

2025년도 공공기관 용역과제
AI개발 수행내역서

과제명	AI기반 와인 품질 예측모델 개발 및 시각화
담당자	임예지

2025년 10 월 17 일

AI개발 수행내용

1. 사업과제 : 와인의 화학적 특성에 따른 품질 예측모델 개발 및 시각화

2. 개요 및 현황

2.1 추진배경 및 목적

- 기존의 와인 품질 평가는 전문가의 주관적 판단에 의존하는 한계가 있었으나, 머신 러닝을 활용하여 화학적 특성 데이터를 기반으로 품질 예측을 하고자 함
- 와인의 산도, 알코올 도수, 황산염, 휘발산 등의 화학적 요소 간의 복잡한 상관관계를 분석하여 품질을 예측하는 AI 모델을 구축하고자 함
- 구축된 모델은 데이터 기반의 품질 관리 시스템 개발 및 생산 효율성 향상에 기여할 수 있으면, 향후 식음료 및 발효 산업 전반으로의 확장 가능성을 모색하고자 함

2.2 과제 범위

과제구분		내용
AI	AI기반 화학적 특성에 따른 와인 품질 예측모델 구현	원시 데이터 수집 및 데이터셋 구축
		데이터 전처리, 표준화, 상관관계 분석 (EDA도구 활용)
		예측모델 선정 및 학습
		R^2 , MAE 등 평가지표를 활용한 모델 성능 평가
		streamlit 프로토타입 구축
		테스트
시각화	데이터 분석 및 예측결과 시각화 구현	사용자 입력 연동 기반 예측 시스템 구현
		데이터 시각화 및 대시보드 구현
		예측모델 시각화
		테스트
		통합테스트 및 시운전

2.3 과제 추진 방법

1) 구축 대상 선정 기준

○ 데이터 접근성 및 활용성

- 데이터 수집 및 관리의 용이성
- 정부 및 공공기관에서 이미 구축된 데이터베이스 활용 여부
- 종속변수(품질 점수)에 영향을 미치는 다양한 독립변수 확보를 통해 모델 학습에 적합한 구조를 가짐

○ 예측모델 개발 효율성

- 비교적 단순한 변수 구조를 갖고 있어 모델 학습 및 평가 과정이 효율적으로 수행 가능

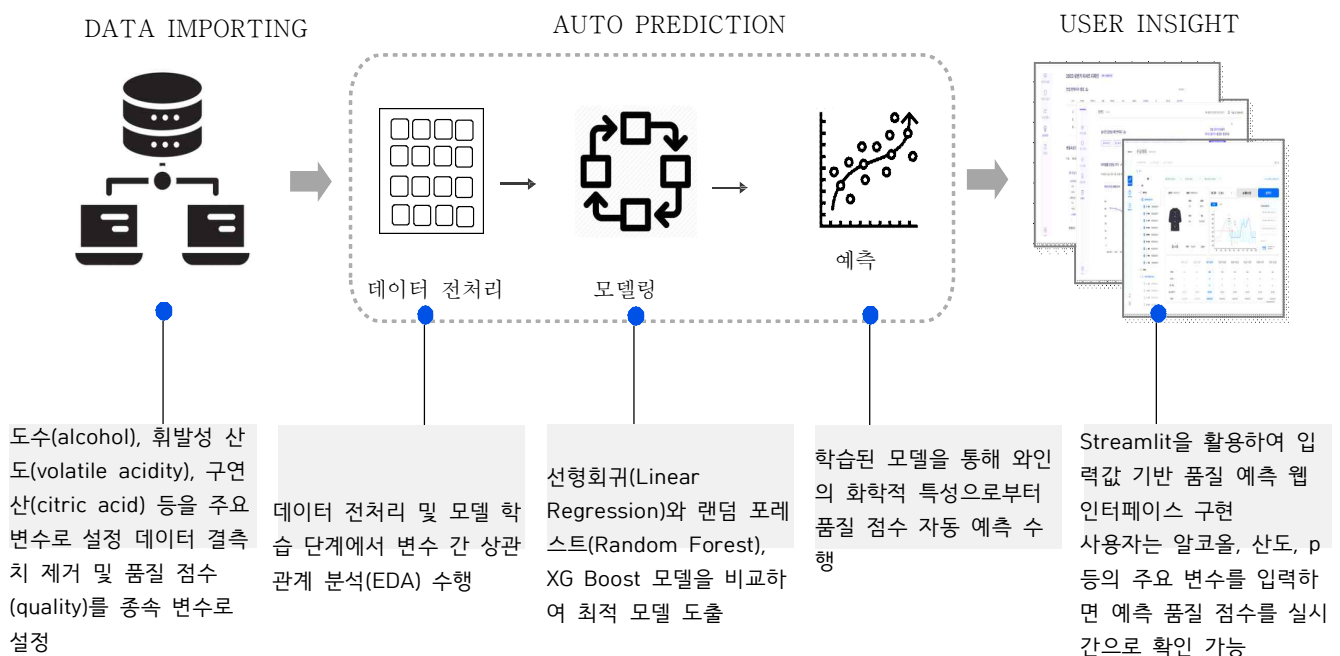
○ 품질 관리 기여도 및 실용성

- 데이터 기반 의사결정을 통해 생산 과정의 효율성 향상 및 품질 편차 최소화 기대
- 와인 산업 내 품질 예측 자동화로 인한 시간·비용 절감 및 소비자 만족도 향상 기여

2) AI 예측 분석모델 적용 대상

구분	수집 데이터	예측모델인자(독립변수)	AI예측 분석 대상
품질 관리 (와인)	<ul style="list-style-type: none"> - 와인의 화학적 성분 및 품질 점수 데이터 - 각 시료별 11개 화학적 특성 수집 	<ul style="list-style-type: none"> - 화학적 변수: 산도(fixed acidity), 휘발산(volatile acidity), 구연산(citric acid), 잔당(residual sugar), 염화물(chlorides), 황산염(sulphates), 알코올(alcohol) 등 - 환경 변수: pH, 밀도(density), 이산화황(SO₂) 농도 등 	<ul style="list-style-type: none"> - 화학적 변수 간 상관관계 분석 - 와인의 품질(quality) 예측 - 각 변수별 품질에 미치는 영향도 분석

3) AI 분석모델 구축 프로세스



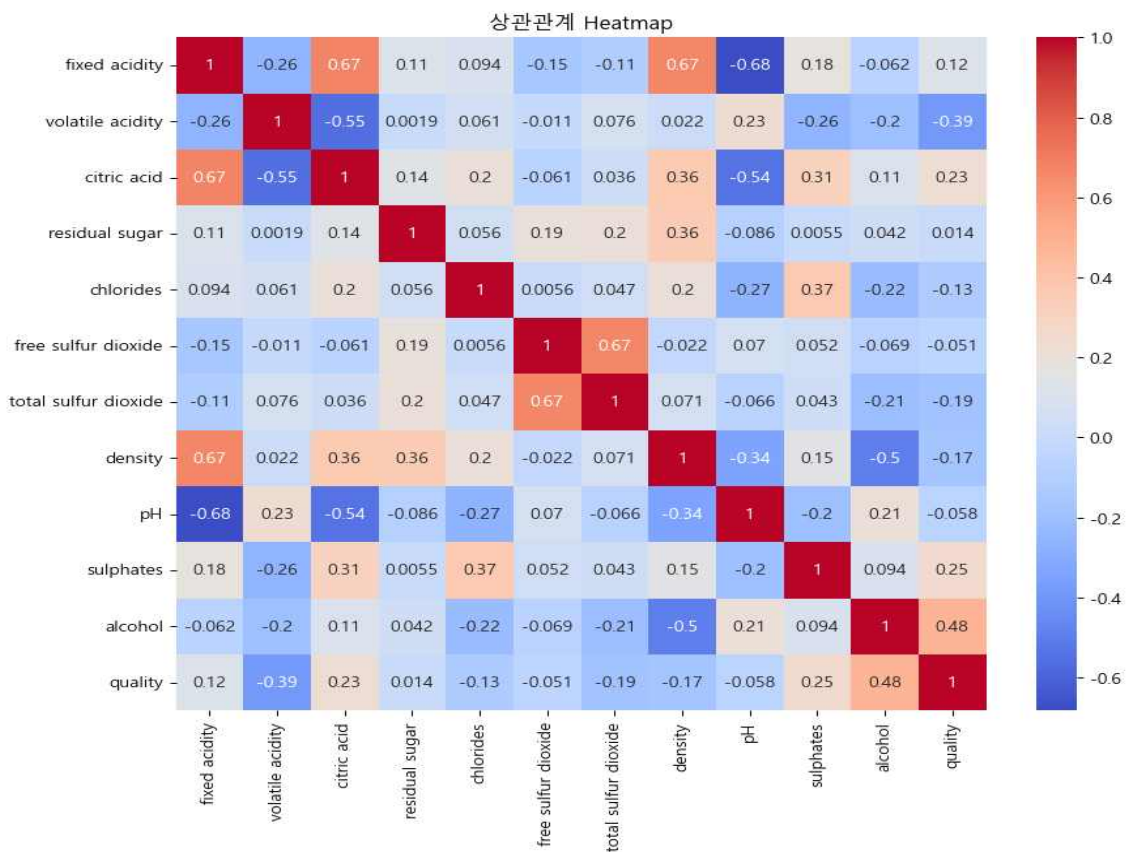
연구개발 주요 결과물

1. 데이터 수집

- 포르투갈산 "비뉴 베르데(Vinho Verde)"의 레드 와인 품질 데이터

2. 데이터 분석

2.1 와인 품질데이터 상관관계(Heatmap)



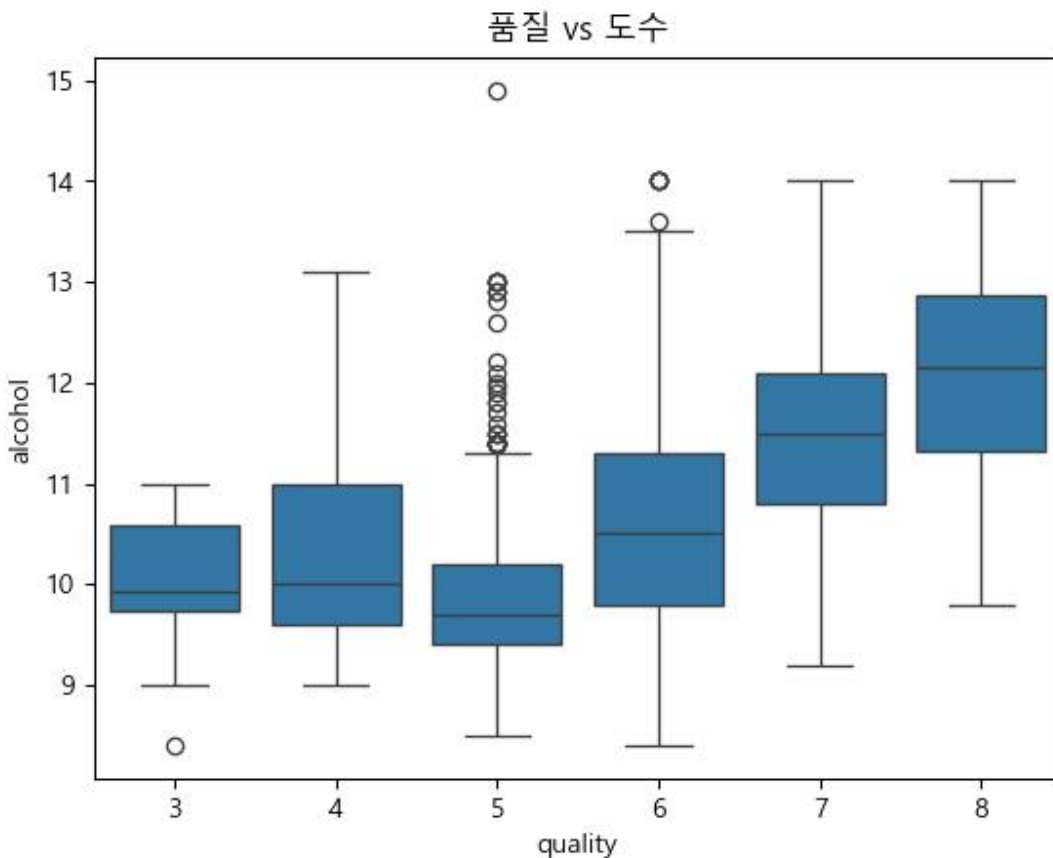
○ 상관계 히트맵을 통해 다음과 같은 관계를 확인함:

- 알코올 도수(alcohol) 와 품질(quality) 은 양의 상관관계(≈ 0.48) 로, 알코올 도수가 높을수록 품질 점수가 높아지는 경향을 보임.
- 휘발산(volatile acidity) 은 품질과 음의 상관관계(≈ -0.39) 를 보여, 산도가 높을수록 와인 품질이 낮아지는 경향이 나타남.
- 황산염(sulphates), 구연산(citric acid) 또한 품질과 비교적 높은 양의 상관관계를 보여 와인의 풍미와 보존성에 영향을 미치는 주요 변수로 확인됨.

○ 이러한 분석 결과를 기반으로, 품질 예측에 영향을 주는 주요 인자 알코올 도수, 휘발성 산도, 구연산, 황산염, 밀도, 총 이산화황, 고정 산도 7개의 변수 중심으로 모델링을 수행함.

2.2 탐색적 데이터 분석

○ 주요 변수(도수)의 분포 및 품질 영향 분석 (박스플롯)



○ 위 박스플롯은 품질(quality) 등급별 알코올(alcohol) 함량의 분포를 나타낸다.

전반적으로 품질 점수가 높을수록 알코올 도수가 증가하는 경향을 보이며, 이는 히트맵 분석 결과(알코올과 품질의 양의 상관관계 +0.48)와 일치한다.

품질 7~8등급 구간에서 알코올 농도가 상대적으로 높고, 품질 3~4 구간에서는 낮은 알코올 함량이 주로 관찰된다.

따라서 알코올 함량은 와인 품질을 결정하는 주요 요인으로 해석된다.

○ 데이터 전처리

First rows		Last rows										
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
5	7.4	0.66	0.00	1.8	0.075	13.0	40.0	0.9978	3.51	0.56	9.4	5
6	7.9	0.60	0.06	1.6	0.069	15.0	59.0	0.9964	3.30	0.46	9.4	5
7	7.3	0.65	0.00	1.2	0.065	15.0	21.0	0.9946	3.39	0.47	10.0	7
8	7.8	0.58	0.02	2.0	0.073	9.0	18.0	0.9968	3.36	0.57	9.5	7
9	7.5	0.50	0.36	6.1	0.071	17.0	102.0	0.9978	3.35	0.80	10.5	5

3. 데이터 학습 및 모델정의

3.1 데이터 분할

○ 학습용(80%)과 테스트용(20%)으로 분할하여 모델의 일반화 성능을 검증

```
# 모델 학습
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

3.2 모델학습 및 학습 시각화

○ 모델 선정

와인 품질 예측을 위해 선형회귀(Linear Regression)와

랜덤 포레스트(Random Forest) 그리고 XG 부스트, SVR, LGBM 다섯 가지 모델을 사용

두 모델을 학습시킨 후 결정계수(R^2), RMSE, MAE 등을 비교하여 최종 모델 결정

XGBoost 기본 - RMSE: 0.6222, R^2 : 0.4077

[Linear Regression] 성능 평가
RMSE: 0.63, MSE: 0.40, R^2 : 0.39

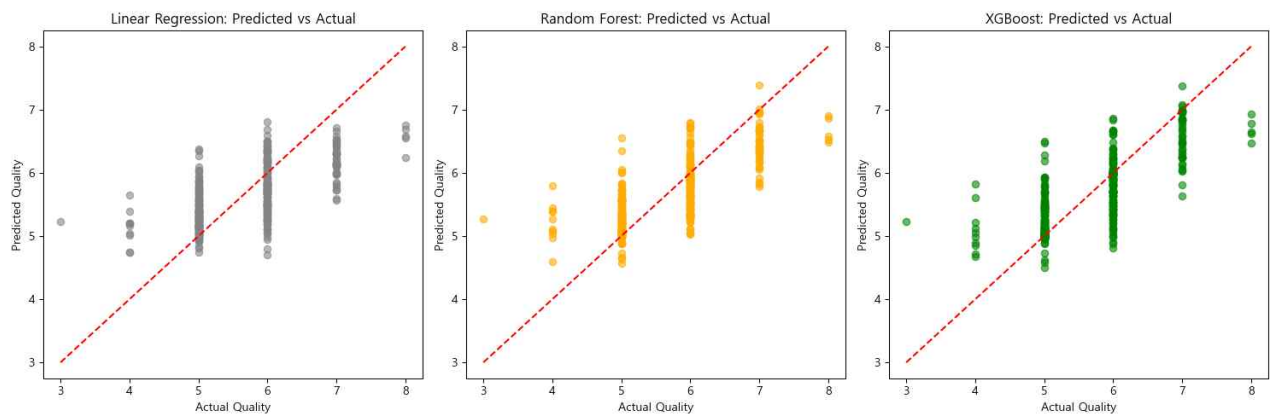
[Random Forest] 성능 평가
RMSE: 0.57, MSE: 0.33, R^2 : 0.50

[XG Boost] 성능 평가
RMSE: 0.57, MSE: 0.32, R^2 : 0.51

SVR 성능: RMSE=0.725, R^2 =0.196
튜닝된 SVR 성능: RMSE=0.599, R^2 =0.452

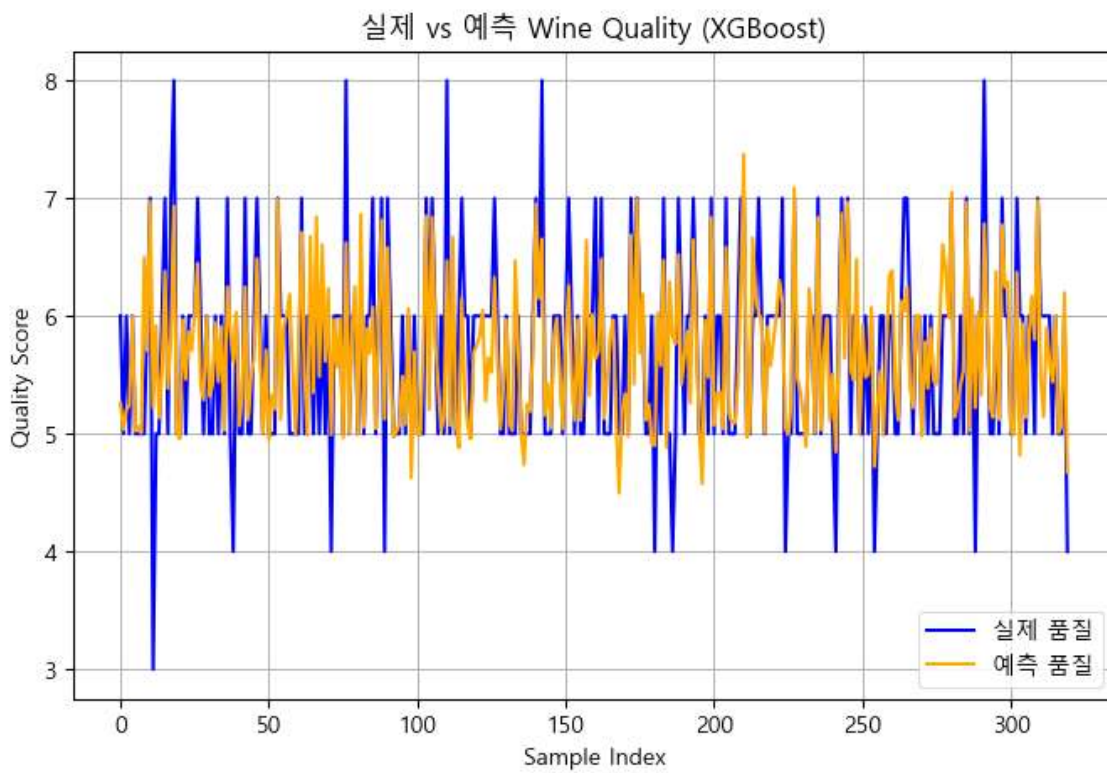
LGBM 성능: RMSE=0.588, R^2 =0.471
튜닝 후 LGBM 성능: RMSE=0.582, R^2 =0.482

○ 선형 회귀VS 랜덤 포레스트 VS XG Boost



3.3 모델 예측

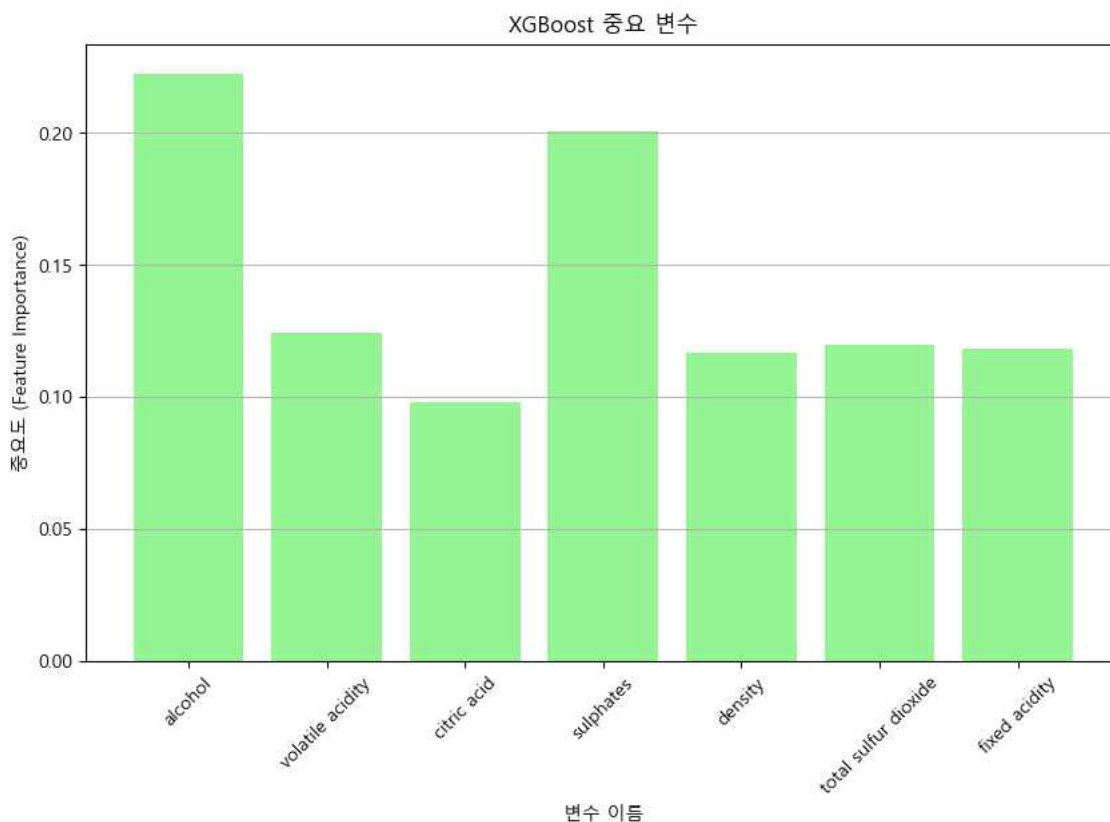
○ 예측값 vs 실제값 비교



- 파란색 선은 실제 품질 데이터, 주황색 선은 XG Boost로 예측한 품질 데이터로 실제 점수의 상승 추세를 거의 따라가며 일부 구간에서 약간의 진폭의 오차만 보임
- 두 그래프의 추세가 유사하게 움직이는 것으로 보아, 모델이 실제 데이터의 패턴을 일정 수준 학습
- 예측값이 과도하게 벗어나는 구간이 없으며, 전반적으로 모델이 품질의 흐름을 잘 반영

3.4 모델 해석(Feature Importance 분석)

- XG Boost 중요 변수 분석



- alcohol(알코올 도수)이 가장 높은 중요도를 보여 품질에 가장 큰 영향을 미치는 변수
- 도수가 높을수록 풍미와 저장성에 영향을 미쳐 품질 점수가 상승하는 경향을 반영한 것으로 해석
- 황산염(sulphates), 휘발성 산도(volatile acidity) 역시 품질 예측에 주요한 영향을 미쳤으며, 이는 산도와 향 안정성이 와인 품질에 직결됨을 시사

4. Streamlit 상호작용 시각화

4.1 변수 입력 예측 페이지

와인 품질 예측 시스템

화학적 특성 값을 입력하고 "예측하기"를 눌러 품질 점수를 확인하세요. 🧐

alcohol(알코올 도수)

9.40

volatile acidity(휘발성 산도)

0.70

citric acid(구연산)

0.00

sulphates(황산염)

0.56

density(밀도)

1.00

total sulfur dioxide(총 이산화황)

34

fixed acidity (고정 산도)

7.40

예측하기 🚀

4. 결론 및 향후 계획

4.1 프로젝트의 결론

- 예측 결과 그래프에서도 실제 품질값과 예측값의 변화 추세가 유사하게 나타나, XG Boost 모델이 비선형적 특성을 효과적으로 학습했음을 확인
- 변수 중요도 분석에서는 alcohol(알코올 도수)가 가장 높은 영향력
- 그 외에 sulphates(황산염), volatile acidity(휘발성 산)도 와인 품질 예측에 주요 요인으로 작용
이를 통해 알코올 도수와 산도 균형이 와인 품질의 주요 판단 기준임을 파악

4.2 한계점 및 개선방향

- 데이터의 한정성
 - 데이터가 특정 지역의 한 종류의 와인에 국한되어 있어, 다양한 생산지·품종의 와인에 대한 일반화가 어렵다.
- 모델의 설명력의 한계
 - $R^2 = 0.55$ 으로, 절반 정도의 품질 변동만 설명 가능
- 향후 개선 방향
 - 보다 다양한 와인 종류와 생산지를 포함한 데이터 확장
 - 고성능 앙상블 모델 적용
 - Feature Engineering(특성 조합 및 비선형 변수 생성) 을 통한 설명력 향상