

**SK네트웍스 Family AI 과정 15기  
 데이터 전처리 인공지능 데이터 전처리 결과서**



| **산출물 단계** | 데이터 전처리 |
| --- | --- |
| **평가 산출물** | 인공지능 데이터 전처리 결과서 |
| **제출 일자** |  |
| **깃허브 경로** | 깃허브 주소 기재 |
| **작성 팀원** |  |

1. **문서 개요**

* 프로젝트명:
* 전처리 목적: (예: 텍스트 분류 모델 학습용 데이터 정제)
* 문제 정의:  
   예: 사용자 채팅 로그를 기반으로 상담 태그를 분류하는 모델을 학습하기 위한 데이터셋 구축

1. **데이터셋 개요**

* 데이터 출처 및 수집 방법:  
   예: LLM 챗봇 로그 / 자체 수집 / 사용자 동의 기반 저장
* 데이터 구성:

| **항목명** | **설명** | **예시** |
| --- | --- | --- |
| user\_id | 사용자 식별자 | u\_1043921 |
| message | 사용자 질문 | “오늘 날씨 어때요?” |
| timestamp | 입력 시각 | 2025-05-01 14:21:03 |
| tag | 상담 주제 | 날씨 |

* 원본 데이터 샘플(5~10건 첨부):  
   (스크린샷 또는 테이블 형태)

1. **전처리 프로세스 개요**

* 전체 흐름도:  
    
   ① 수집 → ② 결측치 처리 → ③ 이상치 탐지 → ④ 정규화 → ⑤ 데이터 분리
* 전처리 파이프라인 요약:

| **단계** | **목적** | **수행 작업** | **사용 도구/라이브러리** |
| --- | --- | --- | --- |
| 결측치 처리 | 누락값 제거 | Null 행 제거, 특수값 대체 | pandas |
| 이상치 처리 | 비정상 데이터 제거 | 단어 수 기준, 이상시간 필터링 | numpy |
| 정규화 | 텍스트 전처리 | 소문자 변환, 불용어 제거 | nltk |
| 표준화 | 수치 표준화 | z-score 정규화 | sklearn |
| 분리 | 학습/검증 분할 | train:test = 8:2 | train\_test\_split |

1. **세부 전처리 단계**

4.1 결측치 처리

* 결측치 존재 여부: 있음 / 없음
* 결측 컬럼 및 비율:

| **컬럼명** | **결측률** | **처리 방법** |
| --- | --- | --- |
| tag | 5.1% | ‘기타’로 대체 |
| message | 0.8% | 해당 행 제거 |

* 코드 예시 (Python):

df = df.dropna(subset=['message'])  
 df['tag'] = df['tag'].fillna('기타')

4.2 이상치 처리

* 정의한 이상치 기준:  
   예: 메시지 길이 1자 이하 / timestamp가 미래 날짜일 경우 등
* 처리 방식 및 영향:

| **항목** | **기준** | **처리 방식** | **제거 수** |
| --- | --- | --- | --- |
| message\_length | < 2자 | 해당 행 제거 | 42건 |
| timestamp | 2026년 초과 | 해당 행 제거 | 5건 |

4.3 정규화 및 표준화

* 텍스트 정규화:  
  + 소문자 변환: “Hi There” → “hi there”
  + 특수문자 제거: “hello!!” → “hello”
  + 불용어 제거: “is, are, the” 등 제외
* 수치형 표준화: z-score, Min-Max
* 사용 라이브러리: re, nltk, sklearn.preprocessing

4.4 데이터 변환 및 생성

* 레이블 인코딩: tag 컬럼 → 숫자 레이블

from sklearn.preprocessing import LabelEncoder  
 encoder = LabelEncoder()  
 df['tag\_id'] = encoder.fit\_transform(df['tag'])

* 파생 변수 생성 (예: 메시지 길이):

df['msg\_length'] = df['message'].apply(len)

1. **학습/검증 데이터 분리**

* 분리 기준 및 방법:  
  + 기준: 무작위 분할
  + 비율: Train 80% / Test 20%
* 분리 코드:

from sklearn.model\_selection import train\_test\_split  
 train, test = train\_test\_split(df, test\_size=0.2, random\_state=42)

* 분리 후 건수:

| 구분 | 데이터 수 |
| --- | --- |
| 학습 데이터 | 8,000건 |
| 테스트 데이터 | 2,000건 |

1. **전처리 결과 요약 및 평가**

* 전처리 후 전체 건수: 10,000건 → 9,842건
* 품질 향상 지표:  
  + 결측값 제거: 158건 정제
  + 이상치 제거: 총 47건 제외
  + 레이블 정리 및 불균형 개선
* 향후 활용 방안:  
  + 텍스트 분류, 대화 응답 분기 조건 설계, LLM 학습데이터 구축 등에 사용

1. **변경 이력**

| 변경일 | 변경자 | 변경 내용 | 비고 |
| --- | --- | --- | --- |
| 2025-05-10 | 홍길동 | tag 결측치 ‘기타’ 대체 | 전처리 안정성 확보 목적 |