 성능을 향상시키는 부트스트랩 샘플링 기반의 앙상블 기법

모델 개발

[알고리즘 소개 (XGBoost)]

```
24 # 모델 훈련 및 예측
25 models = {
26     'XGBoost': xgb.XGBRegressor(objective ='reg:squarederror', random_state=42)
27 }
28
29 results = {}
30
31 for model_name, model in models.items():
32     model.fit(X_train_scaled, y_train)
33     y_pred = model.predict(X_test_scaled)
34     mse = mean_squared_error(y_test, y_pred)
35     r2 = r2_score(y_test, y_pred)
36     results[model_name] = {'MSE': mse, 'R2': r2}
```

경사 부스팅 알고리즘의 확장 버전으로, 성능과 효율성을 높이기 위해 정교한 트리 기반 모델을 사용하는 강력한 앙상블 학습 기법

모델 평가

[교차 검증, 일반화 성능 평가] (랜덤포레스트, 배깅 n_estimators = 100)

	선형 회귀	랜덤 포레스트	XGBoost	배깅
전처리 전	MSE : 0.39 R^2 : 0.4	MSE : 0.3 R^2 : 0.54	MSE : 0.35 R^2 : 0.46	MSE : 0.3 R^2 : 0.54
이상치 제거	MSE : 0.3 R^2 : 0.36	MSE : 0.25 R^2 : 0.46	MSE : 0.3 R^2 : 0.36	MSE : 0.25 R^2 : 0.46
제거 후 정규화	MSE : 0.3 R^2 : 0.36	MSE : 0.25 R^2 : 0.46	MSE : 0.3 R^2 : 0.37	MSE : 0.26 R^2 : 0.46

결과 분석 및 시각화

	ind	MSE	R^2
0	1	0.398833	0.163758
1	2	0.351010	0.331540
2	3	0.387232	0.390914
3	4	0.383841	0.312132
4	5	0.411959	0.278162

배깅

	ind	MSE	R^2
0	1	0.396977	0.167651
1	2	0.350020	0.333426
2	3	0.387902	0.389859
3	4	0.385125	0.309832
4	5	0.409664	0.282183

랜덤 포레스트

	ind	MSE	R^2
0	1	0.380402	0.202402
1	2	0.319036	0.392430
2	3	0.363122	0.428836
3	4	0.365256	0.345437
4	5	0.395805	0.306466

선형회귀

각 예측 모델별 MSE와 R^2 결정계수 값입니다.

선형회귀의 MSE와 결정계수가 다른 모델보다 더 좋은 예측 결과를 나타내고 있습니다.

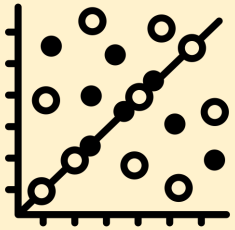
결과 분석 및 시각화

선형회귀로 **K-fold(K=5)**를 진행하며 생성된
모델 5개의 **Quality**에 대한
설명변수의 평균상관계수입니다.

Quality를 예측하는데 높은 영향을 미치는 요인이
alcohol과 **sulphates**
그 다음은 **volatile acidity, pH** 순으로 영향을
미친다는 것을 알 수 있습니다.

	변수	평균 계수
0	fixed acidity	0.081650
1	volatile acidity	-0.767997
2	citric acid	-0.226267
3	residual sugar	0.052973
4	chlorides	-0.093544
5	free sulfur dioxide	0.107950
6	total sulfur dioxide	-0.255846
7	density	-0.071770
8	pH	-0.444555
9	sulphates	1.135623
10	alcohol	1.233556

결론



선택한 모델: 정규화 데이터 기반 선형회귀

- **정규화 데이터 기반 선형 회귀 모델**이 와인의 화학적 특성을 기반으로 품질을 예측하는 데 있어 다른 모델들보다 우수한 성능을 보였다.

- 특히 **예측값과 실제값 간의 차이가 가장 적었기 때문**에, 본 프로젝트의 목적에 가장 부합한다고 판단함.

- 따라서, 최종 모델로 **정규화 데이터 기반 선형 회귀 모델을 채택**하여 와인 품질 예측을 수행.


WINE
Thank
you