📚 🍎 Apple Quality 데이터 분석 실습

> 아이티윌 데이터분석 부트캠프 52기
> 이광호 강사 (leekh4232@gmail.com)

우리는 종종 “이 사과는 좋아 보인다”, “이 사과는 별로다” 라는 판단을 아주 빠르게 내립니다.

그 판단은 과연 무엇에 근거하고 있을까요?

이번 실습에서는 Kaggle의 **Apple Quality 데이터셋**을 활용해 사과의 여러 특성을 바탕으로 **이 사과가 ’좋은 품질’일 확률을 설명하는 모델**을 만들어 보는 것 입니다.

이 과제의 목표는 정확하게 맞히는 것이 아니라, **“왜 이 사과가 좋은 사과일 가능성이 높은지”를 숫자와 말로 설명하는 것**에 있습니다.

※ 본 과제는 팀/개인 단위 모두 수행 가능합니다.

📘 데이터 불러오기

```
load_data("apple_quality")
```

📘 데이터 설명

field	description
A_id	각 과일에 대한 고유 식별자
Size	크기
Weight	무게
Sweetness	단맛 정도
Crunchiness	과일의 아삭한 식감을 나타내는 질감
Juiciness	과일의 과즙 함량 정도
Ripeness	과일이 익은 정도
Acidity	과일의 산도 수준
Quality	과일의 전반적인 품질

📘 과제 수행 미션

📝 1단계. 데이터와 첫인상 - “이 데이터는 어떤 세계를 보여주고 있을까?”

1. 데이터의 전체 구조를 확인하자.
    - 관측치 수
    - 변수 타입 (연속형 / 범주형)
2. 아직 분석을 하지 않은 상태에서 중요해 보이는 변수를 직관적으로 예측해 보자.

> 📌 **출제 의도** 이 단계는 데이터를 바로 분석 대상으로 보기보다, **현실의 대상(사과)을 설명하는 정보로 이해하는 연습**을 하기 위한 단계이다. 변수 이름을 해석하지 못하면 이후 분석도 설득력을 갖기 어렵다.

📝 2단계. 목표 변수 시각화 - “우리가 맞히려는 대상은 어떤 분포를 가지고 있을까?”

1. `Quality`의 분포를 막대그래프로 시각화하자.
2. 좋은 사과와 그렇지 않은 사과의 비율을 확인하자.
3. 이 분포를 보고 정확도 하나만으로 모델을 평가해도 괜찮을지 고민해 보자.

> 📌 **출제 의도** 분류 문제에서는 모델보다 먼저 **무엇을 얼마나 자주 맞혀야 하는지**를 이해하는 것이 중요하다. 이 단계는 성능 지표를 선택하는 이유를 스스로 고민하게 하기 위함이다.

📝 3단계. 단변량 EDA - “사과 하나만 놓고 보았을 때, 무엇이 보일까?”

각 연속형 변수에 대해 다음을 수행하자.

1. 히스토그램과 KDE를 그려 분포를 확인하자.
2. 왜도(skewness)가 있는 변수를 찾아보자.
3. 눈에 띄는 이상치가 있는지 시각적으로 판단하자.

> 📌 **출제 의도** 이 단계는 모델 성능을 높이기 위한 작업이 아니라, **데이터가 가진 기본적인 성질을 이해하는 과정**이다. 분포의 모양을 설명할 수 있다면 충분하다.

📝 4단계. 이변량 EDA - “품질에 따라 무엇이 달라질까?”

1. `Quality`를 기준으로 각 연속형 변수의 분포를 boxplot로 비교하자.
2. 중앙값 차이가 분명한 변수와 분포가 크게 겹치는 변수를 구분하자.
3. 분류에 도움이 될 것 같은 변수를 골라 보자.

> 📌 **출제 의도** 이 단계는 통계 검정 이전에 **그래프를 근거로 판단하는 연습**을 하기 위한 단계이다. “그래서 이 변수가 중요해 보인다”라는 설명이 나오면 충분하다.

📝 5단계. 변수 간 관계 탐색 - “변수들은 서로 독립적일까?”

1. 연속형 변수들의 상관계수를 계산하자.
2. 상관행렬을 히트맵으로 시각화하자.
3. 상관이 높은 변수들을 동시에 사용할 때의 문제점을 고민해 보자.

> 📌 **출제 의도** 여러 변수를 함께 사용할 때 발생할 수 있는 **정보 중복과 해석상의 어려움**을 인식하는 것이 목적이다. 복잡한 용어보다 직관적인 설명이 더 중요하다.

📝 6단계. EDA를 바탕으로 한 모델 설계 - “그려본 결과를 어떻게 모델로 옮길까?”

1. 지금까지의 EDA 결과를 간단히 요약하자.
2. 사용할 독립변수를 최종 선정하자.
3. 각 변수 선택에 대해 다음 문장을 완성하자.

“이 변수는 EDA 단계에서 ㅇㅇ한 패턴을 보였기 때문에 품질을 설명하는 데 도움이 될 것으로 판단했다.”

> 📌 **출제 의도** 변수 선택이 임의가 아니라 **관찰 결과에 근거한 결정**임을 스스로 설명하게 하기 위함이다.

📝 7단계. 로지스틱 회귀 모델 적합 - “확률로 말하는 분류기”

1. 로지스틱 회귀 모델을 적합하자.
2. 회귀계수의 부호와 크기를 확인하자.
3. EDA에서 예상한 방향과 결과가 일치하는지 비교하자.

> 📌 **출제 의도** 모델 결과를 처음부터 해석 대상으로 보기보다, **EDA에서 세운 가설이 어떻게 반영되었는지 확인**하는 과정이다.

📝 8단계. 오즈비를 통한 해석 - “숫자를 다시 사람의 언어로”

1. 회귀계수를 오즈비로 변환하자.
2. 주요 변수 2~3개를 골라 해석 문장을 작성하자.
3. 이 해석이 EDA 결과와 어떻게 연결되는지 설명하자.

> 📌 **출제 의도** 이 단계의 목적은 계산이 아니라 **모델이 말하는 내용을 사람의 언어로 바꾸는 연습**이다.

📝 9단계. 시각화로 보는 모델 성능 - “모델은 어떤 판단을 하고 있을까?”

1. 혼동행렬을 시각화하자.
2. ROC Curve를 그리고 AUC를 계산하자.
3. ROC Curve의 모양을 말로 설명하자.

> 📌 **출제 의도** 성능 지표를 숫자로만 받아들이지 않고, **모델의 판단 특성을 이해하는 도구**로 활용하게 하기 위함이다.

📝 10단계. 같은 조건, 다른 판단 - “모델은 어디서 갈림길을 만들었을까?”

1. 가상의 사과 두 개를 설정하자.
2. 예측 확률이 달라진 이유를 분석하자.
3. EDA → 회귀계수 → 확률의 흐름으로 설명하자.

> 📌 **출제 의도** 이 단계는 분석의 마무리 단계로, **모델의 판단을 대신 설명할 수 있는지**를 확인하기 위한 문제이다.

📘 채점 기준 (총 100점)

단계	배점	평가 포인트	감점 기준
1. 데이터와 첫인상	8	관측치·변수 유형 파악, 변수 의미 해석, 직관적 중요 변수 제시	변수 의미 해석 누락·오해(-2), 중요 변수 직관 언급 없음(-1)
2. 목표 변수 시각화	8	`Quality` 분포·불균형 해석, 지표 선택 필요성 언급	분포 시각화 부재(-3), 불균형/지표 논의 없음(-2)
3. 단변량 EDA	8	분포·왜도·이상치 관찰을 말로 설명	그래프만 제시하고 해석 없음(-3), 왜도/이상치 언급 누락(-2)
4. 이변량 EDA	10	품질별 분포 비교, 겹침/차이 패턴을 근거로 설명	품질 구분 없이 그래프 제시(-3), 겹침/차이 해석 없음(-3)
5. 변수 간 관계	8	상관계수·히트맵 해석, 중복/다중공선성 위험 언급	상관 수치만 제시(-2), 중복 위험 언급 없음(-2)
6. EDA 기반 모델 설계	10	변수 선택 근거를 관찰 결과와 연결해 서술	선택 이유 미기재(-3), EDA 결과와 불일치/설명 없음(-2)
7. 로지스틱 적합	12	모델 적합, 계수 방향/크기 확인, EDA 예상과 비교	계수 해석 없음(-4), EDA 예상과 비교 누락(-2), 적합 결과 미제시(-3)
8. 오즈비 해석	12	주요 변수 오즈비 해석을 사람 언어로 서술, EDA와 연결	오즈비 변환/해석 없음(-4), 사람 언어 설명 부재(-3)
9. 성능 시각화	12	혼동행렬·ROC/AUC 시각화 및 의미 설명	지표만 숫자로 제시(-3), 곡선/혼동행렬 누락(-3), 의미 해석 없음(-2)
10. 같은 조건, 다른 판단	12	가상 사례 확률 차이 설명, EDA→계수→확률 흐름 서술	사례 미제시(-4), 확률 차이 이유 미해석(-3), 흐름 연결 부재(-2)