데이터 전처리 파이프라인 구현 및 문서화

1. 개요 - Kaggle의 타이타닉 데이터를 이용하여 생존자 예측 모델링을 위한 데이터 전처리 파이프라인 구현 및 문서화를 진행합니다
    a. 목표 : 머신러닝 알고리즘 중 로지스틱 회귀를 선택하여 원본 데이터를 결측치가 없고 수치화 정규화된 데이터로 변환하는 전처리 파이프라인을 구현해본다
    b. 데이터셋
        i. https://www.kaggle.com/competitions/titanic
        ii. train.csv / test.csv
    c. 사용도구 : pandas, sklearn
2. Exploratory Data Analysis
    a. 머신러닝을 위한 전처리를 수행하기 위해 EDA를 진행하였습니다.
    b. 데이터 기본 현황
        i. 총 891개의 학습 데이터
        ii. 결측치
            1. Age : 177개(19.9%)
            2. Cabin : 687개(77.1%)
            3. Embarked : 2개(0.2%)
        iii. 데이터 타입
            1. 수치형 : Age, Fare, SibSp, Parch
            2. 범주형 : Sex, Pclass, Embarked
            3. 불필요 : PassengerId, Name, Ticket
    c. 데이터별 분석
        i. Cabin : 객실번호는 결측치가 너무 많아서 유의미한 정보를 추출하기 어려워 보입니다
        ii. Age : 분포가 특정 값에 치우치지 않고 비교적 고른 분포이며 결측치가 20%라서 제거하면 손실이 클것으로 보입니다
        iii. Embarked : 탑승구는 결측치 2건을 제외하고 모두 있습니다
        iv. Age / Fare : 나이와(0~80), 요금(0~512)는 스케일이 달라서 스케일링 작업이 필요할것으로 보입니다 => 로지스틱 회귀가 스케일에 민감하므로
        v. 범주형 데이터는 수치형 변환이 필요합니다

3. 전처리 파이프라인 구성
    a. 수치형 - 결측치O : SimpleImputer, StandardScaler
    b. 수치형 - 결측치X : StandardScaler
    c. 범주형 - 결측치O : SimpleImputer, OneHotEncoder
    d. 범주형 - 결측치X : OneHotEncoder
4. 전처리 파이프라인 구현 (Python) 및 전처리 실행 결과는 별도의 파일 참조
5. 결론
    a. 타이타닉 데이터셋의 결측치, 범주형, 스케일 문제를 해결하기 위한 전처리 파이프라인을 구축하였습니다
    b. 해당 파이프라인은 모델 학습과 평가 과정관련 문서 작업에 재사용할 예정입니다