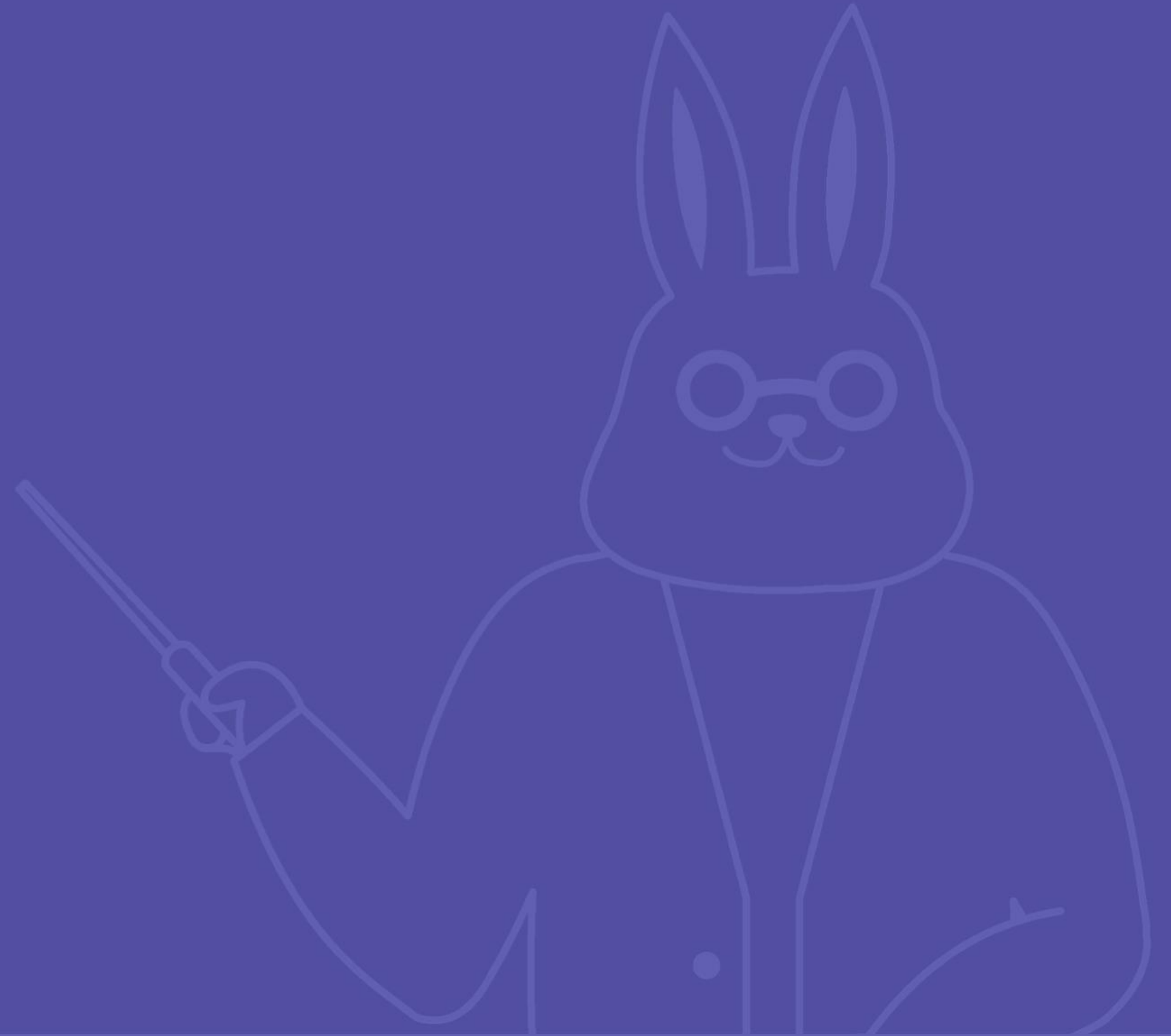




머신러닝 시작하기

02 데이터 전 처리하기



목차

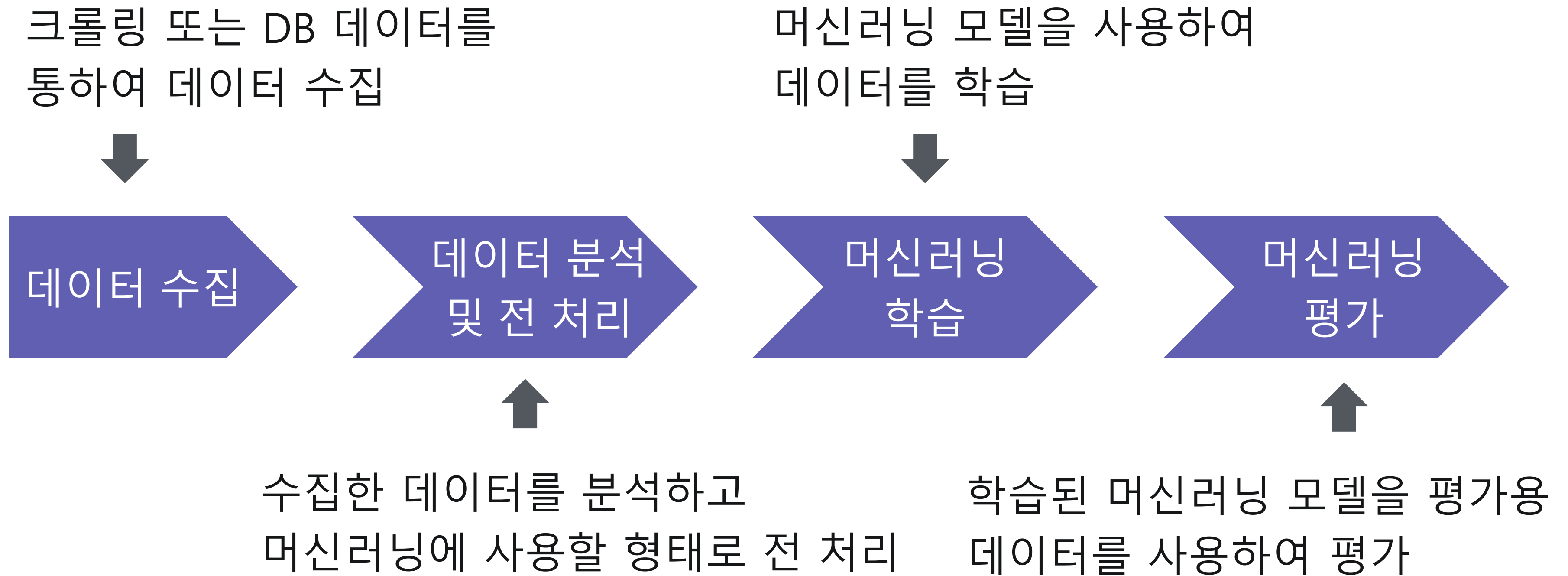
01. 머신러닝을 위한 데이터 전 처리 이해하기
02. 범주형 자료 전 처리
03. 수치형 자료 전 처리
04. 데이터 정제 및 분리하기

01

머신러닝을 위한 데이터 전 처리 이해하기



✓ 머신러닝 과정 이해하기



✓ 데이터 전 처리의 역할

- 1) 머신러닝의 입력 형태로 **데이터 변환 (특성 엔지니어링)**
- 2) 결측값 및 이상치를 처리하여 **데이터 정제**
- 3) 학습용 및 평가용 **데이터 분리**

✔ 왜 데이터 전 처리가 필요할까? - 데이터 변환

대부분의 머신러닝 모델은 숫자 데이터를 입력 받는다

일반적으로 행렬 형태 입력

| | 변수 1 | 변수 2 | 변수 3 | ... | 변수 12 |
|---------|------|------|-------|-----|-------|
| 샘플 1 | 3 | 12.9 | 0.012 | ... | 6 |
| 샘플 2 | 5 | 20.7 | 0.1 | ... | 3 |
| | | ⋮ | | ⋱ | ⋮ |
| 샘플 1000 | 7 | 32.1 | 0.098 | ... | 3 |

➡

머신러닝 모델

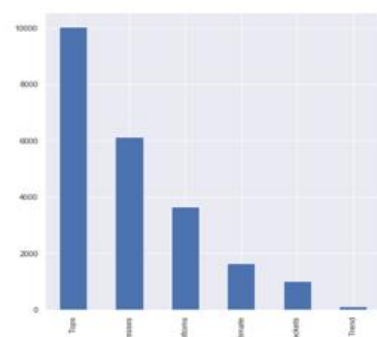
✓ 왜 데이터 전 처리가 필요할까? - 데이터 변환

실제 데이터는 머신러닝 모델이 이해할 수 없는 형태로 되어 있음

실제 데이터 set



이미지 데이터



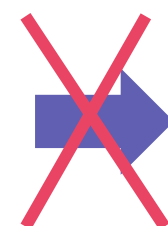
범주형 데이터



자연어 데이터



시계열 데이터



머신러닝 모델

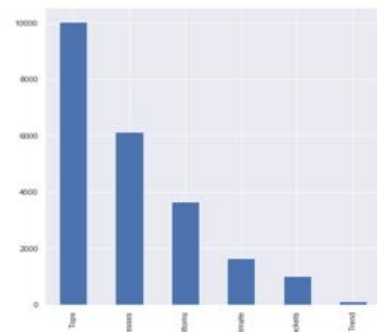
✔ 왜 데이터 전 처리가 필요할까? - 데이터 변환

전 처리를 통하여 머신러닝 모델이 이해할 수 있는 수치형 자료로 변환

실제 데이터 set



이미지 데이터



범주형 데이터



자연어 데이터



시계열 데이터



데이터 전처리



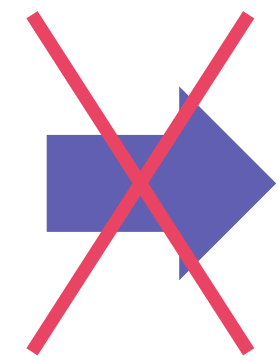
머신러닝 모델

✔ 왜 데이터 전 처리가 필요할까? - 데이터 정제

전 처리를 통하여 데이터의 결측값 및 이상치를 처리

결측값과 이상치가 있는 데이터

| Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|------|-------|-------|------------------|---------|-------|----------|
| 22.0 | 1 | 0 | A/5 21171 | 7.2500 | NaN | S |
| 38.0 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| 26.0 | 0 | 0 | STON/O2. 3101282 | 7.9250 | NaN | S |
| 35.0 | 1 | 0 | 113803 | 53.1000 | C123 | S |
| 35.3 | 0 | 0 | 373450 | 8.0500 | NaN | S |



머신러닝
모델

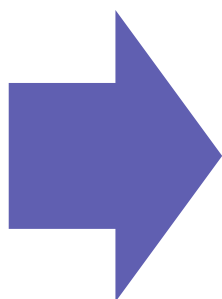
✓ 왜 데이터 전 처리가 필요할까? - 데이터 분리

전 처리를 통하여 학습용과 평가용 데이터를 분리

원본 데이터 (150 샘플)

| | 꽃받침 길이 | 꽃받침 넓이 | 꽃잎 길이 | 꽃잎 넓이 | 클래스 |
|-----|--------|--------|-------|-------|-----|
| 0 | 5.1 | 3.5 | 1.4 | 0.2 | 0 |
| 1 | 4.9 | 3.0 | 1.4 | 0.2 | 0 |
| 2 | 4.7 | 3.2 | 1.3 | 0.2 | 0 |
| 3 | 4.6 | 3.1 | 1.5 | 0.2 | 0 |
| 4 | 5.0 | 3.6 | 1.4 | 0.2 | 0 |
| ... | ... | ... | ... | ... | ... |
| 145 | 6.7 | 3.0 | 5.2 | 2.3 | 2 |
| 146 | 6.3 | 2.5 | 5.0 | 1.9 | 2 |
| 147 | 6.5 | 3.0 | 5.2 | 2.0 | 2 |
| 148 | 6.2 | 3.4 | 5.4 | 2.3 | 2 |
| 149 | 5.9 | 3.0 | 5.1 | 1.8 | 2 |

150 rows × 5 columns



학습용 데이터
(100 샘플)

| | 꽃받침 길이 | 꽃받침 넓이 | 꽃잎 길이 | 꽃잎 넓이 | 클래스 |
|-----|--------|--------|-------|-------|-----|
| 96 | 5.7 | 2.9 | 4.2 | 1.3 | 1 |
| 105 | 7.6 | 3.0 | 6.6 | 2.1 | 2 |
| 66 | 5.6 | 3.0 | 4.5 | 1.5 | 1 |
| ... | ... | ... | ... | ... | ... |
| 71 | 6.1 | 2.8 | 4.0 | 1.3 | 1 |
| 106 | 4.9 | 2.5 | 4.5 | 1.7 | 2 |
| 14 | 5.8 | 4.0 | 1.2 | 0.2 | 0 |

100 rows × 5 columns

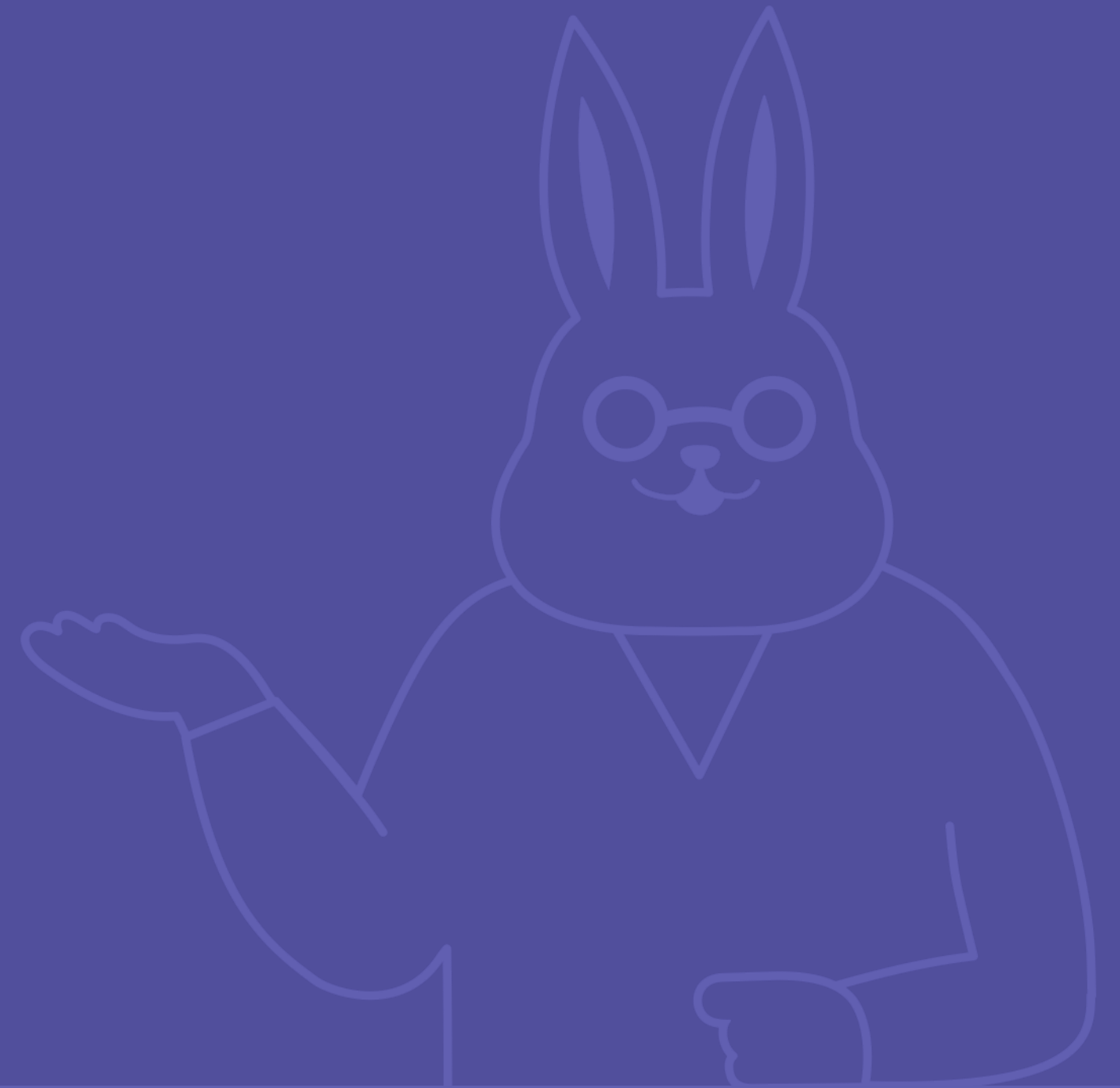
평가용 데이터
(50 샘플)

| | 꽃받침 길이 | 꽃받침 넓이 | 꽃잎 길이 | 꽃잎 넓이 | 클래스 |
|-----|--------|--------|-------|-------|-----|
| 73 | 6.1 | 2.8 | 4.7 | 1.2 | 1 |
| 18 | 5.7 | 3.8 | 1.7 | 0.3 | 0 |
| 118 | 7.7 | 2.6 | 6.9 | 2.3 | 2 |
| ... | ... | ... | ... | ... | ... |
| 19 | 5.1 | 3.8 | 1.5 | 0.3 | 0 |
| 56 | 6.3 | 3.3 | 4.7 | 1.6 | 1 |
| 104 | 6.5 | 3.0 | 5.8 | 2.2 | 2 |

50 rows × 5 columns

02

범주형 자료 전 처리



✔ 타이타닉 생존자 데이터 살펴보기

< 타이타닉 생존자 데이터 변수 확인 >

| 변수 명 | 변수 설명 |
|-------------|--|
| PassengerId | 각 승객의 고유 번호 |
| Survived | 생존 여부 (0: 사망, 1: 생존) |
| Pclass | 객실 등급 (1st: Upper, 2nd: Middle, 3rd: Lower) |
| Name | 이름 |
| Sex | 성별 |
| Age | 나이 |
| SibSp | 동반한 형제자매와 배우자의 수 |
| Parch | 동반한 부모, 자식의 수 |
| Ticket | 티켓의 고유 번호 |
| Fare | 티켓의 요금 |
| Cabin | 객실 번호 |
| Embarked | 승선한 항(C: Cherbourg, Q: Queenstown, S: Southampton) |

✓ 범주형 자료 살펴보기

범주형 데이터는 몇 개의 범주로 나누어진 자료

< 타이타닉 생존자 데이터 >

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|-------------|----------|--------|---|--------|------|-------|-------|------------------|---------|-------|----------|
| 0 | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | NaN | S |
| 1 | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| 2 | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.9250 | NaN | S |
| 3 | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 | 0 | 113803 | 53.1000 | C123 | S |
| 4 | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.0 | 0 | 0 | 373450 | 8.0500 | NaN | S |

✓ 범주형 자료 살펴보기

범주의 크기가 의미 없다면 **명목형 자료**

크기가 의미 있다면 **순서형 자료**

< 타이타닉 생존자 데이터 >

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|-------------|----------|--------|---|--------|------|-------|-------|------------------|---------|-------|----------|
| 0 | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | NaN | S |
| 1 | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| 2 | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.9250 | NaN | S |
| 3 | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 | 0 | 113803 | 53.1000 | C123 | S |
| 4 | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.0 | 0 | 0 | 373450 | 8.0500 | NaN | S |

명목형 자료
순서형 자료

✓ 범주형 자료 변환 방식

대표적인 범주형 자료 변환 방식

명목형 자료:

- 수치 맵핑 방식
- 더미(Dummy) 기법

순서형 자료:

- 수치 맵핑 방식

✓ 범주형 자료 변환하기

1) 명목형 자료 변환하기 – 수치 맵핑 변환

- 일반적으로 범주를 0, 1로 맵핑
- (-1, 1), (0, 100) 등 다양한 케이스가 있지만 모델에 따라 성능이 달라질 수 있음

< 성별(Sex) 데이터 변환 예 >

| Sex | Age | SibSp |
|--------|------|-------|
| male | 22.0 | 1 |
| female | 38.0 | 1 |
| female | 26.0 | 0 |
| female | 35.0 | 1 |
| male | 35.0 | 0 |

변환 전

| Sex | Age | SibSp |
|-----|------|-------|
| 0 | 22.0 | 1 |
| 1 | 38.0 | 1 |
| 1 | 26.0 | 0 |
| 1 | 35.0 | 1 |
| 0 | 35.0 | 0 |

male -> 0, female-> 1 변환 후

✓ 범주형 자료 변환하기

1) 명목형 자료 변환하기 – 수치 맵핑 변환

- 3개 이상인 경우, 수치의 크기 간격을 같게 하여 수치 맵핑 ex) (0,1,2,3,...)

< Embarked 데이터 변환 예 >

| Fare | Cabin | Embarked |
|---------|-------|----------|
| 7.2500 | NaN | S |
| 71.2833 | C85 | C |
| 7.9250 | NaN | S |
| 53.1000 | C123 | S |
| 8.0500 | NaN | S |
| 90.0 | C78 | Q |

변환 전

| Fare | Cabin | Embarked |
|---------|-------|----------|
| 7.2500 | NaN | 0.0 |
| 71.2833 | C85 | 2.0 |
| 7.9250 | NaN | 0.0 |
| 53.1000 | C123 | 0.0 |
| 8.0500 | NaN | 0.0 |
| 90.0 | C78 | 1.0 |

S->0, Q->1, C->2 변환 후

✓ 범주형 자료 변환하기

2) 명목형 자료 변환하기 – 더미(Dummy) 기법

- 더미 기법을 사용하여 **각 범주**를 **0 or 1**로 변환

< 더미 변환 예 >

| | Age | Pclass | SibSp | Parch | Fare | Sex_female | Sex_male | Embarked_C | Embarked_Q | Embarked_S |
|-----|------|--------|-------|-------|---------|------------|----------|------------|------------|------------|
| 0 | 22.0 | 3 | 1 | 0 | 7.2500 | 0 | 1 | 0 | 0 | 1 |
| 1 | 38.0 | 1 | 1 | 0 | 71.2833 | 1 | 0 | 1 | 0 | 0 |
| 2 | 26.0 | 3 | 0 | 0 | 7.9250 | 1 | 0 | 0 | 0 | 1 |
| 3 | 35.0 | 1 | 1 | 0 | 53.1000 | 1 | 0 | 0 | 0 | 1 |
| 4 | 35.0 | 3 | 0 | 0 | 8.0500 | 0 | 1 | 0 | 0 | 1 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 885 | 39.0 | 3 | 0 | 5 | 29.1250 | 1 | 0 | 0 | 1 | 0 |
| 886 | 27.0 | 2 | 0 | 0 | 13.0000 | 0 | 1 | 0 | 0 | 1 |
| 887 | 19.0 | 1 | 0 | 0 | 30.0000 | 1 | 0 | 0 | 0 | 1 |
| 889 | 26.0 | 1 | 0 | 0 | 30.0000 | 0 | 1 | 1 | 0 | 0 |
| 890 | 32.0 | 3 | 0 | 0 | 7.7500 | 0 | 1 | 0 | 1 | 0 |

✓ 범주형 자료 변환하기

3) 순서형 자료 변환하기 – 수치 맵핑 변환

- 수치에 맵핑하여 변환하지만, 수치 간 크기 차이는 커스텀 가능
- 크기 차이가 머신러닝 결과에 영향을 끼칠 수 있음

< 순서형 자료 변환 예 >

| | feature_1 | feature_2 | feature_3 |
|---|-----------|-----------|-----------|
| 0 | 1.2 | 2 | 매우 많음 |
| 1 | 0.1 | 1 | 없음 |
| 2 | -0.1 | 3 | 조금 많음 |

변환 전

| | feature_1 | feature_2 | feature_3 |
|---|-----------|-----------|-----------|
| 0 | 1.2 | 2.0 | 10.0 |
| 1 | 0.1 | 1.0 | 0.0 |
| 2 | -0.1 | 3.0 | 4.0 |

없음->0, 조금 많음->4, 매우 많음->10 변환 후

03

수치형 자료 전 처리



수치형 자료 살펴보기

크기를 갖는 수치형 값으로 이루어진 데이터

< 타이타닉 생존자 데이터 >

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|-------------|----------|--------|---|--------|------|-------|-------|------------------|---------|-------|----------|
| 0 | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | NaN | S |
| 1 | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| 2 | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.9250 | NaN | S |
| 3 | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 | 0 | 113803 | 53.1000 | C123 | S |
| 4 | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.0 | 0 | 0 | 373450 | 8.0500 | NaN | S |

✓ 수치형 자료 살펴보기

머신러닝의 입력으로 바로 사용할 수 있으나,
모델의 성능을 높이기 위해서 데이터 변환이 필요

대표적인 수치형 자료 변환 방식

- 1) 스케일링(Scaling) - 정규화(Normalization), 표준화(Standardization)
- 2) 범주화

✓ 수치형 자료 변환하기 – 정규화 방식

스케일링(Scaling)

- 변수 값의 범위 및 크기를 변환하는 방식
- 변수(feature) 간의 범위가 차이가 나면 사용

1) 정규화(Normalization)

변수 X 를 정규화한 값 X'

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

✓ 정규화를 적용한 데이터 예시

1) 정규화(Normalization)

| | feature_1 | feature_2 | feature_3 |
|----|-----------|-----------|------------|
| 0 | 1.280187 | -1.156924 | -81.977837 |
| 1 | 0.519024 | 0.277231 | -78.493732 |
| 2 | -1.340744 | 0.564647 | 51.682415 |
| 3 | 0.880929 | 1.037069 | 45.883654 |
| 4 | -1.260126 | 1.257954 | 15.080874 |
| 5 | 0.401379 | -1.310234 | 90.150390 |
| 6 | -1.142048 | 0.243710 | 57.606259 |
| 7 | 0.566775 | -0.396015 | 64.846291 |
| 8 | -0.724533 | -0.510327 | -5.383149 |
| 9 | -1.615751 | -0.056775 | 130.638733 |
| 10 | -0.721374 | -0.627100 | 108.228715 |

변환 전

| | feature_1 | feature_2 | feature_3 |
|----|-----------|-----------|-----------|
| 0 | 1.280187 | -1.156924 | 0.000000 |
| 1 | 0.519024 | 0.277231 | 0.016387 |
| 2 | -1.340744 | 0.564647 | 0.628645 |
| 3 | 0.880929 | 1.037069 | 0.601371 |
| 4 | -1.260126 | 1.257954 | 0.456496 |
| 5 | 0.401379 | -1.310234 | 0.809571 |
| 6 | -1.142048 | 0.243710 | 0.656506 |
| 7 | 0.566775 | -0.396015 | 0.690558 |
| 8 | -0.724533 | -0.510327 | 0.360248 |
| 9 | -1.615751 | -0.056775 | 1.000000 |
| 10 | -0.721374 | -0.627100 | 0.894599 |

정규화 변환 후

✓ 수치형 자료 변환하기 – 표준화 방식

스케일링(Scaling)

- 변수 값의 범위 및 크기를 변환하는 방식
- 변수(feature) 간의 범위가 차이가 나면 사용

2) 표준화(Standardization)

변수 X 를 표준화한 값 X'

$$X' = \frac{X - \mu}{\sigma}$$

✓ 표준화를 적용한 데이터 예시

2) 표준화(Standardization)

| | feature_1 | feature_2 | feature_3 |
|----|-----------|-----------|------------|
| 0 | 1.280187 | -1.156924 | -81.977837 |
| 1 | 0.519024 | 0.277231 | -78.493732 |
| 2 | -1.340744 | 0.564647 | 51.682415 |
| 3 | 0.880929 | 1.037069 | 45.883654 |
| 4 | -1.260126 | 1.257954 | 15.080874 |
| 5 | 0.401379 | -1.310234 | 90.150390 |
| 6 | -1.142048 | 0.243710 | 57.606259 |
| 7 | 0.566775 | -0.396015 | 64.846291 |
| 8 | -0.724533 | -0.510327 | -5.383149 |
| 9 | -1.615751 | -0.056775 | 130.638733 |
| 10 | -0.721374 | -0.627100 | 108.228715 |

변환 전

| | feature_1 | feature_2 | feature_3 |
|----|-----------|-----------|-----------|
| 0 | 1.280187 | -1.156924 | -1.707156 |
| 1 | 0.519024 | 0.277231 | -1.656828 |
| 2 | -1.340744 | 0.564647 | 0.223561 |
| 3 | 0.880929 | 1.037069 | 0.139798 |
| 4 | -1.260126 | 1.257954 | -0.305147 |
| 5 | 0.401379 | -1.310234 | 0.779229 |
| 6 | -1.142048 | 0.243710 | 0.309130 |
| 7 | 0.566775 | -0.396015 | 0.413712 |
| 8 | -0.724533 | -0.510327 | -0.600749 |
| 9 | -1.615751 | -0.056775 | 1.364081 |
| 10 | -0.721374 | -0.627100 | 1.040369 |

표준화 변환 후

✓ 수치형 자료 변환하기 - 범주화

범주화

- 변수의 값보다 범주가 중요한 경우 사용

| 시험 점수 | |
|-------|-----|
| 0 | 12 |
| 1 | 100 |
| 2 | 20 |
| 3 | 35 |
| 4 | 92 |
| 5 | 53 |
| 6 | 62 |
| 7 | 78 |
| 8 | 5 |
| 9 | 90 |
| 10 | 54 |



평균 : 54.63
 평균 이상 -> 1
 평균 이하 -> 0

| 시험 점수 | |
|-------|---|
| 0 | 0 |
| 1 | 1 |
| 2 | 0 |
| 3 | 0 |
| 4 | 1 |
| 5 | 0 |
| 6 | 1 |
| 7 | 1 |
| 8 | 0 |
| 9 | 1 |
| 10 | 0 |

04

데이터 정제 및 분리하기



✓ 결측값(Missing data) 처리하기

일반적인 머신러닝 모델의 입력 값으로 결측값을 사용할 수 없음
따라서 **Null, None, NaN** 등의 결측값을 **처리** 해야함

대표적인 결측값 처리 방식

- 1) 결측값이 존재하는 **샘플 삭제**
- 2) 결측값이 많이 존재하는 **변수 삭제**
- 3) 결측값을 **다른 값으로 대체**

✔ 이상치(Outlier) 처리하기

이상치가 있으면, 모델의 **성능을 저하**할 수 있음
이상치는 일반적으로 전 처리 과정에서 **제거**하며,
어떤 값이 이상치 인지 판단하는 기준이 중요함

이상치 판단 기준 방법

- 1) 통계 지표(카이제곱 검정, IQR 지표 등)를 사용하여 판단
- 2) 데이터 분포를 보고 직접 판단
- 3) 머신러닝 기법을 사용하여 이상치 분류

✓ 데이터 분리는 왜 필요할까?

머신러닝 모델을 **평가**하기 위해서는 **학습에 사용하지 않은 평가용 데이터**가 필요
약 7:3 ~ 8:2 비율로 학습용 평가용 데이터를 분리함

학습용
데이터

평가용
데이터

✓ 지도학습 데이터 분리

지도학습의 경우 feature 데이터와 label 데이터를 분리하여 저장합니다.

Feature 데이터: label을 예측하기 위한 입력 값

Label 데이터: 예측해야 할 대상이 되는 데이터

✔ 지도학습 데이터 분리

타이타닉 데이터를 바탕으로 생존자를 예측한다면?

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|-------------|----------|--------|---|--------|------|-------|-------|------------------|---------|-------|----------|
| 0 | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | NaN | S |
| 1 | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| 2 | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.9250 | NaN | S |
| 3 | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 | 0 | 113803 | 53.1000 | C123 | S |
| 4 | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.0 | 0 | 0 | 373450 | 8.0500 | NaN | S |

✔ 지도학습 데이터 분리

Feature 데이터: 승객 나이, 가족 정보, 표 가격 등등

| | PassengerId | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|-------------|--------|---|--------|------|-------|-------|------------------|---------|-------|----------|
| 0 | 1 | 3 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | NaN | S |
| 1 | 2 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| 2 | 3 | 3 | Heikkinen, Miss. Laina | female | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.9250 | NaN | S |
| 3 | 4 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 | 0 | 113803 | 53.1000 | C123 | S |
| 4 | 5 | 3 | Allen, Mr. William Henry | male | 35.0 | 0 | 0 | 373450 | 8.0500 | NaN | S |

Label 데이터: 생존 여부

| | Survived |
|---|----------|
| 0 | 0 |
| 1 | 1 |
| 2 | 1 |
| 3 | 1 |
| 4 | 0 |