Iris.py

문제 개요

- Iris 꽃 관련 데이터를 수집하여 학습용/테스트용 데이터로 분할한 뒤, 이를 분석하려고 함
- Iris 꽃 관측 데이터는 iris.csv 에 저장되어 있으며, 꽃 종류에 대한 데이터는 iris_metadata.csv 에 저장되어 있으며 내용은 다음과 같음

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
5	5.4	3.9	1.7	0.4	setosa
6	4.6	3.4	1.4	0.3	setosa
7	5.0	3.4	1.5	0.2	setosa
8	4.4	2.9	1.4	0.2	setosa
9	4.9	3.1	1.5	0.1	setosa
10	5.4	3.7	1.5	0.2	setosa

(iris.csv 내용의 일부)

	name	toxic
0	setosa	0
1	versicolor	2
2	virginica	5

(iris_metadata.csv 내용

(0) 뼈대 코드

- iris.py 모듈 내에 네 가지 함수를 정의할 것
- read_dataset()함수는 iris.csv 와 iris_metadata.csv 를 pandas dataframe 객체로 읽는 함수이며, 수업시간에 다루지 않았기 때문에 코드를 제공함. 수정하지 말 것
- main() 함수는 complete here 주석 전에 선언된 코드는 수정하지 말 것

(1) merge_dfs

- iris dataframe 과 metadata dataframe 을 merge 함수를 통해 병합하려고 함
- 파라미터
  - iris: 꽃 관측 데이터가 저장되어 있는 pandas dataframe
  - metadata: 꽃 종류에 대한 데이터가 저장되어 있는 pandas dataframe
- iris dataframe 의 species 열과 metadata dataframe 의 name 열을 기준으로 병합을 수행
- 반환값
  - Iris 의 species, metadata 의 name 을 기준으로 merge 된 pandas dataframe

(2) split_dfs

- 파라미터로 받는 데이터프레임을 학습/테스트용 데이터로 나눔
- 파라미터
  - df: pandas dataframe
  - ratio: 분할 비율을 나타내는 float 값 (0-1 범위)
- ratio 를 바탕으로 분할 기준으로 삼을 index 값을 계산
  - int 와 float 간의 연산의 결과값이 float 이므로 이를 int 로 변환할 필요가 있음
  - 이 때 Python 내장함수인 round 함수를 사용하거나, int 로 casting
- 위에서 구한 기준 index 값을 기준으로 파라미터로 받은 df 를 학습용/테스트용 data-frame 으로 분할
  - 이 때 기준 index 값보다 작은 index 는 학습용 dataframe 으로, index 값보다 같거나 큰 index 는 테스트용 dataframe 으로 분할되도록 해야 함
- 반환값
  - (학습용 pandas dataframe, 테스트용 pandas dataframe)으로 이루어진 tuple 객체

(3) truncate_non_toxics

- 파라미터
  - ■ df: pandas dataframe
- 파라미터로 받는 df 를 읽은 뒤 toxic 열의 값이 0 보다 큰 데이터만 추출하는 slic-ing 을 진행
- 반환값
  - ■ Toxic 열의 값이 0 이상인 데이터만 존재하는 pandas dataframe

(4) main

- 위에서 정의한 함수를 이용해 다음 내용을 코드로 구현
- merge_dfs 함수를 이용해 iris 와 metadata 가 병합된 dataframe 을 생성
- 병합된 dataframe 에 대해 split_df 함수를 이용하여 학습용 dataframe 와 테스트용 dataframe 을 구함
- 학습용 dataframe 와 테스트용 dataframe 각각에 대해 truncate_non_toxics 함수를 적용하여 toxic 이 0 보다 큰 데이터만 남도록 슬라이싱 진행
- pandas 내의 describe()함수를 사용하여 학습용 dataframe 와 테스트용 dataframe 에 대한 descriptive statistics 출력
- 출력 결과

	sepal_length	sepal_width	...	petal_width	toxic
count	55.000000	55.000000	...	55.000000	55.000000
mean	5.978182	2.789091	...	1.396364	2.272727
std	0.525530	0.310718	...	0.302437	0.870388
min	4.900000	2.000000	...	1.000000	2.000000
25%	5.600000	2.600000	...	1.200000	2.000000
50%	6.000000	2.800000	...	1.300000	2.000000
75%	6.300000	3.000000	...	1.500000	2.000000
max	7.100000	3.400000	...	2.500000	5.000000

(학습용 dataframe에 대한 describe()함수 출력 결과)

	sepal_length	sepal_width	petal_length	petal_width	toxic
count	45.000000	45.000000	45.000000	45.000000	45.0
mean	6.608889	2.973333	5.537778	2.017778	5.0
std	0.652555	0.333984	0.570601	0.276577	0.0
min	4.900000	2.200000	4.500000	1.400000	5.0
25%	6.200000	2.800000	5.100000	1.800000	5.0
50%	6.500000	3.000000	5.500000	2.000000	5.0
75%	6.900000	3.200000	5.800000	2.300000	5.0
max	7.900000	3.800000	6.900000	2.500000	5.0

(테스트용 dataframe에 대한 describe()함수 출력 결과)