|  |  |
| --- | --- |
| 교육 제목 | **Pandas 데이터 사전 처리** |
| 교육 일시 | 2021.10.01 |
| 교육 장소 | YGL-C6 |
| **교육 내용** | |

1. 누락 데이터처리

머신러닝 등 데이터 분석의 정확도는 데이터의 품질에 좌우 된다. 품질을 높이기 위해 누락데이터 중복 데이터 등 오류를 수정하고 분석 목적에 맞게 변형하는 과정이 필요

* 1. 누락 데이터 확인 (seaborn titanic 자료 사용)

data.info() 메소드를 통해 누락 데이터 확인

테이블이(가) 표시된 사진

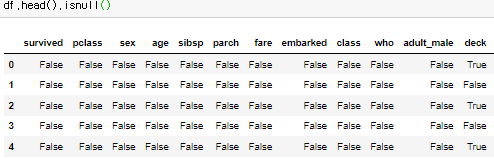
자동 생성된 설명

age, deck,embark\_town에서 누락데이터 확인

텍스트이(가) 표시된 사진

자동 생성된 설명

head().isnull() 메소드 사용으로 누락 값 확인



True인 행은 누락 된 값

notnull() 메소드를 사용하면 누락 되지않은 데이터

* 1. 누락 데이터 제거

data.dropna()메소드 사용 thresh = 500은 누락 데이터가 500개 이상인 데이터만

테이블이(가) 표시된 사진

자동 생성된 설명

* 1. 누락 데이터 치환

데이터를 무작정 삭제해 버리면 어렵게 수집한 데이터를 활용하지 못함

누락데이터를 바꿔서 대체할 값으로는 데이터의 분포와 특성을 잘 나타낼 수 있는 평균값, 최빈값 등을 활용

fillna()메소드 활용

텍스트이(가) 표시된 사진

자동 생성된 설명 age컬럼의 평균값으로 대체

* 1. 중복 데이터 처리

하나의 데이터셋에서 동일한 관측값이 2개 이상 중복되는 경우 중복 데이터를 찾아서 삭제해야함 중복으로 존재할 경우 분석 결과 왜곡 가능성 농후

텍스트이(가) 표시된 사진

자동 생성된 설명 c2와 c3의 데이터가 중복됨

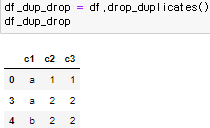
duplicated()메소드 활용 : 이전 행과의 레코드가 중복되는지 확인

테이블이(가) 표시된 사진

자동 생성된 설명

* 1. 중복 데이터 처리

drop\_duplicates() 메소드 사용 : 중복되는 행을 제거하고 고유한 관측값만 남김

 duplicated()에서 True인 값 제거

drop\_duplicates()메소드의 subset옵션에서 ‘열 이름의 리스트’를 전달할 수 있다. 데이터의 중복 여부를 판별할 때 subset옵션에 해당하는 열을 기준으로 판단한다.

텍스트이(가) 표시된 사진

자동 생성된 설명c2,c3열에서는 1,2만 존재

1. 데이터 표준화
   1. 단위 환산

같은 데이터셋 안에서 서로 다른 측정 단위를 사용한다면 데이터의 일관성 측면에서 문제가 발생한다. 따라서 측정 단위를 맞출 필요가있다

auto-mpg데이터셋의 마일을 킬로미터로 변경

테이블이(가) 표시된 사진

자동 생성된 설명

* 1. 자료형 변환

숫자가 문자열로 저장된 경우에는 정수형이나 실수형을 변환해야 한다. 먼저 dtyptes를 사용해 자료형을 확인

테이블이(가) 표시된 사진

자동 생성된 설명마력(horsepower)이 문자열로 나타내져 있기 때문에 변환해야함

텍스트, 영수증이(가) 표시된 사진

자동 생성된 설명

unique()를 통해 데이터 확인

‘?’로 입력 된 값을 NaN값으로 변환 한다. 이 후 dropna(axis=0) 메소드로 NaN값이 들어 있는 모든 행을 삭제한다. 이 후 astype()메소드를 이용해 실수형으로 변경

텍스트이(가) 표시된 사진

자동 생성된 설명

같은 방법으로 정수형 데이터를 문자형 데이터로 바꾸어 줄 수 있다.

텍스트이(가) 표시된 사진

자동 생성된 설명

* 1. 범주형(카테고리) 데이터 처리

연속적인 데이터를 그대로 사용하기 보다는 일정한 구간으로 나눠서 분석하는 것이 효율 적이다. 연속 변수를 일정한 구간으로 나누고 각 구간을 범주형 이산 변수로 변환하는 과정을 구간 분할이라고 한다.

numpy의 histogram()함수를 사용하면 나누려는 구간(bin)의 개수를 입력했을때 각 구간에 속하는 값의 개수와 경계값을 반환 한다.

텍스트이(가) 표시된 사진

자동 생성된 설명

pandas의 cut() 함수의 옵션을 설정한다. 앞에서 구한 경계값 리스트를 bin옵션에 할당하고 각 구간의 이름을 label 옵션에 할당한다.

\텍스트이(가) 표시된 사진

자동 생성된 설명

위 처럼 범주형 데이터로 변환 하여도 회귀 분석이나 머신러닝 알고리즘에 바로 사용할 수가 없다. 이를 컴퓨터가 인식 할 수 있도록 더미 변수를 사용해 변환해준다.

더미변수란 0또는 1로 표현되고 어떤 특성이 있는지 없는지만 판단한다.

get\_dummies() 메소드 사용

sklearn 라이브러리를 활용해서 원핫인코딩을 편하게 할 수 있다. 앞선 데이터의 hp\_bin 컬럼에 들어있는 범주형 데이터를 0,1원소를 가지는 원핫 벡터로 변환하고 그 결과는 희소행렬로 정리된다.

텍스트이(가) 표시된 사진

자동 생성된 설명

* 1. 정규화

각 데이터의 상대적 크기 차이 때문에 머신러닝 분석 결과가 달라질 수 있다. 따라서 숫자 데이터의 상대적인 크기 차이를 제거할 필요가 있다. 각 데이터 값을 동일한 크기 또는 기준으로 나타내는 것을 정규화(Normalization)이라고 한다.

텍스트이(가) 표시된 사진

자동 생성된 설명

최댓값 정규화 진행 후 최대값은 1 최솟값은 0.2로 변환되었다. 여기서 최댓값이 아닌 최솟값과 최댓값을 모두 이용하면 0~1 범위로 변경 가능하다. (data – data.min)/(data.max -data.min)

테이블이(가) 표시된 사진

자동 생성된 설명

* 1. 시계열 데이터

Pandas는 시계열 데이터를 다루는 여러 가지 유용한 기능을 제공한다.