와인 품질 예측 모델
https://www.kaggle.com/datasets/ruthgn/wine-quality-data-set-red-white-wine/data

변수명		내용
type of wine	와인 종류	와인 종류 **(분류 : '레드', '화이트')**
fixed acidity	고정산도	**와인을 발효하는 데 사용된 포도에서 자연적으로 발생하여 와인으로 옮겨지는 산. (g / dm^3)** 주로 와인을 발효하는 데 사용된 포도에서 유래하는 타르타르산, 사과산, 구연산 또는 석신산으로 구성됨. 또한 쉽게 증발되지 않음.
volatile acidity	휘발성 산도	**낮은 온도에서 증발하는 산. (g / dm^3)** 주로 아세트산으로, 매우 높은 수준에서 불쾌한 식초와 같은 맛을 낼 수 있음
citric acid	구연산	**구연산은 와인의 산도를 높이는 산 보충제로 사용됨. (g / dm^3)** 일반적으로 소량으로 발견되며 와인에 '신선함'과 풍미를 더할 수 있습니다.
residual sugar	잔류당	**발효가 멈춘 후 남은 당의 양. (g / dm^3)** 1g/L 미만인 와인을 찾는 것은 드물. 잔류당 수치가 45g/L 이상인 와인은 달콤한 것으로 간주됨 반면에 달콤한 맛이 나지 않는 와인은 드라이한 것으로 간주됨
chlorides	염화물	**와인에 존재하는 염화물 염(염화나트륨)의 양. (g / dm^3)**
free sulfur dioxide	자유 이산화황	자유 형태의 SO2는 분자 SO2(용해 가스)와 중아황산염 이온 사이에 평형 상태로 존재하며, 미생물의 성장과 와인의 산화를 방지. 다른 모든 것이 일정하다면, 자유 이산화황 함량이 높을수록 보존 효과가 더 강함. **(mg / dm^3)**
total sulfur dioxide	총 이산화황	**SO2의 자유 형태와 결합 형태의 양 (mg/dm^3)** 낮은 농도에서 SO2는 대부분 와인에서 감지되지 않지만, 자유 SO2 농도가 50ppm을 초과하면 SO2가 와인의 향과 맛에서 분명해짐.
density	밀도	**와인 주스의 밀도**는 **알코올 함량과 설탕 함량에 따라 달라짐. (g / cm^3)** 일반적으로 물의 밀도와 비슷하지만 더 높음. (와인은 '더 진하다').
pH	pH	**와인의 산도를 측정하는 척도** 대부분 와인은 **pH 척도에서 3-4 사이임.** pH가 낮을수록 와인의 산성도가 더 높고, pH가 높을수록 와인의 산성도가 낮음. (pH 척도는 기술적으로 와인에 떠다니는 자유 수소 이온의 농도를 측정하는 대수 척도) (**pH 척도의 각 지점**은 **10의 인수**임. 즉, **pH가 3인 와인은 pH가 4인 와인보다 10배 더 산성**입니다.)
sulphates	황산염	**항균 및 항산화제로 작용함.(g/dm3)** 와인 첨가물로서 황산칼륨의 양은 **이산화황 가스(SO2) 수치에 영향**을 줄 수 있음.
alcohol	알코올	주어진 와인 부피에 얼마나 많은 알코올이 포함되어 있는가(ABV). 와인은 일반적으로 5~15%의 알코올을 포함. **(부피 기준 %)**
quality	품질	와인 전문가가 매긴 사이의 점수 (0(매우 나쁨)에서 10(매우 우수))

물리화학적 실험에 근거한 연속 변수
고정산도, 휘발성 산도, 구연산, 잔류당, 염화물, 자유 이산화황, 총 이산화황, 밀도, pH, 황산염, 알코올, 품질

출력변수 : 품질

1. 데이터와 데이터의 출처 → 데이터 선정배경 왜 이 데이터를 선정하게 되었는가**?**
kaggle에서 와인 품질 예측 데이터를 가져왔음
포르투갈의 미뉴 지역에서 생산되는 비뉴 베드의 레드와인과 화이트 와인에 대한 데이터
(https://www.kaggle.com/datasets/ruthgn/wine-quality-data-set-red-white-wine/data)

---

2. 데이터 각 변수의 의미 → 데이터를 보고 중요도 판단 해야함**.**

변수명		내용
type of wine	와인 종류	와인 종류 **(분류 : '레드', '화이트')**
fixed acidity	고정산도	**와인을 발효하는 데 사용된 포도에서 자연적으로 발생하여 와인으로 옮겨지는 산. (g / dm^3)** 주로 와인을 발효하는 데 사용된 포도에서 유래하는 타르타르산, 사과산, 구연산 또는 석신산으로 구성됨. 또한 쉽게 증발되지 않음.
volatile acidity	휘발성 산도	**낮은 온도에서 증발하는 산. (g / dm^3)** 주로 아세트산으로, 매우 높은 수준에서 불쾌한 식초와 같은 맛을 낼 수 있음
citric acid	구연산	**구연산은 와인의 산도를 높이는 산 보충제로 사용됨. (g / dm^3)** 일반적으로 소량으로 발견되며 와인에 '신선함'과 풍미를 더할 수 있습니다.
residual sugar	잔류당	**발효가 멈춘 후 남은 당의 양. (g / dm^3)** 1g/L 미만인 와인을 찾는 것은 드물. 잔류당 수치가 45g/L 이상인 와인은 달콤한 것으로 간주됨 반면에 달콤한 맛이 나지 않는 와인은 드라이한 것으로 간주됨
chlorides	염화물	**와인에 존재하는 염화물 염(염화나트륨)의 양. (g / dm^3)**
free sulfur dioxide	자유 이산화황	자유 형태의 SO2는 분자 SO2(용해 가스)와 중아황산염 이온 사이에 평형 상태로 존재하며, 미생물의 성장과 와인의 산화를 방지. 다른 모든 것이 일정하다면, 자유 이산화황 함량이 높을수록 보존 효과가 더 강함. **(mg / dm^3)**
total sulfur dioxide	총 이산화황	**SO2의 자유 형태와 결합 형태의 양 (mg/dm^3)** 낮은 농도에서 SO2는 대부분 와인에서 감지되지 않지만, 자유 SO2 농도가 50ppm을 초과하면 SO2가 와인의 향과 맛에서 분명해짐.
density	밀도	**와인 주스의 밀도는 알코올 함량과 설탕 함량에 따라 달라짐. (g / cm^3)** 일반적으로 물의 밀도와 비슷하지만 더 높음. (와인은 '더 진하다').
pH	pH	**와인의 산도를 측정하는 척도** 대부분 와인은 **pH 척도에서 3-4 사이임.** pH가 낮을수록 와인의 산성도가 더 높고, pH가 높을수록 와인의 산성도가 낮음. (pH 척도는 기술적으로 와인에 떠다니는 자유 수소 이온의 농도를 측정하는 대수 척도) (**pH 척도의 각 지점**은 **10의 인수**임. 즉, **pH가 3인 와인은 pH가 4인 와인보다 10배 더 산성**입니다.)
sulphates	황산염	**항균 및 항산화제로 작용함.(g/dm3)** 와인 첨가물로서 황산칼륨의 양은 **이산화황 가스(SO2) 수치에 영향**을 줄 수 있음.
alcohol	알코올	주어진 와인 부피에 얼마나 많은 알코올이 포함되어 있는가(ABV). 와인은 일반적으로 5~15%의 알코올을 포함. **(부피 기준 %)**
quality	품질	와인 전문가가 매긴 사이의 점수 (0(매우 나쁨)에서 10(매우 우수))

중요하지 않다고 생각되는 변수
➡️아직 데이터 중요도를 예측하는 과정 진행하지 않아 확실하진 않음
해당 성분에 대한 설명을 바탕으로 중요도가 낮다고 판단한 변수들
데이터 전처리 과정에서 중요도가 낮은 것이 확인 되면 제외할 계획

염화물
: 와인의 짠맛을 결정짓는 성분인데 와인에 아주 미량만 포함되어 있어 품질에 미치는 영향이 적다고 판단

자유이산화황**,** 총이산화황**,** 황산염
: 와인의 보존성을 높이는 역할을 하는 성분들인데 이 수치들로 와인의 품질이 결정되기 어렵다고 판단

---

3. 귀무가설**(**영가설**),** 대립가설 설정

귀무가설(영가설) ➡️ 원래있던 것 / ~차이가 없다, ~효과가 없다, ~관계가 없다
대립가설 ➡️ 내가 입증하려고 하는 것 / 확실한 근거에 의하여 입증하고자하는 가설

귀무가설**(**영가설**)** : 와인의 품질도는 설명변수와 관련이 없다.
대립가설 : 와인의 품질도는 설명변수와 관련이 있다.

4. 설명변수**,** 반응변수 설정 및 의미
반응 변수(= 종속 변수(y)) : 품질도(품질 변수를 토대로 만든 새로운 변수)
설명 변수(= 독립 변수(x)) : 품질도를 제외한 변수

---

5. 대립가설에 따른 설명변수 또는 반응변수의 데이터 가공 계획 및 방법

반응변수

품질도 **:** 와인 품질 예측에 가장 중요한 변수품질 변수를 이용해 새롭게 만든 변수
- 변수를 이용해 새롭게 만든 변수
- quality 변수로 이진 레이블 변환
- 1~5 : 0 (기본 품질)
- 6~10 : 1 (우수 품질)

설명변수
rpart 함수를 이용해 와인의 품질에 영향을 미치는 변수들 **3~4**가지 정할 예정

• **type (**와인의 종류**)**
- 해당 변수는 chr값임
- factor 변수로 변환 (as.factor 사용)

• **fixed acidity (**고정산도**)**
- 산도의 범위가 모델 학습에 영향을 미칠 것이라고 예측
- 연속형 변수의 상대성 중요성 조정 필요
- 스케일링(표준화 또는 정규화) 진행

• **volatile acidity(**휘발성 산도**)**
- 고정 산도와 동일
- 산도 관련 변수들이 비슷한 단위로 학습될 수 있도록 조정
- 스케일링(표준화 또는 정규화) 진행

• **citric acid (**구연산**)**
- 품질에 영향을 미칠 가능성이 있는 변수라고 판단
- 다른 변수들과의 균형 맞춤
- 스케일링(표준화 또는 정규화) 진행

• **residual sugar (**잔류당**)**
- 다른 변수들에 비해 값의 크기가 큼
- 분포를 정규화하고 극단값의 영향을 줄이기 위해 로그 변환 진행

• **chlorides (**염화물**)**
- 적은 값이지만 품질에 영향을 줄 가능성 있음
- 값의 분포 조정을 위해 역수변환 진행
- 큰 값일수록 영향을 적게 주게끔 설정

• **free sulfur dioxide (**자유 이산화황**)**
- 다른 변수들에 비해 분포가 비대칭적일 가능성 있음
- 정규성 개선을 위해 로그 변환 진행

• **total sulfur dioxide (**총 이산화황**)**
- 자유 이산화황과 동일
- 정규성 개선을 위해 로그 변환 진행

• **density (**밀도**)**
- 연속형 변수임
- 스케일링 진행

• **pH (pH)**
- 일정 범위 내 존재하는 값
- 데이터 간 균형을 위해 스케일링 진행

• **sulphates (**황산염**)**
- 작은 값에 치우쳐 있을 가능성이 있는 변수
- 정규성 개선을 위해 로그 변환 진행

• **alcohol (**알코올**)**
- 와인 품질에 영향을 줄 가능성이 가장 높은 변수
- 스케일링 진행

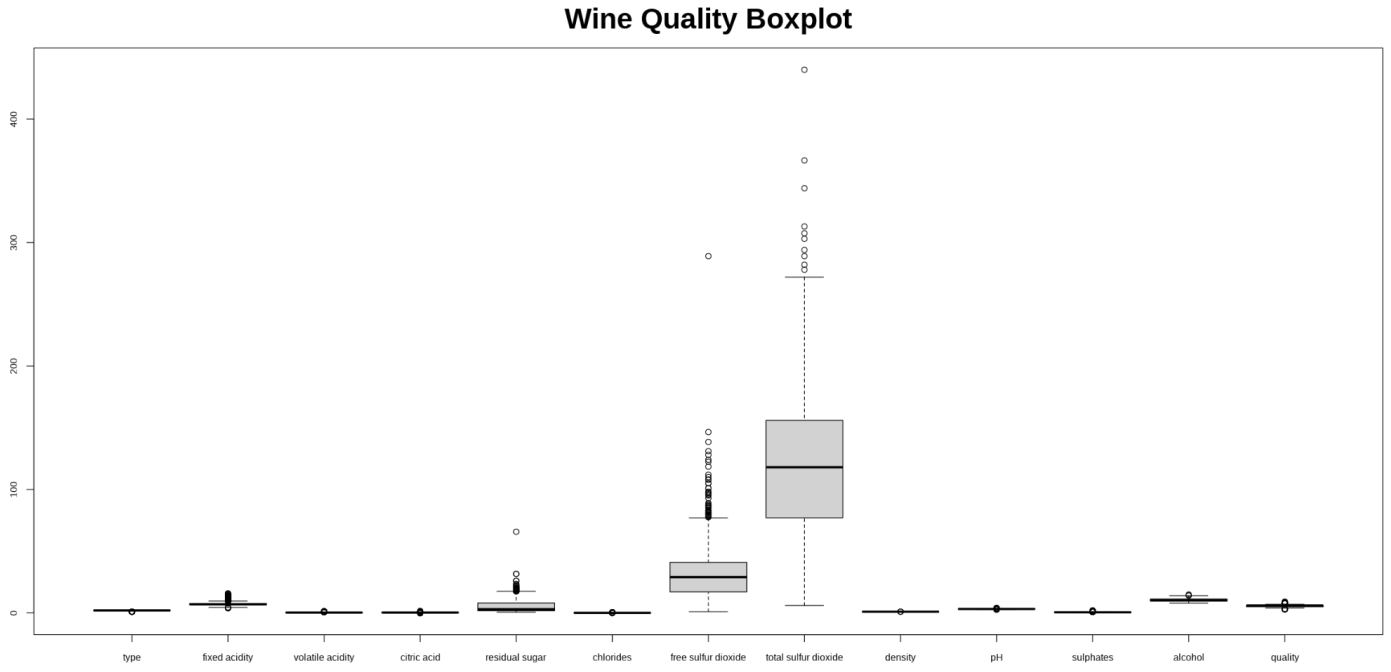
변수들 어떤 방식으로 전처리 할 것인지
ex)원핫 인코딩 / 로그변환 등등

---

6. 이상값**,** 결측값 처리 방법

결측값은 존재하지 않습니다.

이상값의 제외 여부는 데이터의 특성과 분석 목적에 따라 결정되어야한다고 생각합니다. 아래 Boxplot를 보시면 자유_이산화황, 총_이산화황 이 두분이 가장 두드러지는 것을 볼 수 있습니다.



그래서 저희는 지금 이 자리에서는 계획 발표 자리이기 때문에 아직 데이터를 완전히 분석하지 않아서 정확하게 제외를 할지 말지는 추후 프로젝트 결과 발표때 말하겠습니다.

---

7. 모델이 만들어졌을 경우 무엇을 예측하려고 하는지에 대한 설명 이 모델을 뭘 위해 만들었나**?**

저희는 반응변수인 와인의 품질에 영향을 미치는 와인의 성분들의 함량을 기준으로 와인의 품질을 예측하려고 합니다.