📚 🚢 Titanic 생존 분석 캡스톤

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

Kaggle의 **Titanic 데이터셋**을 활용해 데이터 전처리 → 탐색적 분석 → 통계적 추론 → 로지스틱 회귀까지 생존 분석의 전 과정을 경험하는 평가 과제입니다.

아래 미션들은 "정답을 맞히는" 예측 대회가 아니라, **"왜 어떤 사람은 살아남고, 왜 어떤 사람은 그렇지 못했는가"를 설명하는 탐구 과제**입니다.

각 미션은 독립적이지만, 모두 합쳐 하나의 질문으로 수렴합니다.

> **"타이타닉에서 누가 살아남았으며, 그 사회적·구조적 이유는 무엇일까?"**

이 과제의 목적은 **"분류모형 = 점수"가 아니라 "분류모형 = 사회적 구조를 설명하는 세계관"**을 만들어 보는 데 있습니다.

※ 본 과제는 팀/개인 단위 모두 수행 가능합니다.

📘 데이터 불러오기

```
load_data("titanic")
```

📘 데이터 설명

관측치: 약 890명 승객 정보

주요 변수

변수명	설명
PassengerId	탑승객의 ID(인덱스와 같은 개념)
Survived	생존유무(0은 사망 1은 생존)
Pclass	객실의 등급(1=1등급, 2=2등급, 3=3등급)
Name	이름
Sex	성별
SibSp	동승한 형제 혹은 배우자의 수
Parch	동승한 자녀 혹은 부모의 수
Ticket	티켓번호
Fare	요금

Cabin	선실
Embarked	탑승지 (C = Cherbourg, Q = Queenstown, S = Southampton)

📘 과제 수행 미션

📝 미션 1. “이 데이터는 얼마나 믿을 수 있을까?”

본격적인 분석 전, 데이터 신뢰도를 점검하자.

- 어떤 변수에 결측치가 얼마나 존재하는가? (`age`, `cabin`, `embarked` 등)
- `fare`가 0이거나 극단적으로 큰 값은 없는가? `age`의 최소·최대는 현실적인가?
- `cabin`과 같이 결측이 많은 변수는 그대로 둘 것인지, 파생 범주(유/무)로 바꿀 것인지?
- 중복 행이나 비상식적 레코드는 없는가?

전처리 전·후를 비교하며 **어떤 기준으로 값을 제거/대체했는지, 그 선택이 결과에 미칠 영향**을 말로 정리하자.

> 📌 **출제 의도** 모델을 돌리기 전에 “이 기록이 믿을 만한가?”를 먼저 묻는 연습입니다. 결측·이상치를 처리한 기준을 말로 남겨야 이후 해석이 설득력을 가집니다.

📝 미션 2. “생존 데이터는 어떤 모습일까?”

`survived`의 분포를 시각화하고, 기초 통계를 확인해 보자.

- 생존자 비율은 얼마인가? (베이스라인: 모두 사망/모두 생존일 때 정확도는?)
- 클래스 불균형 문제는 어느 정도인가? 단순 정확도 지표만으로 충분할까?
- 데이터 수집 방식(생존자 기록의 편향 가능성)을 고려하면 어떤 한계가 있을까?

> 📌 **출제 의도** “무엇을 얼마나 자주 맞혀야 하는가”를 먼저 보는 단계입니다. 베이스라인과 불균형을 이해해야 적절한 성능 지표를 고를 수 있고, 기록 편향을 인식해야 해석에 겸손해집니다.

📝 미션 3. “성별은 생존에 어떤 역할을 했을까?”

성별(`sex`)에 따른 생존률을 비교하고 시각화하자.

- 남녀 생존률 격차의 크기와 방향을 수치로 제시
- 비율 차이에 대한 통계 검정(예: 두 집단 비율 검정)을 적용해 보고, 효과 크기까지 해석
- “여성과 아이 먼저” 규칙이 실제로 작동했는지, 데이터가 이를 뒷받침하는지 논의

> 📌 **출제 의도** 숫자로 차이를 확인하고, “우연이냐 구조냐”를 검정과 효과 크기로 이야기하게 합니다. 역사적 규칙과 데이터가 맞닿는 지점을 직접 해석해 보세요.

📝 미션 4. “객실 등급은 곧 계급일까?”

`pclass`별 생존률과 분포를 살펴보자.

- 1·2·3등실 간 생존률 차이가 얼마나 뚜렷한가?
- 요금(`fare`) 분포와 함께 보면 어떤 계층적 구조가 보이는가?
- 단순 등급 효과인지, 성별·연령과 얽힌 교란 요인인지 고민해 보자.

> 📌 **출제 의도** 숫자 위에 얹힌 사회적 맥락을 보게 하는 단계입니다. 계급·요금·등급이 섞여 있을 때 교란을 의식하며 “왜 이런 서열이 보일까?”를 설명해 봅니다.

📝 미션 5. “나이는 생존에 불리했을까?”

`age` 분포를 전체·생존/사망 그룹으로 나란히 비교하자.

- 어린이/성인/노년층을 어떻게 구분할지 기준을 정하고, 그룹별 패턴을 설명
- 평균·중앙값만으로 충분한가? 분포 꼬리나 이상치가 해석에 미치는 영향은?
- 결측 `age`를 어떻게 처리하는지에 따라 생존률 해석이 어떻게 달라지는지 실험해 보자.

> 📌 **출제 의도** 연속형 변수는 분포와 결측 처리에 따라 이야기가 달라집니다. “연령대 나누기” 같은 선택이 해석을 어떻게 움직이는지 체감해 보라는 의도입니다.

📝 미션 6. “함께 탄 사람은 도움이 되었을까?”

`sibsp`, `parch`를 활용해 동반자 효과를 탐색하자.

- 두 변수를 합쳐 `family_size` 같은 파생 변수를 만들어도 좋다.
- 가족 규모에 따른 생존률 패턴이 단조로운지, 특정 구간에서 뒤집히는지 관찰
- “혼자” vs “가족 동반” 구분이 설명력을 높이는지 논의

> 📌 **출제 의도** 단순 변수 두 개보다, 맥락 있는 파생변수가 설명을 풍부하게 할 수 있음을 경험하게 합니다. “가족 규모”처럼 비선형 패턴을 찾는 눈을 기르는 단계입니다.

📝 미션 7. “어디서 탔는지가 운명을 갈랐을까?”

탑승 항구(`embarked`)에 따른 생존률과 특징을 비교하자.

- 항구별 승객 구성이 다르다면(요금, 등급, 성별), 이것이 생존률 차이에 미친 영향은?
- 단순 비율 비교 외에 교차표·시각화로 패턴을 보여주자.
- 항구 정보가 결측인 경우 어떻게 처리했는지, 그 결정이 결과에 주는 영향도 기록

> 📌 **출제 의도** 같은 범주라도 안에 누가 타고 있는지 구성이 다르면 해석이 달라집니다. 결측 처리 선택까지 포함해 “왜 이 차이가 나왔는지”를 말로 남기는 연습입니다.

📝 미션 8. “어떤 변수들이 생존과 가장 가까울까?”

연속형·범주형 변수를 구분해 적절한 방법으로 `survived`와의 관계를 탐색하자.

- 연속형(`age`, `fare`)은 분포·박스플롯·상자곱비율 등으로 경향을 보고, 비선형성을 체크
- 범주형(`sex`, `pclass`, `embarked`, `cabin 유무`, `family_size 구간`)은 교차표와 시각화
- 단순 상관계수 외에 **어떤 통계량이 관계를 잘 설명하는지** 선택 이유를 남기자.

> 📌 **출제 의도** 변수 성격에 맞는 도구를 고르고, 왜 그 도구를 골랐는지 설명해 보는 단계입니다. 상관계수 하나에 기대기보다 “이 관계를 보여주기 좋은 지표/그래프가 뭘까?”를 스스로 결정해 보세요.

📝 미션 9. “상관관계는 곧 원인일까?”

지금까지의 탐색 결과를 두고 해석의 한계를 짚어보자.

- 예: `sex` 효과가 실제 구조인지, 구조적·문화적 규칙(대피 우선순위)의 결과인지?
- `pclass`와 `fare`처럼 얽힌 변수들 사이에서 인과성을 말할 수 있는가?
- 데이터 수집 편향, 측정 누락(직업, 건강상태 등)이 결론에 주는 한계를 서술

> 📌 **출제 의도** “상관이 높다 = 원인이다”라는 함정을 피하기 위한 단계입니다. 기록되지 않은 요인과 구조적 규칙을 인정하며, 결론의 한계를 스스로 적어보게 합니다.

📝 미션 10. “생존을 설명하는 로지스틱 회귀모형을 만들어보자”

`survived`를 종속변수로 하는 로지스틱 회귀모형을 설계하자.

- 어떤 변수를 선택/제외했는지, 그 이유를 “설명하고 싶어서”라는 관점으로 명시
- 범주형 인코딩 방식, 파생변수(`family_size`, `cabin 존재`, `age*sex`, `pclass*sex` 등) 설계
- 학습/검증 데이터 분리 방식과 클래스 불균형을 다루는 방법(가중치, 언더/오버샘플링) 고민

> 📌 **출제 의도** “넣을 수 있어서”가 아니라 “설명하고 싶어서” 넣는 모델을 설계하는 연습입니다. 인코딩·불균형 대응까지 포함해 선택의 이유를 언어로 남기도록 합니다.

📝 미션 11. “이 모형은 믿을 만할까?”

모형을 진단하고, 숫자를 넘어 의미를 해석하자.

- 혼동행렬, 정확도 외에 재현율·정밀도·F1·ROC-AUC·PR-AUC를 함께 제시
- 회귀계수(로그 오즈) 해석: 어떤 변수가 생존 확률을 얼마나 변화시키는가?
- 모형 가정 점검: 선형성(로그 오즈), 다중공선성, 영향력 큰 관측치, 캘리브레이션 곡선
- 개선 아이디어: 변수 변환, 상호작용 추가, 비선형 모형 비교 등

> 📌 **출제 의도** 점수 하나로 끝내지 않고, 여러 지표와 가정 진단을 통해 “이 모델이 어디서 잘하고 어디서 부족한지”를 말로 설명하게 하는 단계입니다.

📝 미션 12. “같은 조건인데 왜 결과는 달랐을까?”

지금까지의 분석을 엮어 한 편의 이야기로 마무리하자.

- 가상의 두 승객을 설정하고, 예측 확률을 제시한 뒤 **비전공자에게 설명하듯** 서술
- “왜 이 사람이 더 높은 확률을 갖게 되었는가?”를 사회적·구조적 맥락과 연결
- 모델이 설명하지 못한 요인(운, 기록 누락 등)도 함께 언급해 설득력을 높이자.

> 📌 **출제 의도** 숫자를 사람의 언어로 풀어내는 마무리입니다. 모델이 말하는 것과 말하지 못하는 것을 구분해, 비전공자도 납득할 이야기를 만드는 연습입니다.

📘 채점 기준 (총 100점)

미션	배점	평가 포인트	감점 기준
1. 데이터 신뢰도	8	결측·이상치·편향 확인, 처리 기준과 영향 서술	결측/이상치 언급 누락(-3), 처리 기준·영향 설명 없음(-2)
2. 생존 분포	6	생존/사망 비율·불균형 해석, 지표 선택 논리 제시	분포 시각화 없이 수치만 제시(-2), 불균형·지표 논의 없음(-2)
3. 성별 효과	6	생존률 격차 시각화·비율 검정, 효과 크기와 역사적 맥락 해석	검정/효과 크기 미제시(-2), 맥락 해석 없음(-2), 시각화 부재(-1)
4. 객실 등급	8	pclass별 패턴, 교차분석/검정 근거, 계층적 해석	교차분석·검정 없음(-3), 계층/교란 논의 없음(-2), 그래프만 있고 해석 없음(-2)
5. 나이	7	연령대 구분·분포 비교, 결측 처리 영향 설명	연령대 기준 미제시(-2), 결측 처리 영향 언급 없음(-2), 해석 없이 그래프만 제시(-2)
6. 가족 동반	7	family_size 파생·패턴 해석, 비선형/역전 구간 언급	파생 변수 미사용(-2), 패턴 해석 없음(-2), 비선형/역전 논의 부재(-1)
7. 탑승 항구	7	embarked별 생존률·구성 차이, 결측 처리 영향 기록	결측 처리 설명 없음(-2), 구성 차이 고려 없이 비율만 제시(-2), 시각화/해석 부재(-2)
8. 변수-생존 탐색	7	변수 유형별 적합한 통계/시각화 선택 근거와 해석	변수 유형 구분 없이 상관만 제시(-2), 선택 근거 없음(-2), 해석 미흡(-2)
9. 상관 vs 인과	6	상관 해석 한계, 누락 요인·구조적 규칙 언급	한계/누락 요인 언급 없음(-3), 상관=인과로 단정(-3)
10. 로지스틱 설계	14	변수 선택·인코딩·불균형 대응 논리, 계수 방향 예상	선택·인코딩·불균형 처리 이유 미기재(-4), 계수 방향 예상/설명 없음(-3), 설계 재현 불가(-3)
11. 모형 진단	12	혼동행렬·ROC/PR, 가정·캘리브레이션 점검과 개선안	지표만 나열하고 해석 없음(-3), 곡선/혼동행렬 누락(-3), 가정/개선안 언급 없음(-3)
12. 스토리텔링	12	가상 승객 사례 설명, 사회·구조 맥락 연결, 모델 한계 포함	사례 미제시(-4), 확률 차이 이유 해석 없음(-3), 한계·맥락 언급 없음(-3)