

빅데이터 분석 기사 시험 분석

데사라면

빅데이터 분석기사 시험 분석

주요항목	세부항목	세세항목
데이터 수집 작업	데이터 수집하기	정형, 반정형, 비정형 등 다양한 형태의 데이터를 읽을 수 있다.
		필요시 공개 데이터를 수집할 수 있다.
데이터 전처리 작업	데이터 정제하기	정제가 필요한 결측값, 이상값 등이 무엇인지 파악할 수 있다.
		결측값과 이상값에 대한 처리 기준 을 정하고 제거 또는 임의의 값으로 대체 할 수 있다.
	데이터 변환하기	데이터의 유형을 원하는 형태 로 변환할 수 있다.
		데이터의 범위를 표준화 또는 정규화 를 통해 일치시킬 수 있다.
		기존 변수를 이용하여 의미 있는 새로운 변수를 생성 하거나 변수를 선택 할 수 있다.

빅데이터 분석기사 시험 분석

주요항목	세부항목	세세항목
데이터 모형 구축 작업	분석모형 선택하기	다양한 분석 모형 을 이해 할 수 있다.
		주어진 데이터와 분석 목적 에 맞는 분석모형을 선택할 수 있다.
		선정모형 에 필요한 가정 등을 이해할 수 있다.
	분석모형 구축하기	모형 구축에 부합하는 변수 를 지정 할 수 있다.
		모형 구축에 적합한 형태로 데이터 를 조작 할 수 있다.
		모형 구축에 적절한 매개변수 를 지정 할 수 있다.
데이터 모형 평가 작업	구축된 모형 평가하기	최종 모형을 선정하기 위해 필요한 모형 평가 지표 들을 잘 사용할 수 있다.
		선택한 평가지표를 이용하여 구축된 여러 모형을 비교 하고 선택 할 수 있다.
		성능 향상을 위해 구축된 여러 모형 을 적절하게 결합 할 수 있다.
	분석결과 활용하기	최종모형 또는 분석결과를 해석 할 수 있다.
		최종모형 또는 분석결과를 저장 할 수 있다.

빅데이터 분석기사 기출 분석

- 문제 유형
- 주관식 단답형(3점) 10문제
 - 단순 작업형(10점) 3문제

[붙임 : 유형별 예시문제]

○ 단답형

단답형 예시문제

여러 명의 사용자들이 컴퓨터에 저장된 많은 자료들을 쉽고 빠르게 조회, 추가, 수정, 삭제할 수 있도록 해주는 소프트웨어는 무엇인가?

○ 작업형 제1유형 : 데이터 처리 영역

작업형 제1유형 예시문제

mtcars 데이터셋(mtcars.csv)의 qsec 컬럼을 최소최대 척도(Min-Max Scale)로 변환한 후 0.5보다 큰 값을 가지는 레코드 수를 구하시오.

작업형 제 2유형 예시 문제 (40점)

작업형 제2유형 예시문제

아래는 백화점 고객의 1년 간 구매 데이터이다.

아 래

(가) 제공 데이터 목록

- ① **y_train.csv** : 고객의 성별 데이터 (학습용), CSV 형식의 파일
- ② **X_train.csv, X_test.csv** : 고객의 상품구매 속성 (학습용 및 평가용), CSV 형식의 파일

(나) 데이터 형식 및 내용

- ① **y_train.csv** (3,500명 데이터)

	cust_id	gender
0	0	0
1	1	0
2	2	1
3	3	1
4	4	0
5	5	0
6	6	0
7	7	0
8	8	0
9	9	1

- * custid: 고객 ID
- * gender: 고객의 성별 (0: 여자, 1: 남자)

- ② **X_train.csv** (3,500명 데이터), **X_test.csv** (2,482명 데이터)

	cust_id	총구매액	최대구매액	환불금액	주구매상품	주구매지점	내정일수	내정당구매건수	주알방문세율	구매주기
0	0	68282840	11264000	6860000.0	기타	강남점	19	3.894737	0.527027	17
1	1	2136000	2136000	300000.0	스포츠	잠실점	2	1.500000	0.000000	1
2	2	3197000	1636000	NaN	기타	관악점	2	2.000000	0.000000	1
3	3	16077620	4935000	NaN	기타	광주점	10	2.444444	0.310182	16
4	4	29060000	24000000	NaN	기타	본점	2	1.500000	0.000000	85
5	5	11379000	9652000	462000.0	디자이너	일산점	3	1.666667	0.200000	42
6	6	10066000	7612000	4682000.0	소프트웨어	강남점	6	2.400000	0.333333	42
7	7	514570000	27104000	26624000.0	영품	본점	60	2.634621	0.222862	5
8	8	683243360	173086000	NaN	기타	본점	18	5.944444	0.411215	15
9	9	26640860	13728000	NaN	농산물	대전점	1	12.000000	0.000000	0

고객 3,500명에 대한 학습용 데이터(y_train.csv, X_train.csv)를 이용하여 성별예측 모델을 만든 후, 이를 평가용 데이터(X_test.csv)에 적용하여 얻은 2,482명 고객의 성별 예측값(남자일 확률)을 다음과 같은 형식의 CSV 파일로 생성하시오.(제출한 모델의 성능은 ROC-AUC 평가지표에 따라 채점)

<제출형식>

custid,gender
3500,0.267
3501,0.578
3502,0.885
...

<유의사항>

성능이 우수한 예측모형을 구축하기 위해서는 적절한 데이터 전처리, Feature Engineering, 분류 알고리즘 사용, 초매개변수 최적화, 모형 앙상블 등이 수행되어야 한다.

시험 환경

문제 1

문제 1

solution.py

```
1 import collections as col
2 def solution(participant, completion):
3     answer = list((col.Counter(participant) - col.Counter(completion)).keys())
4     print("빅데이터 분석기사는 이와 같은 환경으로 print로 데이터를 확인해야 합니다.")
5     return answer
```

실행 결과

출력 > 빅데이터 분석기사는 이와 같은 환경으로 print로 데이터를 확인해야 합니다.

실행결과 창에 출력 결과가 나옴

코드 실행

제출 후 채점하기

시험 환경

작업형 제2유형
예시문제

아래는 백화점 고객의 1년 간 구매 데이터이다.

아 래

(가) 제공 데이터 목록

- ① `y_train.csv` : 고객의 성별 데이터 (학습용), CSV 형식의 파일
- ② `X_train.csv`, `X_test.csv` : 고객의 상품구매 속성 (학습용 및 평가용), CSV 형식의 파일

(나) 데이터 형식 및 내용

① `y_train.csv` (3,500명 데이터)

	cust_id	gender
0	0	0
1	1	0
2	2	1
3	3	1
4	4	0
5	5	0
6	6	0
7	7	0
8	8	0
9	9	1

- * `custid`: 고객 ID
- * `gender`: 고객의 성별 (0: 여자, 1: 남자)

② `X_train.csv` (3,500명 데이터), `X_test.csv` (2,482명 데이터)

	cust_id	종구매액	최대구매액	환불금액	주구매상품	주구매지점	내정일수	내정당구매건수	주알방문비용	구매주기
0	0	68282840	11264000	6860000.0	기타	강남점	19	3.894737	0.527027	17
1	1	2136000	2136000	305000.0	스포츠	잠실점	2	1.500000	0.000000	1
2	2	3197000	1636000	NaN	기타	관악점	2	2.000000	0.000000	1
3	3	16077620	4935000	NaN	기타	광주점	10	2.444444	0.310102	16
4	4	29060000	24000000	NaN	기타	본 점	2	1.500000	0.000000	85
5	5	11379000	9652000	462000.0	디자이너	일산점	3	1.666667	0.200000	42
6	6	10066000	7612000	4682000.0	소프트웨어	강남점	6	2.400000	0.333333	42
7	7	514570000	27104000	26624000.0	영품	본 점	60	2.634621	0.222862	5
8	8	663243360	173086000	NaN	기타	본 점	18	5.944444	0.411215	15
9	9	26640660	13728000	NaN	농산물	대전점	1	12.000000	0.000000	0

고객 3,500명에 대한 학습용 데이터(`y_train.csv`, `X_train.csv`)를 이용하여 성별예측 모델을 만든 후, 이를 평가용 데이터(`X_test.csv`)에 적용하여 얻은 2,482명 고객의 성별 예측값(남자일 확률)을 다음과 같은 형식의 CSV 파일로 생성하시오.(제출한 모델의 성능은 ROC-AUC 평가지표에 따라 채점)

<제출형식>

```
custid,gender
3500,0.267
3501,0.578
3502,0.885
...
```

<유의사항>

성능이 우수한 예측모형을 구축하기 위해서는 적절한 데이터 전처리, Feature Engineering, 분류 알고리즘 사용, 초매개변수 최적화, 모형 앙상블 등이 수반되어야 한다.

solution.py

```
1 import collections as col
2 def solution(participant, completion):
3     answer = list((col.Counter(participant) - col.Counter(completion)).keys())
4     print("빅데이터 분석기사는 이와 같은 환경으로 print로 데이터를 확인해야 합니다.")
5     return answer
```

실행 결과

출력 > 빅데이터 분석기사는 이와 같은 환경으로 print로 데이터를 확인해야 합니다.