KDT 인공지능 기초 (실기)

2023.01

멀티캠퍼스

문제 유형1

* 다음은 음식 조리 시간 관련 데이터셋이다. 이 데이터셋을 활용하여 음식 조리 시간에 대한 예측을 수행하시오.

파일명: indian_food2.csv (구분자: 쉼표(,), 255 행, 10 컬럼, UTF-8 Encoding, 결측치: Empty string(nan, 빈칸(' '), 문자열(''))

변수명	설명	데이터 타입
id	음식 ID	Double
name	음식 이름	String
ingredients	재료 목록(", "로 구분, 예: 고추장, 고추 가루, 된장)	String
diet	고기 포함 여부(vegetarian/non vegetarian)	String
prep_time	재료 준비 시간	Double
cook_time	음식 조리 시간(Target 변수)	Double
flavor_profile	음식 맛 구분("-1", "spicy", "sweet", "bitter", "sour")	String
course	음식 종류(dessert/main course/snack/starter)	String
state	지역(Andhra Pradesh ~ West Bengal)	String
region	음식 지역("-1", "Central", "East", "North", "North East", "South", "West", "nan")	String

*** 다음의 전처리를 먼저 수행한 후 분석 진행하시오.**

단계 1-1. 지역(state) 컬럼을 제거하시오.

단계 1-2. 조리 시간(cook_time)과 재료 준비 시간(prep_time)이 양수(>0)인 데이터만 추출하시오.

단계 1-3. region의 값이 "-1"인 경우와 빈문자(empty string)는 제거한 후 'North East' 지역을 'North'로 변경하시오(결측치는 제거하시오).

단계 1-4. 음식 맛 구분(flavor_profile)을 다음과 같이 2가지로 정리한 변수(변수명: ed_flavor_profile)를 추가하시오.

IF flavor_profile= 'sweet' then 'sweet' else 'not sweet'

단계 1-5. 음식 종류(course)를 2가지로 정리한 변수(변수명: ed_course)를 추가하시오.

IF course = 'main course ' then 'main course' else 'not main course'

단계 1-6. 재료 목록(ingredients)를 이용하여 재료 수 변수(변수명: ingredients_no)를 추가하시오.

예: 고추장, 마늘 → ingredients_no=2

*** 상기 전처리를 완료한 데이터 프레임(데이터 프레임명: basetable1, 212 Rows)으로 다음 분석(문제 1~3) 수행**

1. (basetable1을 이용) 2범주로 정리된 음식 맛(ed_flavor_profile)이 5개로 정리된 음식 지역(region) 간에 통계적으로 유의미한 차이가 있는지 적절한 검정을 수행하고, 검정 결과 중 검정 통계량을 기술하시오.

- 검정 통계량은 소수점 넷째 자리 이하는 버리고 소수점 셋째 자리까지 기술. (답안 예시) 0.123

2. (basetable1을 이용) 5개 지역(region)에 따라 재료 준비 시간(prep_time)의 평균이 통계적으로 유의미한 차이가 있는지 적절한 검정을 수행하고, 검정 결과 중 유의 확률(P-Value) 값을 기술하시오.

- 주어진 데이터는 정규성과 등분산성을 가정하며, 유의 확률(P-Value) 값은 소수점 넷째 자리 이하는 버리고 소수점 셋째 자리까지 기술. (답안 예시) 0.123

3. (basetable1을 이용) 음식 조리 시간(cook_time)을 예측하는 모델을 생성하고자 한다. 이를 위하여 basetable1에 다음의 전처리를 추가로 수행하시오.

단계 1-7. 다음 변수들을 추가하시오(총 2개).

- ed_course으로부터 생성한 가변수(1개), ed_flavor_profile로부터 생성한 가변수(1개)
- 가변수화 Hint

Brightics Studio	One Hot Encoder 사용, Drop Last=True

단계 1-8. 음식 조리 시간(cook_time) 예측을 위한 선형 회귀 모형을 생성하고자 한다. 다음 가이드에 따라 Train Data로 학습시킨 모형을 Test Data에 적용하여 음식 조리 시간(cook_time)을 예측하시오.

- Train Data: id가 3의 배수가 아닌 데이터, Test Data: id가 3의 배수인 데이터
- 독립변수(3개): **단계 1-7**의 가변수들(2개), prep_time
- 선형 회귀 가이드

Brightics Studio	Default 사용

생성한 회귀 분석 모델을 Test Data에 적용하여 음식 조리 시간(cook_time)를 예측한 후 예측값들에 대한 평균제곱근오차(Root Mean Square Error, RMSE)를 기술하시오.

- RMSE는 소수점 둘째 자리 이하는 버리고 소수점 첫째 자리까지 기술. (답안 예시) 1.2

문제 유형 2

보험사에서 고객(총 29,132명) 신상 정보와 21개의 보험 상품 구매 정보를 이용하여 보험 상품 추천 시스템을 개발하고자 하며, 제공된 데이터셋 정보는 다음과 같다.

Train2.csv (구분자: ","(Comma), 29,132 Rows, 26 Columns, UTF-8 Encoding)

변수명	설명	데이터 타입
ID	고객 ID	String
join_date	가입 일자, DAY/MONTH/YEAR 형식 예: 2021년 4월 8일 가입->8/4/2021	String
sex	성별(M/F)	String
birth_year	출생 연도	Double
branch_code	모집 사무소	String
P1~ P21	보험 상품 1~21에의 가입 여부(1: 가입, 0: 미가입)	Double

*** 다음의 전처리를 먼저 수행한 후 분석 진행하시오.**

단계 1-1. 가입 일자(join_date) 컬럼의 연도 정보를 이용하여 가입 연령(변수명: reg_age) 컬럼을 추가하시오(join_date)를 기준으로 결측치를 제거한 후 변수 생성하시오).

reg_age = 가입 연도 - 출생 연도 + 1

단계 1-2. 가입한 보험 상품의 수가 4개 이상(>=)인 고객은 VIP로 분류하는 VIP 컬럼(변수명: VIP)을 추가하시오.

IF 가입한 보험 상품의 수 >= 4 then 'VIP' else 'Not VIP' (VIP는 총 1,231명)

(가입한 보험 상품 수는 P1~P21 상품 중 가입한 상품의 수를 집계한 후 사용)

*** 상기 전처리를 완료한 데이터 프레임(데이터 프레임명: basetable1, 29132(결측치 포함)/29130(결측치 제거) Rows)으로 다음 분석(문제 1~3) 수행**

1. (basetable1을 이용) 모집 사무소(branch_code) 별 가입 연령의 평균을 집계하고, 이 값이 가장 큰 모집 사무소를 기술하시오. (답안 예시) ABC

2. (basetable1을 이용) 고객의 VIP 여부가 성별(sex)과 서로 독립이라 할 수 있는지 적절한 검정을 수행하고, 검정 결과 p-value를 기술하시오.

- p-value는 소수점 셋째 자리 이하는 버리고 소수점 둘째 자리까지 기술 (답안 예시) 0.1234

3. (basetable1 을 이용) 고객들이 어떤 보험 상품들을 같이 가입했는지 연관성 분석을 다음 조건에 따라 수행하고자 한다.

- 21 개의 보험 상품(P1~P21 활용)을 분석 대상으로 하며, 1 개의 보험상품에만 가입한 고객은 분석에서 제외
- Hint

cust_id	P1	P2	P3
4WK	1	1	0
CP5	0	1	1

→

(가입한 상품)
{"P1", "P2"}
{"P2", "P3"}

- (Association Rule 모델 가이드) Min Support: 0.01, Min Confidence: 0.01, 그 외: Default

연관성 분석 결과 중, 선행(antecedent)과 후행(consequent)이 단일 아이템으로 구성된 rule 만 추출하시오. 이 결과를 활용하여, 보험상품 'P15'에 가입한 고객에게 추천하기 가장 적합한 보험 상품을 Lift 기준(내림차순)으로 선정하고, 그 때의 Lift 를 기술하시오.

- Lift 는 소수점 넷째 자리 이하는 버리고 셋째 자리까지만 기술하시오.

(답안예시) 1.234

문제 유형 3

다음은 영국에 위치한 온라인 쇼핑몰 U 사의 2010 년~2011 년의 판매 데이터이며, 판매 영수증 정보는 다음과 같다.

Online2.txt (구분자: tab('₩t'), 541,909 Rows, 8 Columns, UTF-8 Encoding)

변수명	설명	데이터 타입
InvoiceNo	영수증 번호	Double
StockCode	물품 번호	String
Description	제품 설명	String
Quantity	구매 수량	Double
InvoiceDate	구매 시간, 2010 년의 경우 12 월만 존재	String
UnitPrice	물품 단가(단위: 파운드)	Double
CustomerID	구매자 ID	Double
Country	배송지 국가	String

*** 다음의 전처리를 먼저 수행한 후 분석 진행하시오.**

단계 1-1. 영수증 번호(InvoiceNo)가 수치형이 아닌 데이터 Row는 제거하시오.

단계 1-2. 물품 단가(UnitPrice) 또는 구매 수량(Quantity)이 0 이하(<=0)인 Row는 제거하시오.

단계 1-3. 구매액(변수명: TotOrder) 변수를 추가하시오. TotOrder = Quantity*UnitPrice

*** 상기 전처리를 완료한 데이터 프레임(데이터 프레임명: basetable1, 530,103 Rows)으로 다음 분석(문제 1~3) 수행**

1. (basetable1 을 이용) '영수증 번호(InvoiceNo)별 총 구매액(변수명: InvTotOrder)'의 평균이 국가별로 차이가 있는지 다음의 조건들을 참고하여 검정을 수행하고 이 때의 검정 통계량(F-value)을 기술하시오. 이때 처리 순서는 다음 제시된 조건 순서대로 진행하시오.

[조건]

- 하나의 InvoiceNo 는 하나의 Country 에 대응
- 영수증 번호별 총 구매액(InvTotOrder)은 각 InvoiceNo 에 대한 TotOrder 의 합으로 정의 (19959 개)
- 국가별 구매 건수(InvoiceNo 수)가 50 이상(>=50) 이고 400 이하(<=400)인 국가만을 대상으로 검정 수행, 자유도(df)=14
- 주어진 데이터는 등분산 조건을 만족한다고 가정
- 검정 통계량은 소수점 넷째 자리 이하는 버리고 셋째 자리까지만 기술 (답안 예시) 1.234

2. (basetable1 을 이용) '주문날짜(InvoiceDate)'와 '영수증 번호(InvoiceNo)별' 총 구매액'(InvTotOrder, 1 번 참고) 테이블을 구성한 후 요일(Weekday)변수를 생성하시오. (InvoiceNo, InvTotOrder Weekday 컬럼을 가지는 데이터

프레임, 19959 rows). 이 때, 요일별 총 구매액이 상위 20%에 속하는 데이터만을 대상으로 총 구매액의 요일별 평균을 구하시오. 평균이 가장 낮은 요일과 평균 값을 기술하시오

- 평균 값은 정수 부분만 기술하시오.
- 영수증 번호별 총 구매액(InvTotOrder)은 각 InvoiceNo 에 대한 TotOrder 의 합으로 정의
- quantile(0.8) 함수를 사용하되 함수의 parameter 는 기본값 사용
(답안예시) 일요일, 1

3. (basetable1 을 이용) 구매자(CustomerID)가(와) 주문 날짜(InvoiceDate) 기준으로 동시에 구매한 물품(StockCode)들에 대한 연관성 분석을 다음 조건과 같이 수행하고자 합니다
- 분석 대상: 물품 단가(UnitPrice)의 평균이 4 이상(>=4)이고 구매(판매) 수량(Quantity)의 합이 600 이상(>=600)인 물품
- 구매자(CustomerID)와 주문 날짜(InvoiceDate)를 기준으로 물품 목록은 동일목록 존재 시 한 품목만 포함
- 구매자(CustomerID)와 주문 날짜(InvoiceDate)를 기준으로 한 종류의 물품만을 구매한 구매자는 제외
- 연관성 분석 시 구매자(CustomerID)와 주문 날짜(InvoiceDate)를 기준으로 물품목록을 추출한 후 연관성 분석(basket, transaction) 단위로 함
Note: 물품(StockCode)의 UnitPrice 는 구매 시점마다 상이할 수 있습니다.
- 구매자 (CustomerID)를 확인할 수 없는 경우(결측치)는 분석에서 제외합니다.
- (Association Rule 모델 가이드) Min Support: 0.01, Min Confidence: 0.01, 그 외: Default

연관성 분석 결과 중, 선행(antecedent)과 후행(consequent)이 단일 아이템으로 구성된 rule 만 추출하시오. 이 결과를 활용하여, 물품 '23012'을 주문한 고객에게 추천하기 적합한 품목을 Lift 기준(내림차순)으로 선정하고, 그 때의 Lift 를 기술하시오.
- Lift 는 소수점 넷째 자리 이하는 버리고 셋째 자리까지만 기술하시오.
(답안예시) 1.234

문제 유형4

실시간 중계 Tweet이 방문자들의 반응지표들과 관계가 있는지에 대한 분석을 수행하고자 한다. 제공된 데이터셋 정보는 다음과 같다.(2017년 9월 29일부터 10 월 26일까지 개인 계정에서 보낸 313개의 트윗 정보 수집)

tweets2.csv (구분자: ","(comma), 313 Rows, 13 Columns, UTF-8 Encoding):

변수명	설명	데이터 타입
No	Index	Double
atConference	실시간 중계 여부(TRUE/FLASE)	String
day	트윗 일자(YYYY-MM-DD)	String
impressions	트윗 조회 수	Double
retweets	리-트윗 횟수	Double
likes	"좋아요" 클릭수	Double
userprofileclicks	프로필 조회 횟수	Double
urlClicks	URL 클릭 횟수	Double
hashtagClicks	트윗에서 해시 태그를 클릭한 횟수	Double
detailExpands	트윗 확장 횟수	Double
follows	Follower 수	Double
mediaViews	트윗에 포함된 미디어를 본 횟수	Double
mediaEngagements	트윗에 포함된 미디어를 클릭한 횟수	Double

*** 다음의 전처리를 먼저 수행한 후 진행하시오.**

단계 1-1. atConference, day, impressions 를 제외한 나머지 9 개 변수의 합 변수(변수명: totalSum)를 추가하시오.

*** 상기 전처리를 완료한 데이터 프레임(데이터 프레임명: basetable1)으로 다음 분석(문제 1~3)을 수행하시오.**

1. (basetable1 을 이용) 리-트윗 횟수(retweets)가 0 이 아닌 트윗을 대상으로 실시간 중계 여부(atConference)에 따라 좋아요('likes') 클릭 수에 평균 차이가 있는지 적절한 검정을 수행하고 검정 결과의 t-statistic 를 기술하시오.
 - 주어진 데이터가 정규성과 등분산 조건을 만족한다는 가정을 하며, t-statistic 은 절대치를 취한 후 소수점 다섯째 자리 이하는 버리고 넷째 자리까지만 기술

* 실시간 중계 여부(atConference)를 예측하는 모델을 생성하기 위해 **basetable1**에 다음의 전처리를 수행하시오.

단계 1-2. No, atConference, day, totalSum 를 제외한 10개 변수에 Min-Max Scaler를 적용하시오.

단계 1-3. day 변수를 이용하여 요일변수(변수명: weekday, 월요일/화요일/.../일요일)를 추가하시오.
- 파이썬의 경우 Series.dt.day_name(locale='ko_kr') 함수 활용

단계 1-4. 다음 조건에 따라 train 과 test 데이터셋을 생성하시오.
- train: No 가 4 의 배수가 아닌 데이터, test: No 가 4 의 배수인 데이터

2. 실시간 중계 여부(atConference)를 예측하는 K-NN 분류 모델을 만들고자 한다.

- 독립 변수(총 10 개): Min-Max Scaler 를 적용한 변수 사용
- 분류 기준: Euclidean 거리를 이용하여 가까운 5 개 이웃(Neighbor)의 실시간 중계 여부(atConference) 참조

도출된 KNN 분류 모델을 통하여 Test 데이터셋의 실시간 중계 여부(atConference) 예측 시, 실제로는 실시간 중계지만(atConference="TRUE") 실시간 중계가 아닌 것으로 예측한(predicted="FALSE") 데이터의 수는 몇 개인가?

3. (단계 1-4 에서 생성된 train/test dataset 활용) 실시간 중계 여부(atConference)를 예측하는 로지스틱 회귀 분류 모델을 만들고자 한다.

- 독립 변수(16 개) : No, atConference, day, totalSum 를 제외한 수치형 변수(minmax 표준화 적용), weekday 의 더미 변수(6 개)
- 수치형 변수: 'impressions', 'retweets', 'likes', 'userprofileclicks', 'urlClicks', 'hashtagClicks', 'detailExpands', 'follows', 'mediaViews', 'mediaEngagements'
- 모델 가이드 : Inverse of Regularization= 100000, seed=123, Penalty: l2, solver: newton-cg
- 모델 도출 시 train 데이터셋 사용

도출된 모델에 test 데이터셋을 적용하여 실시간 중계 여부(atConference)가 "TRUE"인 데이터에 대한 Log Probability TRUE 의 합계를 구하시오.(절대값을 취한 후 소수점 다섯째 자리 이하는 버리고 넷째 자리까지만 기술)

- Brightics 와 파이썬 패키지 버전 차이로 약간의 차이가 발생할 수 있음

문제 유형5

- 휴대폰 음질 측정 데이터로 음질의 특징을 분석하고자 한다. MOS(Mean Opinion Score) 변수는 두 지점간의 통신망과 테스트 장비를 이용하여 측정한 음질 측정 지표이며, 다음은 데이터셋에 대한 정보이다.

phone.csv (구분자: "," , 105,828 Rows, 10 Columns)

변수명	설명	데이터 타입
Date	테스트 실행 시점	String
Signal	테스트 시 측정된 신호 강도 (단위: dBm)	double
Speed	테스트 시 측정된 전송 속도 (단위: m/s)	double
Distance	테스트 시 두 지점 간의 거리 (단위: m)	double
Manufacturer	제조사	String
TestResult	테스트 결과 (SUCCESS/FAILURE - DROP CALL/FAILURE - SETUP FAIL)	String
TestTechnology	테스트 시 사용된 이동 통신망 유형 (GSM/LTE/UMTS)	String
SetupTime	테스트 시 소요된 통화 연결 시간(s)	double
MOS	테스트 시 측정된 음질 측정 지표	double
Acceptability	테스트 시 서비스 만족여부(0[불만족/1[만족]/nan[결측치]]	double/string

*** 다음의 전처리를 먼저 수행한 후 분석 진행하시오.**

단계 1-1. 데이터 중 테스트 결과(TestResult)가 SUCCESS이고 신호 강도(Signal) 값이 결측치(null, N/A 등)가 아닌 데이터만 추출

단계 1-2. 결측치 보정

1) 두 지점 간의 거리(Distance) 변수의 결측치는 통신망 유형별(TestTechnology)로 두 지점 간의 거리(Distance) 값이 결측치가 없는 데이터로 다음과 같은 선형식을 만들고 이를 이용하여 결측치를 보정 (OLS(Ordinary Least Squares) 적용)

$$Dis\ tan\ tan\ c\ e = b_1 \times Signal$$

- 모델 생성 시 거리(Distance) 변수를 기준으로 결측치가 없는 데이터를 train용으로 구성
- 결측치 보정 시 거리(Distance) 변수를 기준으로 결측치가 있는 데이터를 생성된 모델에 적용해서 보정
- 이때 생성된 모델은 통신망 유형(TestTechnology)별로 생성되어야 하며, 결측치 보정 시에도 적용해야 함

2) 통화 연결 시간(SetupTime)이 0인 데이터는 통신망 유형(TestTechnology)별로 '통화 연결 시간이 0이 아닌 데이터들의 평균 값'으로 보정

3) 전송 속도(Speed)가 0 이하인 데이터는 1) 단계와 동일하게 통신망 유형별(TestTechnology)로 전송 속도(Speed) 값이 0 초과(>)인 데이터를 사용하여 선형식을 만들고 이를 이용하여 추정해 보정

$$Speed = b_2 \times Signal$$

- 모델 생성 시 전송 속도(Speed) 변수를 기준으로 0초과인 데이터를 train용으로 구성

- 전송 속도(Speed) 0 이하인 데이터 보정은 전송 속도(Speed) 변수 기준으로 0이하인 데이터를 생성된 모델에 적용해서 보정
- 이때 생성된 모델은 통신망 유형(TestTechnology)별로 생성되어야 하며, 전송 속도(Speed) 0 이하인 데이터 보정 시에도 반영해서 보정해야 함

*** 상기 전처리를 완료한 데이터 프레임(데이터 프레임명: basetable1, 105,147 rows)으로 다음 분석(문제 1~3) 진행하시오.**

1. (basetable1 이용) 통신망 유형(TestTechnology)별로 통화 연결 시간의 이상치(Outlier)를 제거한 후 통신망 유형(TestTechnology)별로 통화 연결 시간(SetupTime)에 차이가 있는지 검정하고 검정 통계량(F Statistic)을 기술하시오.(이때 이상치는 Tukey 방식 적용하며, 검정 통계량은 정수부분만 기술)

- 이때 집단 간의 평균 차이 검정 시 정규성과 등분산을 가정함
- [Tukey 방식]
  Q1: 제 1사분위수, Q3: 제 3사분위수, IQR: Q3 – Q1 일 때,
  (Y < Q1 - 1.5*IQR) or (Y > Q3 + 1.5*IQR) 이면 Y 이상치

단계 1-3. (basetable1 이용) 테스트 시점의 날짜(Date)의 요일 변수(Day) 생성 후 요일 변수(Day)와 통신망 유형(TestTechnology) 변수의 가변수(Dummy)를 생성하시오.
- Brightics의 경우 drop_last=True, 파이썬의 경우 drop_first=True 적용

단계 1-4. (단계1-3. 이용) Split Data 함수를 사용하여 Train과 Test 데이터를 분할하시오.

- Train:Test = 7:3 비율, 음질의 만족 여부(Acceptability)에 따라 층화추출 적용, seed(random_state) : 123
- Python인 경우 Acceptability 변수를 기준으로 결측치를 제외한 데이터셋으로 train과 test 분할하며, 이때 sklearn.model_selection.train_test_split() 함수 사용
-

2. (단계 1-3. 생성된 데이터셋 이용) 음질의 만족 여부(Acceptability)를 예측하는 KNN 분류 모델을 만드시오.
- Train Data 사용
- 독립변수(총 6개): 신호 강도(Signal), 전송 속도(Speed), 두 지점 간의 거리(Distance), 통화 연결 시간(SetupTime), 통신망 유형(TestTechnology)의 가변수(2개)

생성된 모델에 Test 데이터를 적용하여 음질의 만족 여부(Acceptability)를 예측한 후 모델의 성능평가로 AUC(the Area Under a ROC Curve) 값을 기술하시오.
- KNN 분류 모델 가이드 : K=5
- AUC 값은 소수점 셋째 자리 이하는 버리고 둘째 자리까지 기술

3. 음질의 만족 여부(Acceptability)를 예측하는 모델을 Decision Tree 알고리즘을 이용하여 만들고자 한다. 통신망 유형(TestTechnology)별로 분리해서 모델을 다음 조건에 따라 만드시오.

- Train Data 사용
- 독립변수(총 10개): 신호 강도(Signal), 전송 속도(Speed), 두 지점 간의 거리(Distance), 통화 연결 시간(SetupTime), 요일의 가변수(6개)

이 모델에 Test데이터를 적용하여 음질의 만족 여부(Target)를 예측한 후 정확도(Accuracy) 값을 기술하시오.

- [Decision Tree 모델 가이드]

분순도 기준: Gini, Max Depth: 6, Min Samples Splits: 5, Seed: 1234, 그 외: Default

- 정확도(Accuracy) 값은 소수점 둘째 자리까지 기술, 이후 자리는 절삭.

문제 유형5

- 휴대폰 음질 측정 데이터로 음질의 특징을 분석하고자 한다. MOS(Mean Opinion Score) 변수는 두 지점간의 통신망과 테스트 장비를 이용하여 측정한 음질 측정 지표이며, 다음은 데이터셋에 대한 정보이다.

phone.csv (구분자: "," , 105,828 Rows, 10 Columns)

변수명	설명	데이터 타입
Date	테스트 실행 시점	String
Signal	테스트 시 측정된 신호 강도 (단위: dBm)	double
Speed	테스트 시 측정된 전송 속도 (단위: m/s)	double
Distance	테스트 시 두 지점 간의 거리 (단위: m)	double
Manufacturer	제조사	String
TestResult	테스트 결과 (SUCCESS/FAILURE - DROP CALL/FAILURE - SETUP FAIL)	String
TestTechnology	테스트 시 사용된 이동 통신망 유형 (GSM/LTE/UMTS)	String
SetupTime	테스트 시 소요된 통화 연결 시간(s)	double
MOS	테스트 시 측정된 음질 측정 지표	double
Acceptability	테스트 시 서비스 만족여부(0[불만족/1[만족]/nan[결측치]]	double/string

*** 다음의 전처리를 먼저 수행한 후 분석 진행하시오.**

단계 1-1. 데이터 중 테스트 결과(TestResult)가 SUCCESS이고 신호 강도(Signal) 값이 결측치(null, N/A 등)가 아닌 데이터만 추출

단계 1-2. 결측치 보정

4) 두 지점 간의 거리(Distance) 변수의 결측치는 통신망 유형별(TestTechnology)로 두 지점 간의 거리(Distance) 값이 결측치가 없는 데이터로 다음과 같은 선형식을 만들고 이를 이용하여 결측치를 보정 (OLS(Ordinary Least Squares) 적용)

$$Dis\ tan\ tan\ c\ e = b_1 \times Signal$$

- 모델 생성 시 거리(Distance) 변수를 기준으로 결측치가 없는 데이터를 train용으로 구성
- 결측치 보정 시 거리(Distance) 변수를 기준으로 결측치가 있는 데이터를 생성된 모델에 적용해서 보정
- 이때 생성된 모델은 통신망 유형(TestTechnology)별로 생성되어야 하며, 결측치 보정 시에도 적용해야 함

5) 통화 연결 시간(SetupTime)이 0인 데이터는 통신망 유형(TestTechnology)별로 '통화 연결 시간이 0이 아닌 데이터들의 평균 값'으로 보정

6) 전송 속도(Speed)가 0 이하인 데이터는 1) 단계와 동일하게 통신망 유형별(TestTechnology)로 전송 속도(Speed) 값이 0 초과(>)인 데이터를 사용하여 선형식을 만들고 이를 이용하여 추정해 보정

$$Speed = b_2 \times Signal$$

- 모델 생성 시 전송 속도(Speed) 변수를 기준으로 0초과인 데이터를 train용으로 구성

- 전송 속도(Speed) 0 이하인 데이터 보정은 전송 속도(Speed) 변수 기준으로 0이하인 데이터를 생성된 모델에 적용해서 보정
- 이때 생성된 모델은 통신망 유형(TestTechnology)별로 생성되어야 하며, 전송 속도(Speed) 0 이하인 데이터 보정 시에도 반영해서 보정해야 함

*** 상기 전처리를 완료한 데이터 프레임(데이터 프레임명: basetable1, 105,147 rows)으로 다음 분석(문제 1~3) 진행하시오.**

1. (basetable1 이용) 통신망 유형(TestTechnology)별로 통화 연결 시간의 이상치(Outlier)를 제거한 후 통신망 유형(TestTechnology)별로 통화 연결 시간(SetupTime)에 차이가 있는지 검정하고 검정 통계량(F Statistic)을 기술하시오.(이때 이상치는 Tukey 방식 적용하며, 검정 통계량은 정수부분만 기술)

- 이때 집단 간의 평균 차이 검정 시 정규성과 등분산을 가정함
- [Tukey 방식]
  Q1: 제 1사분위수, Q3: 제 3사분위수, IQR: Q3 – Q1 일 때,
  (Y < Q1 - 1.5*IQR) or (Y > Q3 + 1.5*IQR) 이면 Y 이상치

단계 1-3. (basetable1 이용) 테스트 시점의 날짜(Date)의 요일 변수(Day) 생성 후 요일 변수(Day)와 통신망 유형(TestTechnology) 변수의 가변수(Dummy)를 생성하시오.
- Brightics의 경우 drop_last=True, 파이썬의 경우 drop_first=True 적용

단계 1-4. (단계1-3. 이용) Split Data 함수를 사용하여 Train과 Test 데이터를 분할하시오.

- Train:Test = 7:3 비율, 음질의 만족 여부(Acceptability)에 따라 층화추출 적용, seed(random_state) : 123
- Python인 경우 Acceptability 변수를 기준으로 결측치를 제외한 데이터셋으로 train과 test 분할하며, 이때 sklearn.model_selection.train_test_split() 함수 사용
-

2. (단계 1-3. 생성된 데이터셋 이용) 음질의 만족 여부(Acceptability)를 예측하는 KNN 분류 모델을 만드시오.
- Train Data 사용
- 독립변수(총 6개): 신호 강도(Signal), 전송 속도(Speed), 두 지점 간의 거리(Distance), 통화 연결 시간(SetupTime), 통신망 유형(TestTechnology)의 가변수(2개)

생성된 모델에 Test 데이터를 적용하여 음질의 만족 여부(Acceptability)를 예측한 후 모델의 성능평가로 AUC(the Area Under a ROC Curve) 값을 기술하시오.
- KNN 분류 모델 가이드 : K=5
- AUC 값은 소수점 셋째 자리 이하는 버리고 둘째 자리까지 기술

3. 음질의 만족 여부(Acceptability)를 예측하는 모델을 Decision Tree 알고리즘을 이용하여 만들고자 한다. 통신망 유형(TestTechnology)별로 분리해서 모델을 다음 조건에 따라 만드시오.

- Train Data 사용
- 독립변수(총 10개): 신호 강도(Signal), 전송 속도(Speed), 두 지점 간의 거리(Distance), 통화 연결 시간(SetupTime), 요일의 가변수(6개)

이 모델에 Test데이터를 적용하여 음질의 만족 여부(Target)를 예측한 후 정확도(Accuracy) 값을 기술하시오.

- [Decision Tree 모델 가이드]

분순도 기준: Gini, Max Depth: 6, Min Samples Splits: 5, Seed: 1234, 그 외: Default

- 정확도(Accuracy) 값은 소수점 둘째 자리까지 기술, 이후 자리는 절삭.

문제 유형 6

여행자 보험 청구 여부를 예측하는 모델을 개발하고자 한다. 데이터셋은 다음과 같다.

travel_insurance2.csv (구분자: 쉼표(,), 63,326 Rows, 12 Columns, UTF-8 Encoding, 결측치 존재(Null, N/A, Empty Cell)

변수명	설명	타입	변수명	설명	타입
NO	ID	Double	Duration	구매자 여행 기간	Double
Agency	여행상품 판매 기관	String	Destination	구매자 여행 국가	String
AgencyType	여행상품 판매처 타입	String	NetSales	순매출액	Double
DistributionChannel	유통 채널	String	Commision	수수료율(%)	Double
ProductName	상품명	String	Gender	구매자 성별	String
Claim	보험 청구 여부(Yes/No)	String	Age	구매자 연령	Double

*** 분석 수행하기 전, 다음 전처리 단계들을 순서대로 수행하시오.**

단계 1-1 ProductName, NetSales 변수를 제거하시오.

단계 1-2. 성별(Gender)에 결측치가 존재한다. 다음 규칙에 따라 성별의 결측치를 보정하시오.

If (NO=짝수) then (Gender='M')

Else if (NO=홀수) then (Gender='F')

Else 성별(Gender)이 결측치가 아닌 경우 현재 값 유지

*** 상기 전처리 완료 후(데이터 프레임명: basetable1) 다음 분석(문제 1~3)을 수행하시오.**

1. (basetable1 이용) 40대(40<=Age<50)이고 여성(Gender='F')인 여행자가 두 번째로 많이 방문하는 여행지 국가(Destination)를 기술하시오(여행지 국가(Destination)는 대문자로 기술, 예시: ITALY)

2. (basetable1 이용) 유통 채널(DistributionChannel)에 따라 여행 기간(Duration)의 평균이 통계적으로 차이가 있는지 적절한 검정을 수행한 후, 검정 통계량(t-value)의 절댓값을 기술하시오.
   - 등분산 가정하고 유의수준 0.05 하에서 양측검정
   - 검정 통계량(t-value)의 절댓값은 소수점 둘째 자리까지만 기술(이후 자리는 버림, 예시: 0.12)

3. (basetable1 이용) 마지막으로 청구(Claim) 여부를 예측하는 로지스틱 회귀 모형을 만들고자 합니다. 다음의 전처리를 수행한 후 분석을 수행하시오.

단계 1-3. 범주형 변수인 성별(Gender)와 유통 채널(DistributionChannel)에 대한 가변수(총 2개)를 추가하시오.

- 가변수화 Hint

  (Brightics Studio) : One Hot Encoder 사용, Drop Last=True

  (Python) : pd.get_dummies() 함수 사용, drop_first=True

단계 1-4. 다음과 같이 Train과 Test 데이터셋을 분리하시오.

- Train DataSet: NO가 3의 배수가 아닌 데이터

- Test DataSet: NO가 3의 배수인 데이터

단계 1-5. Train DataSet으로 로지스틱 회귀분석 모델(Logistic Regression Model)을 학습하시오.

- 독립변수(총 5개): Duration, Commision, Age,

  성별(Gender)과 유통 채널(DistributionChannel)의 가변수(단계 1-3.에서 생성된 변수)

- 종속변수: 청구 여부(Claim)

- 함수 가이드(Brightics Studio)

Inverse of Regularization=100,000, Seed=1234, Penalty: l2, Solver: newton-cg, 나머지: Default

단계 1-6. **단계 1-5.**에서 학습한 모델을 **단계 1-4.**에서 분리한 'Test DataSet'에 적용하여 청구 확률 (Claim=Yes 일 확률)을 구한 후, Lift 변수를 추가하시오.

Lift = 청구 확률/0.015

(※ 0.015: 전체 데이터 셋에서 Claim=Yes인 비율, 반올림 적용)

데이터셋을 청구 확률 (Claim=Yes 일 확률) 역순으로 (내림 차순, DESC ORDER) 정렬한 뒤, 상위 100등 안쪽(<=, 이하)에 포함되는 데이터들에 대한 Lift의 합계를 기술하시오.(Lift의 합계는 정수 부분만 기술(예시: 12)

- Hint(Brightics Studio) : rank() over (partition by A order by B desc/asc) 활용

□ 대문항 세트 1

마케팅 전략을 수립하기 위해 신용 카드 고객을 대상으로 고객 세분화(Customer Segmentation) 및 예측 모델링을 수행하고자 한다.

DS_Sample_1.csv (구분자: comma(","), 1,000 Rows, 18 Columns, UTF-8 인코딩)

컬럼	**정의**	**Type**
CUST_ID	고객 ID	Double
BALANCE	연간 평균 잔고액	Double
BALANCE_FREQUENCY	연중 잔고액 갱신 개월 수 비율 (0~1 사이값)	Double
PURCHASES	구매 총액	Double
ONEOFF_PURCHASES	일시불 구매 총액	Double
INSTALLMENTS_PURCHASES	할부 구매 총액	Double
CASH_ADVANCE	현금서비스 구매 총액	Double
PURCHASES_FREQUENCY	연중 구매 개월 수 비율 (0~1 사이값)	Double
ONEOFF_PURCHASES_FREUQUENCY	연중 일시불 구매 개월 수 비율 (0~1 사이값)	Double
PURCHASES_INSTALLMENTS_FREQUENCY	연중 할부 구매 개월 수 비율 (0~1 사이값)	Double
CASH_ADVANCE_FREQUENCY	연중 현금서비스 구매 개월 수 비율	Double
CASH_ADVANCE_TRX	현금 서비스 구매 횟수	Double
PURCHASES_TRX	구매 횟수	Double
CREDIT_LIMIT	신용카드 한도	Double
PAYMENTS	지불 총액	Double
MINIMUM_PAYMENTS	기한 내 최소 지불 금액	Double
PRC_FULL_PAYMENT	연중 기한 내 전액 지불 개월 수 비율 (0~1 사이값)	Double
TENURE	신용카드 서비스 이용기간	Double

- 필요 패키지/라이브러리 목록

Brightics	

R	dplyr, data.table, tidyr, cluster, tree
Python	import pandas as pd import numpy as np from sklearn.metrics import silhouette_samples, silhouette_score from sklearn.cluster import KMeans from sklearn.tree import DecisionTreeRegressor

분석을 수행하기 전, 상기 데이터를 이용하여 아래의 전처리를 수행하시오.

단계 1: '신용카드 한도(CREDIT_LIMIT)'와 '기한 내 최소 지불 금액(MINIMUM_PAYMENTS)'의 결측 값(Null)을 각 컬럼의 평균값으로 대체하시오.

상기 전처리를 완료한 데이터셋(데이터셋명: **card1**)을 이용하여 다음 1~3 번 문제에 답하시오.

1. (card1 을 이용하여) '연간 평균 잔고액(BALANCE)'이 많을수록, 그리고 '신용카드 서비스 이용기간(TENURE)'이 길수록 '신용카드 한도(CREDIT_LIMIT)' 역시 높을 것으로 예상해볼 수 있다. 이들 변수 간의 관계를 파악하여, 추후 고객의 신용카드 한도 조정에 근거 자료로 활용하고자 한다.

'신용 카드 서비스 이용기간(TENURE)' 별로 '연간 평균 잔고액(BALANCE)'과 '신용카드 한도(CREDIT_LIMIT)' 간 피어슨(Pearson) 상관 계수를 계산하고, 이 중 가장 큰 값을 구하시오.

- 소수점 셋째 자리에서 반올림하여 소수점 둘째 자리까지 기술하시오.

(답안예시) 0.12

2. (card1을 이용하여) 신용카드 판매 전략을 수립하기 위해 고객 세분화(Customer Segmentation) 를 수행할 수 있다. 일시불 구매 금액이 높은 고객 Segment를 도출하기 위해 다음 단계에 따라 분석을 수행하고 질문에 답하시오.

단계 1: '고객 ID(CUST_ID)'를 제외한 모든 변수(17 개)에 대해 Z-score 표준화(Standardization) 한다.
단계 2: 표준화된 변수들에 대해 K-means 군집 분석을 수행한다. 이 때, 군집 수는 2~5 개 중 K-means Silhouette 를 통해 구한 최적의 K 로 설정한다.
단계 3: 단계 2 에서 도출한 각 군집 별로 '일시불 구매 총액(ONEOFF_PURCHASES)'의 평균을 계산한다.

Brightics	Seed=1234 문제 지시 외 Default 값 사용
R	library(cluster) set.seed(12345) 표준화 : scale() 함수의 center=T, scale=T 옵션 사용 Silhouette : silhouette() 함수의 sil_width 평균값 기준

	문제 지시 외 Default 값 사용
Python	from sklearn.metrics import silhouette_samples, silhouette_score from sklearn.cluster import KMeans random_state=1234 문제 지시 외 Default 값 사용

군집 별 '일시불 구매 총액(ONEOFF_PURCHASES)'의 평균 중 가장 큰 값을 구하시오.

- 소수점 셋째 자리에서 반올림하여 소수점 둘째 자리까지 기술하시오.

(답안예시) 1200.34

3. (card1을 이용하여) 이번에는 '일시불 구매 총액(ONEOFF_PURCHASES)' 예측 모델을 Target Marketing에 활용하고자 한다. 다음 단계에 따라 분석을 수행하고 질문에 답하시오.

단계 1: '고객 ID(CUST_ID)'가 4의 배수가 아닌 데이터를 Train Set으로, 4의 배수인 데이터를 Test Set으로 분할한다.

단계 2: Train Set으로 아래 조건에 따라 의사결정나무 회귀모델을 학습한다.

- 독립 변수(총 16개): '고객 ID(CUST_ID)', '일시불 구매 총액(ONEOFF_PURCHASES)'을 제외한 모든 컬럼
- 종속 변수: '일시불 구매 총액(ONEOFF_PURCHASES)'
- 툴별 가이드

Brightics	Seed=1234 문제 지시 외 Default 값 사용
R	set.seed(1234) library(tree) Decision Tree Regression : tree() 함수 사용 문제 지시 외 Default 값 사용
Python	from sklearn.tree import DecisionTreeRegressor random_state=1234 문제에서 지시한 것 외에는 Default 값 사용

단계 3: 생성된 모델을 Test Set에 적용하여 '일시불 구매 총액(ONEOFF_PURCHASES)'을 예측한다.

단계 3에서 얻은 예측 결과를 평가하기 위해, 아래 정의된 Measure B를 구하시오.

$$\mathrm{B} = \left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right)^{\frac{1}{2}}$$

$\hat{y}_i$: 예측값, y_i: 실제값

- B는 소수점 둘째 자리에서 반올림하여 소수점 첫째 자리까지 기술하시오.

(답안예시) 1200.3

□ 대문항 세트 2

교육 전문기관 분석팀에서는 교육 수강자의 정보를 바탕으로 진로설계 프로그램을 운영하기 위해 다음 정보를 수집하였다.

DS_Sample_2.csv (구분자: comma(","), 19,158 Rows, 15 Columns, UTF-8 인코딩)

컬럼	**정의**	**Type**
enrollee_id	수료자 ID	Double
city	도시 코드	String
city_development_index	도시 발전 지표	Double
gender	성별	String
relevant_experience	관련 분야 경험 여부	String
enrolled_university	수강 과목명	String
education_level	학력	String
major_discipline	전공	String
experience	경력	String
company_size	현 직장 직원 수	String
company_type	현 직장 유형	String
last_new_job	전 직장 근속연수	String
training_hours	수료 시간	Double
target	이직 희망 여부 (0: 비희망, 1: 희망)	Double
Xgrp	Train/Test Set 구분	String

- 필요 패키지/라이브러리 목록

Brightics	
R	data.table, dplyr, tidyr, class
Python	Import pandas as pd Import numpy as np from sklearn.linear_model import LogisticRegression from sklearn.neighbors import KNeighborsClassifier

분석을 수행하기 전, 상기 데이터를 이용하여 아래의 전처리를 수행하시오.

단계 1: 분석에 사용하지 않을 city, company_size, company_type 컬럼을 제거하시오.
단계 2: 각 문자형(String Type) 컬럼에 결측치(null/empty space)가 하나라도 존재하는 행(row)은 모두 제거하시오.
단계 3: experience 컬럼의 값이 '>20' 또는 '<1'인 값을 제거하고 experience 컬럼의 타입을 정수형(Integer)으로 변환하시오.
단계 4: last_new_job 컬럼의 값이 '>4' 또는 'never'인 값을 제거하고 last_new_job 컬럼의 타입을 정수형(Integer)으로 변환하시오.

상기 전처리를 완료한 데이터셋(데이터셋명: **job2**, 7,522 Rows)을 이용하여 다음 4~6 번 문제에 답하시오.

4. (job2 를 이용하여) '관련 분야 경험 여부(relevant_experience)'에 따른 '이직 희망 여부(target)'를 기술통계량으로 확인하고자 한다.
관련 분야 경험이 없는(relevant_experience='No relevant experience') 수료자 중 이직을 희망(target='1')하는 수료자의 비율을 A, 관련 분야 경험이 있는(relevant_experience='Has relevant experience') 수료자 중 이직을 희망(target='1')하는 수료자의 비율을 B 라 할 때, A/B 를 구하시오.

- 소수점 셋째 자리에서 반올림하여 소수점 둘째 자리까지 기술하시오.

(답안예시) 12.34

5. (job2 를 이용하여) '이직 희망 여부(target)'에 영향을 주는 변수들을 확인하고자 한다. 다음 절차에 따라 로지스틱 회귀분석(Logistic Regression)을 수행하고 질문에 답하시오.
단계 1: gender, relevant_experience, enrolled_university, education_level, major_discipline 변수로부터 더미 변수들을 생성한다. 단, 각 변수로부터 더미 변수를 생성할 때 마지막으로 등장하는 범주는 제외하도록 한다.
(여기서 마지막으로 등장하는 범주란, 각 컬럼의 값을 사전 순으로 나열하였을 때 마지막으로 등장하는 값이다. 예를 들어, 'col' 변수의 범주가 ['A','C','B','B','C','A']의 값을 가진다면 사전 상의 마지막 값인 C 가 제외된다)
단계 2: 단계 1 에서 생성한 더미 변수와 city_development_index, experience, last_new_job, training_hours, target, Xgrp 변수를 결합하여 새로운 데이터셋(데이터셋명: **job2_2**, 이 데이터셋은 문제 6 에서도 활용)을 구성한다. 이 때, target, Xgrp 를 제외한 데이터셋의 컬럼은 아래 순서에 따르도록 한다.

- city_development_index
- experience
- last_new_job
- training_hours
- gender 의 더미 변수
- relevant_experience 의 더미 변수
- enrolled_university 의 더미 변수
- education_level 의 더미 변수
- major_discipline 의 더미 변수

단계 3. 단계 2 에서 구성한 데이터셋 **job2_2** 로 다음 조건에 따라 상수항(Intercept)이 포함된 로지스틱 회귀분석을 수행한다.

- 종속 변수 : target
- 독립 변수(총 16 개) : target 과 Xgrp 를 제외한 나머지 변수

- 회귀식에 포함되는 독립 변수의 순서를 컬럼의 순서와 일치시킨다.

Brightics	Inverse of Regularization=100000 Seed=1234 문제 지시 외 Default 값 사용
R	glm() 문제 지시 외 Default 값 사용
Python	import pandas as pd pd.get_dummies() 사용 from sklearn.linear_model import LogisticRegression C=100000, random_state=1234 문제 지시 외 Default 값 사용

상수항을 제외한 나머지 변수들에 대한 Odds Ratio 중 가장 큰 값을 기술하시오.

$$x_i\text{의 Odds Ratio} = \frac{odds(P(Y=1|x_1,\ldots,x_i+1,\ldots,x_n))}{odds(P(Y=1|x_1,\ldots,x_i,\ldots,x_n))}$$

- 소수점 셋째 자리에서 버림하여 둘째 자리까지 기술하시오.

(답안예시) 12.34

6. (job2_2를 이용하여) 전체 데이터를 Train과 Test Set으로 나누고, Train Set으로 학습한 모델을 Test Set에 적용하여 모델을 평가하고자 한다. 다음 절차에 따라 분석을 수행하고 질문에 답하시오.

단계 1: 문제 5번 2단계에서 구성한 데이터셋 job2_2에서 Xgrp 컬럼의 값이 'train'인 경우 Train Set으로, 'test'인 경우 Test Set으로 정의하여 분할한다.

단계 2: 아래 가이드에 따라 Train Set으로 K-NN 분류 모델을 학습하고, 이 모델을 Test Set에 적용한다.

- 종속 변수: 이직 희망 여부(target)
- 독립 변수(총 16개): 이직 희망 여부(target)와 Train/Test set 구분 변수(Xgrp)를 제외한 모든 변수
- Euclidean 거리 기준 가장 가까운 5개 데이터의 '이직 희망 여부(target)'를 활용하여 예측

Brightics	문제에서 지시한 것 외에는 Default 값 사용
R	set.seed(1234) library(class)

	KNN : knn() 함수 사용 문제에서 지시한 것 외에는 Default 값 사용
Python	from sklearn.neighbors import KNeighborsClassifier 문제 지시 외 Default 값 사용

단계 3: 예측 결과를 바탕으로 아래 정의된 지표 A를 계산하여 기술하시오.

$$A = \frac{(\#\ \text{of true positive}) + (\#\ \text{of true negative})}{(\#\ \text{of total data})}$$

- A는 소수점 셋째 자리에서 반올림하여 소수점 둘째 자리까지 기술하시오.

(답안예시) 0.12

□ 대문항 세트3

영화 스트리밍 사이트 운영진은 등록된 영화들에 대한 데이터를 분석하여, 개봉 년도, 평점, 투표 수, 감독 등 변수들 간 관계에 대해 알아 보고자 한다.

DS_Sample_3.csv (구분자: comma(","), 188 Rows, 5 Columns, UTF-8 인코딩)

컬럼	정의	Type
Title	영화 제목	String
AirDate	영화 공개 날짜	String
Rating	평점	Double
Num_Votes	투표 수(평점을 남긴 수)	Double
DirectedBy	감독	String

- 필요 패키지/라이브러리 목록

Brightics	
R	data.table, dplyr, tidyr
Python	import pandas as pd import numpy as np import datetime from statsmodels.stats.anova import anova_lm from statsmodels.formula.api import ols from sklearn.linear_model import LinearRegression

분석을 수행하기 전, 상기 데이터를 이용하여 아래의 전처리를 수행하시오.
단계 1: 영화 공개 날짜(AirDate)의 연도 정보를 기준으로, 2005 년에서 2007 년은 'A', 2008 년에서 2010 년은 'B', 2011 년에서 2013 년은 'C' 값을 가지는 공개년도그룹(변수명: group) 변수를 생성하시오.
상기 전처리를 완료한 데이터셋(데이터셋명: **movie3**)을 이용하여 다음 7~9번 문제에 답하시오.

7. **(movie3 을 이용하여)** 감독 별로 영화 흥행 결과를 알아보고자 한다. 여기서 영화의 흥행 변수 'success'는 다음과 같이 정의된다.

$$\text{success}_i = \text{Rating}_i \times \text{Num\_Votes}_i$$

'감독(DirectedBy)' 별로 '흥행(success)' 변수의 평균을 계산하고, 이 중 가장 큰 값을 소수 부분을 버리고 정수 부분만 기술하시오.

(답안예시) 1234

8. (movie3 을 이용하여) 영화 평가에 참여한 총 투표 규모는 영화의 컨텐츠 뿐만 아니라 서비스 공개 시기에 따라 달라질 가능성이 있다. 영화 공개 년도에 따라 투표 수에 차이가 존재하는지 검정하기 위해 수립한 가설은 다음과 같다.
대립 가설(H_1): 공개년도그룹(group)에 따라 평균 투표 수(Num_Votes)는 달라진다.
적절한 검정 방법을 택하여 위 가설을 검정할 때, 검정통계량 값(A)과 요인의 자유도(B)의 합(A+B)을 기술하시오.

- 그룹 간 등분산성 및 정규성은 만족한다고 가정한다.
- 툴별 가이드

Brightics	문제 지시 외 Default 값 사용
R	aov() 함수 사용 문제 지시 외 Default 값 사용
Python	from statsmodels.stats.anova import anova_lm from statsmodels.formula.api import ols 문제 지시 외 Default 값 사용

소수점 셋째 자리에서 반올림하여 소수점 둘째 자리까지 기술하시오.

(답안예시) 12.34

9. (movie3 을 이용하여) 행동 경제학에서는 개인의 의사와 무관하게 집단 전체가 동일한 방향으로 움직이는 군집 행동(Herding Behavior)이 존재한다고 본다. 이 이론에 따르면 특정 영화 관람객이 많아지고 대중적 인기를 끌수록 그 영화에 대한 평균 평점은 계속해서 높아질 가능성이 있다. 이 현상이 실제로 나타나는지 알아보기 위해 '투표 수(Num_Votes)'와 '평점(Rating)' 간 관계를 분석하고자 한다.
아래 가이드에 따라 OLS(Ordinary Least Squares) 방식의 선형회귀 모델링을 수행하고 질문에 답하시오.
- 종속 변수: 평점(Rating)
- 독립 변수(총 3 개): 공개년도그룹(group)를 one-hot encoding 한 결과 컬럼(2 개), 투표 수(Num_Votes)
- 회귀식에 상수항을 포함할 것

Brightics	문제 지시 외 Default 값 사용
R	문제에서 지시한 것 외에는 Default 값 사용
Python	from sklearn.linear_model import LinearRegression 문제 지시 외 Default 값 사용

추정된 모델을 기반으로, 투표 수(Num_Votes)가 5000, 공개년도그룹(group)이 C 인 경우 평점(Rating)의 예측 값을 기술하시오.

- 소수점 셋째 자리에서 반올림하여 소수점 둘째 자리까지 기술하시오.

(답안예시) 12.34

	배점	Brightics	Python	R