|  |  |
| --- | --- |
| 교육 제목 | **판다스 데이터 사전처리, 데이터프레임 응용** |
| 교육 일시 | 2021-10-01 |
| 교육 장소 | 집 |
| **교육 내용** | |
| 오전 | 누락데이터 처리  : 유효 데이터 값 존재 하지 않으면 NaN 표시해야 알고리즘 왜곡 현상방지  사용방법     1. 원하는 데이터 출력      1. 누락된 데이터 확인     2\_1. isnull().sum(axis=0) 을 활용해 누락 데이터 개수 구하기     1. 누락된 데이터 확인      1. Nan값 데이터 삭제     4-1. 누락된 데이터 지우기     1. 중복된 데이터 지우기   단계별로 지우는 이유 : 한번에 다지우면 반드시 실수, 없어진 데이터 값으로  분석 값에 영향을 크게 미침  데이터표준화  : 같은 데이터셋 안에서 서로 다른 측정단위를 사용한다면, 전체데이터 일관성 측면애서 문제 발생  -> 따라서 측정 단위 동일하게 맞춰야 됨      같은 단위로 변환 |
| 오후 | 범주형 : 알고리즘 따라 연속 데이터를 그래도 사용하기 보다는 일정 구간 나누어 분석하는 것이 효율적  구간분할 : 이때 구간을 분할하는 것을 의미  더미변수 : 연속 데이터를 열의 범주형으로 변환 하였 어도 컴퓨터는 인식하지 못 할 경우가 대부분 따라서 인식가능한 0 1 값으로 표현    데이터 구간 분할   * 4부분으로 나누어서 출력     더미 변수 사용    원 핫인 코딩 : 하나는 True, 다른 하나는 False 값 만들어주는 기법  사용 이유 : scikit-learn에서 제공하는 머신러닝 알고리즘은 문자열 값을 입력 값으로 허락하지 않기 때문에 모든 문자열 값들을 숫자형으로 인코딩하는 전처리 작업 후에 머신러닝 모델에 학습을 시켜야 한다.  데이터프레임의 다양한 응용  함수 매핑 : 시리즈, 데이터프레임의 개별 원소들 특정 함수에 일대일 대응시키는 과정을 뜻함  시리즈 객체에 apply() 메소드를 적용하여 모든 원소를 하나씩 입력하고, 함수의 리턴값을 돌려받음    Apply() 메서드를 이용한 함수 매핑    Apply() 메소드를 이용하여 age 값과 fare 값을 10으로 매핑  이후 적용된 값 간추려서 출력    이후 대응대는 add\_10값을 적용  총 정리 문제   1. titanic 데이터를 load, 2. age와 fare컬럼나 추출, 3. age에서 평균 나이를 차감한 나이를 age\_avg컬럼으로 추가, 정수로 4. 나이의 구간을 4단계로 나누어서 [ 청소년, 장년, 노년]으로 컬럼추가 5. one\_hot\_encoding       출력 값    Plot()으로 함수 그래프화 출력 |